Sound Events and Emotions: Investigating the Relation of Rhythmic Characteristics and Arousal

Konstantinos Drossos*, Rigas Kotsakis†, George Kalliris†
and Andreas Floros*
*Audiovisual Signal Processing Laboratory
Dept. of Audiovisual Arts, Ionian University, Corfu, Greece
Email: kdrosos@ionio.gr, floros@ionio.gr
†Laboratory of Electronic Media
Dept. of Journalism and Mass Communication, Aristotle University of Thessaloniki, Thessaloniki, Greece
Email: rkotsakis@gmail.com, gkal@jour.auth.gr

***Abstract*—A variety of recent researches in Audio Emotion Recognition (AER) outlines high performance and retrieval accuracy results. However, in most works music is considered as the original sound content that conveys the identified emotions. One of the music characteristics that is found to represent a fundamental means for conveying emotions are the rhythm-related acoustic cues. Although music is an important aspect of everyday life, there are numerous non-linguistic and non-musical sounds surrounding humans, generally defined as sound events (SEs). Despite this enormous impact of SEs to humans, a scarcity of investigations regarding AER from SEs is observed. There are only a few recent investigations concerned with SEs and AER, presenting a semantic connection between the former and the listener's triggered emotion. In this work we analytically investigate the connection of rhythm-related characteristics of a wide range of common SEs with the arousal of the listener using sound events with semantic content. To this aim, several feature evaluation and classification tasks are conducted using different ranking and classification algorithms. High accuracy results are obtained, demonstrating a significant relation of SEs rhythmic characteristics to the elicited arousal.**

***Keywords*-Audio Emotion Recognition; Sound Events; Arousal; Rhythm Related Features; Audio Emotion Classification**

I. Introduction

Music is likely to be one of the primary audiovisual means used to express and convey emotions by extending and mimicking voice's characteristics [1]. Its impact on listeners' emotions is being throughly studied through various disciplines, like Music Emotion Recognition (MER), Music Psychology and Music Information Retrieval (MIR). Focusing on the former one, currently published emotion recognition accuracy results are likely to imply a connection between sound's technical characteristics and the conveyed emotions [2].

One of the main components in MER is the affective model employed for describing emotions in a qualitative manner. In the literature, two abstract affective model categories are defined, namely the discrete and the continuous models [3]. The former ones assign specific verbal descriptions for particular emotions, like "Happiness", "Fear", "Sadness" etc. The models that belong to the second category consider emotions as a resultant of two or more emotional states, illustrated as continuous values [4], [5]. Typical choices for the above emotional conditions are the Arousal and Valence, the employment of which result into a two dimensional affective space [3]. The consequent emotion can be described either as the ensuing vector from the combination of the aforementioned values or with a verbal description from the discrete category. The mapping of the discrete verbal categories to continuous values is performed through clustering of the latter categories' values, as presented for example in [2].

Ordinary acoustic cues used for MER typically include features related with the measured energy, timbre, tonality and the rhythm characteristics of the music signal [3]. Moreover, additional technical characteristics have been associated with the conveyance of specific emotions, e.g. dissonance, mode and loudness [5]. Focusing particularly on the arousal dimension, there are published works stating a direct relation between arousal and energy or rhythm of a musical piece [2], [6].

Music appears to be only a segment of the heard sounds [3], since there are numerous non-linguistic and non-musical sounds that comprise an acoustic environment, defined as sound events (SEs) [7]. These events communicate to the human listener information regarding attributes of the source and its surroundings, such as the size, direction and speed of the source and/or the nature of the sound production mechanism, or even the texture of the adjacent surfaces [8]. In addition, there is the possibility for the SEs to carry also semantic content [3]. Conveyed information can have an impact on listener's perception and thus can affect his emotion [3].

This work focuses on emotion recognition from SEs. More specifically, as a starting investigation point, we examine the influence of sound events rhythm characteristics on the listener's arousal. Towards this aim, the continuous model was employed in order to avoid introducing addi-

tional complexity of the verbal descriptions of emotions and their interpretation. The selection of the rhythm characteristics was based on the fact that the connection of rhythm with arousal is also supported by many psychological researches [5], [6]. For this cause, we use the only existing, to the best of authors' knowledge, SEs database with pre-annotated affective data, the International Affective Digital Sounds (IADS) [9]. In particular, we extracted several rhythm related features, using the MIR Toolbox [14], and performed feature evaluation and data classification, with WEKA software [16], in order to evaluate the selected features and, in parallel, to examine the possibility of correct identification of arousal from SEs using only rhythm-related characteristics.

The rest of the paper is organized as follows: Section II contains a brief overview of existing researches related to the aim of the present work. Next, Section III outlines the methodology and the experimental procedure followed for deriving the obtained results that are analytically presented in Section IV. Finally, Section V concludes this work.

II. Related Research

As mentioned previously, there is only a small number of published works investigating emotion recognition from SEs. In the majority of them, the continuous model approach is considered. In a recent research [10], an accuracy of nearly 62% for arousal and 50% for valence is reported. These results were obtained by employing non-annotated SEs and four (4) human annotators. Features regarding low level descriptors of sound, voicing related characteristics, statistical, regression and local minima and maxima functionals were used. For the classification process, automated regression was employed. In addition, other researches have employed the IADS data base which incorporates a connection between the semantic content of the SEs and the conveyed emotion [3]. In these works, various technical features were used, such as the timbral, energy and rhythm characteristics of sound. For the classification task, Support Vector Machine (SVM) and Artificial Neural Network (ANN) were utilized, along with the arousal and valence dimensions for affective modeling.

At a larger research extent, focusing on music-only content, existing MER and MIR approaches exhibit relative high accuracy results for emotion recognition. Typical classification results can reach up to 85% [2]. Moreover, SVM [11], Gaussian Mixture Model (GMM) [2] and Decision Trees [12] are frequently utilized as common classification schemes. In particular, commonly employed features in MIR and MER include timbral, rhythmic, pitch and energy related technical characteristics [2], [11]–[13]. As far as affective models are concerned, both discrete and continuous models are employed in MIR and MER [3]. Although continuous models tend to offer the ability for clustering the resulting values according to specified verbal descriptions of emotions (if needed), discrete models seem to be preferred, due to their ability for targeted emotion depiction, for example in terms of arousal and valence listener rating against "Fear" and "Happiness" verbal descriptions [3]. Nevertheless, the usage of discrete models seems to introduce an discontinuity between researches due to the potential consideration of different verbal descriptions for the same emotion, e.g. "Joy" - "Enjoyment" and "Cheerfulness" - "Happiness" [1].

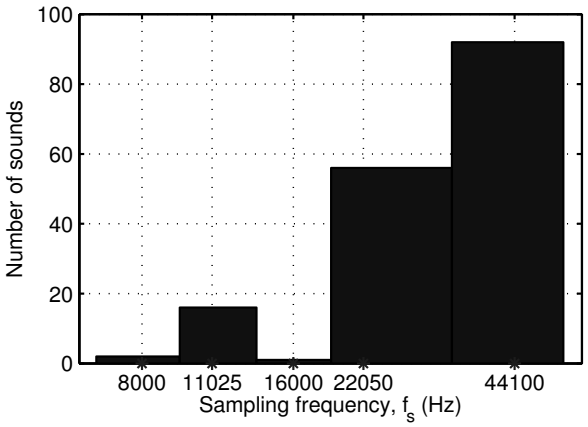

Figure 1. f_s distribution for the used data set

III. Experimental Procedure

In brief, the experimental procedure followed consisted of five stages, namely the a) data pre-processing, b) arousal values clustering, c) feature extraction, d) features' evaluation, and e) classification using the features extracted in stage (b). The IADS data set employed provides a total of 167 sounds with emotional annotation for the arousal, valence and dominance dimensions. All signal pre-processing and feature extraction tasks were performed using the MIR Toolbox [14]. Finally, for the features' evaluation and classification, the WEKA environment was used [16].

A. Data Pre-Processing

The IADS database includes sounds with different sampling frequencies (f_s). In order not to introduce any additional artificial information, the original sounds' f_s was retained in all tests. The distribution of the different f_s values within the data set is illustrated as a histogram in Figure 1. All sounds were also peak-normalized prior to any further processing, in order to avoid any perceptual side effects introduced from different signal level amplitudes.

For each sound waveform, 6 additional copies were created, resulting in a total set of 167×7 SEs. Next, the sounds were clustered in seven groups. Each group contained a single copy of each sound. Thus, the complete test data set was formed in terms of 7 clusters of 167 unique sound events. For each group, the values presented in Table I were

used in order to segment the SEs' waveforms in shorter time frames. Frames' overlap was 20% and hamming windowing function was used for all groups .

Table I
SOUNDS CLUSTERS AND FRAMES' TIME LENGTH

Group	Frame length (sec)
1	0.8
2	1.0
3	1.2
4	1.4
5	1.6
6	1.8
7	2.0

The lowest value of 0.8 seconds was chosen due to the inability of calculating the complete set of features (analytically presented in the following subsection) when smaller frame lengths are utilized. Moreover, the maximum value of 2.0 seconds was used as higher frame lengths were found to obscure any details in the variation of the rhythm related features. Finally, 20% frame overlap was employed as one of the most widely employed window overlap values.

B. Arousal values clustering

The emotional annotation over the IADS data set was performed using the Self Assessment Manikin (SAM) method [15]. This method applied annotation scores within the range $[1, 9]$ for each SE. In practice, arousal annotated values in the IADS data set range from 2.88 to 8.16. The distribution of annotated arousal values is depicted in Figure 2. In order to use these values as nominal classes in the classification process, a clustering scheme was applied.

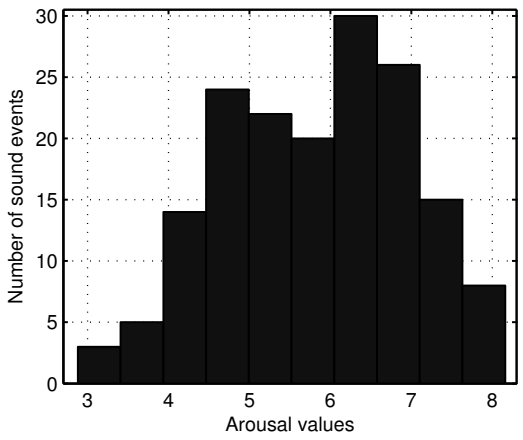

Figure 2. The arousal values distribution in the IADS SEs data base

Two groups of values were created with values below or higher/equal to the half of the maximum annotation's value, i.e. $9/2$ respectively. The amount of SEs corresponding to each cluster can be seen in Table II. Each group was used as a separate class in the classification process.

Table II
THE AMOUNT OF SOUND EVENTS FOR EACH GROUP

Group	Amount of sound events
1	24
2	143

C. Feature Extraction

Feature extraction was performed using the segmented SEs copies presented in Section III-A. The complete list of the extracted features is presented in the first column of Table III. This process resulted in a cluster of values for each sound of each group. Subsequently, for each feature's values and for all sounds in all groups, the statistical measures presented in Table III were calculated. The resulting values were divided by the total length of the corresponding SE. This lead to a 26 dimensions feature space for each sound in each group.

Table III
THE EXTRACTED FEATURES & STATISTICAL MEASURES

Extracted Features	Statistical Measures
Beat spectrum	Mean
Onsets	Standard deviation
Tempo	Gradient
Fluctuation	Kurtosis
Event density	Skewness
Pulse clarity	

The aforementioned statistical measures were used as a means to univocally describe variation's characteristics for each feature along the time axis, corresponding to the sequential sound frames. In addition, the impact of the differences in time lengths of the original sound data were minimized by the ratio of the statistical measure and the total number of sound samples in the original signal.

D. Features' information evaluation

Prior to features' evaluation, an initial correlation analysis was conducted, in order to examine dependencies and reveal cross-correlations between the extracted features. Due to the paper's limited length, a letter/identifier was assigned to each of the 26 extracted features for further reference, as presented in Table IV.

Features' correlation matrix, shown in Table V, clearly portrays that the extracted features are relatively uncorrelated, as most of the cross-correlation factors appear values around zero. Therefore, the whole extracted feature set was utilized for the classification task. The investigation of classification efficiency for the presented features, regarding SEs and the respective impact and contribution, was examined for each feature. This formulated a descending order feature vector. The evaluation of the 26 features for each window length was conducted with the utilization of two different WEKA ranking algorithms, namely the "InfoGainAttributeEval" and the "SVMAttributeEval" [16]. The

Table IV
REPRESENTED FEATURE SET

Feature	Letter	Feature	Letter
beatspectrumstd	A	onsetskurtosis	N
eventdensitystd	B	beatspectrumskewness	O
eventdensityskewness	C	pulseclaritygradient	P
onsetsgradient	D	beatspectrumkurtosis	Q
fluctuationkurtosis	E	pulseclaritykurtosis	R
beatspectrumgradient	F	eventdensitykurtosis	S
eventdensitymean	G	beatspectrummean	T
tempomean	H	eventdensitygradient	U
pulseclaritystd	I	pulseclaritymean	V
fluctuationmean	J	onsetsmean	W
fluctuationstd	K	pulseclarityskewness	X
fluctuationskewness	L	onsetsstd	Y
onsetsskewness	M	fluctuationgradient	Z

former evaluates the importance of each attribute separately by estimating the information gain with respect to the class using entropy metrics from Information Theory. The latter represents an SVM technique that examines the efficiency of each feature while assigning each class separately [16]. Table VI exhibits the feature ranking for both algorithms and for each frame length.

E. Classification

Three different training algorithms were employed in the classification task, namely: ANN implementations, Logistic Regression (LR) and the K-Nearest-Neighbor (KNN) technique. Their performance results were compared in order to determine the most efficient classification method. Multilayer perceptrons with one or more hidden layers, based on linear and sigmoid activation functions, are often utilized in semantic analysis problems, providing increased classification rates and achieving the formulation of efficient generalization rules and conclusions [17], [18]. In the experiments of the current work Artificial Neural Systems (ANS) with a network topology that included two sigmoid hidden layers and a linear output layer were implemented.

LR, on the other hand, is a statistical training technique that has been exploited during supervised implementations and which forms a non-linear regression model that relates the classification decision to the output probability result [17], [19]. Finally, KNN is a popular heuristic algorithm for immediate classification results that compares the attributes of each new instance to the attributes of the already classified instances, determining the k nearest similarities and classifying the sample to the respective class of the $k-neighbors$ [20]. Several experiments were carried out before deriving the optimal selection of $k = 5$ neighbors. It also has to be noted that during the experiments performed, additional training models were tested, like linear regressions, decision trees structures and SMO algorithm, but the classification rates were below 70%.

During the training process of the above classification algorithms, the $k-fold$ validation technique was employed. It is an iterative method that divides the whole set of input instances in k subsets and uses $k-1$ subsets for training purposes and the remaining set for testing the developed model. Since the total number of input samples is a prime number (167), the selected number of folds were selected to be $k = 3,\ 8,\ 24$, which are the nearest integer multiples of 168, in order to equally, as possibly, divide the initial sample set. The $8-fold$ and $24-fold$ validations offer the balanced segmentation of the input instances, while $3-fold$ validation favors the generalization potentials of the classification scheme with limited number of iterations. Finally, $167-fold$ validation (Leave-One-Out - LOO technique) was employed, in order to utilize the maximum number of input samples in the process of developing the classification model. The classification performance/recognition rate of each algorithm is defined as the ratio of the number of correctly classified instances to the total number of input instances, deriving from the correspondent confusion matrices. Table VII presents the obtained classification rates for all the frame lengths considered.

IV. RESULTS & DISCUSSION

As Table V shows, there is a relative un-correlation of the features. Thus, the total feature set can be considered as valuable and no feature can be omitted in the classification process. Regarding classification's results, from Table VII it can be seen that considerable variations in the evaluation results have been derived by both methods, while utilizing different temporal windows. The lowest accuracy score obtained was 71.26%. On the other hand, the highest accuracy score was 88.37%, justifying the notion that the rhythm of a sound stimulus can affect the arousal of the listener. This observation also illustrates that the connection of rhythm and arousal is also applicable to SEs and not only music. More specifically, LR exhibits the highest classification accuracy, regardless frame length and the number of folds. Its accuracy ranges from 81.44%, when 1 second frame length and $3-fold$ was used, up to 88.37%, obtained for 1 second frame lengths under the LOO technique. KNN depicts the second highest classification performance and ANS the third.

Considering the used features, it can be observed that in most evaluation results in Table VI two groups are formed, dividing the feature set in clusters of 13 features, for both utilized evaluation algorithms. In particular, for the case that the highest accuracy was observed (1 second frame lengths), the highest 13 features were A, B, D, E, F, I, J, K, L, M, R, S and W. This fact implies that the most informative feature was rhythm's periodicity at the auditory channels (fluctuation) [14]. The second most informative feature was the onsets in the signal, followed by event density, beat spectrum and pulse clarity. Since the test data considered consisted of sounds with different semantic content, the above results also show that the impact of

Table V
FEATURE CORRELATION MATRIX

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	1.0	0.3	-0.1	0.0	0.1	-0.7	0.0	0.1	0.3	0.2	0.1	0.2	0.0	-0.1	0.1	-0.1	-0.3	-0.1	-0.1	-0.6	-0.1	0.1	-0.1	0.0	0.2	0.2
B	0.3	1.0	-0.1	0.1	-0.1	-0.2	0.6	-0.1	0.2	0.0	0.0	-0.1	-0.3	-0.3	0.0	-0.1	-0.1	-0.2	-0.3	-0.1	0.0	0.3	0.2	0.1	0.3	0.0
C	-0.1	-0.1	1.0	-0.1	0.1	0.0	-0.4	-0.2	0.1	-0.2	-0.3	-0.1	0.3	0.2	0.2	0.1	-0.2	0.1	0.5	0.0	-0.1	-0.1	-0.3	0.1	0.1	-0.3
D	0.0	0.1	-0.1	1.0	-0.1	-0.1	0.1	0.1	-0.2	0.0	0.1	0.0	-0.2	-0.1	-0.1	0.1	0.1	0.1	0.0	0.0	0.5	0.0	0.1	0.1	0.0	-0.1
E	0.1	-0.1	0.1	-0.1	1.0	0.0	-0.2	0.0	0.1	-0.1	0.0	0.8	0.2	0.2	0.0	-0.1	0.0	0.1	0.1	0.0	-0.1	-0.3	-0.2	0.0	0.1	-0.1
F	-0.7	-0.2	0.0	-0.1	0.0	1.0	0.0	-0.1	-0.2	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.2	0.1	0.1	0.5	0.0	-0.1	0.0	0.0	-0.1	0.0
G	0.0	0.6	-0.4	0.1	-0.2	0.0	1.0	0.1	-0.1	0.0	0.1	-0.3	-0.6	-0.4	-0.1	0.0	0.0	-0.2	-0.2	0.1	0.1	0.5	0.7	0.0	-0.3	0.1
H	0.1	-0.1	-0.2	0.1	0.0	-0.1	0.1	1.0	-0.1	0.0	0.0	-0.1	0.1	0.0	-0.1	0.1	0.0	0.1	-0.1	0.0	0.0	0.2	0.0	-0.2	-0.2	0.0
I	0.3	0.2	0.1	-0.2	0.1	-0.2	-0.1	-0.1	1.0	0.0	-0.2	0.1	0.2	0.1	0.2	-0.1	-0.1	-0.3	0.1	-0.2	-0.1	0.0	-0.3	-0.1	0.2	0.0
J	0.2	0.0	-0.2	0.0	-0.1	0.0	0.0	0.0	0.0	1.0	0.9	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	-0.4	-0.1	0.0	0.0	-0.1	-0.1	0.7
K	0.1	0.0	-0.3	0.1	0.0	0.1	0.1	0.0	-0.2	0.9	1.0	0.2	-0.1	0.0	-0.1	0.0	0.1	0.1	0.0	-0.2	-0.1	0.1	0.1	-0.1	-0.1	0.7
L	0.2	-0.1	-0.1	0.0	0.8	0.0	-0.3	-0.1	0.1	0.1	0.2	1.0	0.3	0.2	0.0	-0.1	0.0	0.1	0.0	-0.1	-0.2	-0.3	-0.4	0.0	0.2	0.1
M	0.0	-0.3	0.3	-0.2	0.2	0.0	-0.6	0.1	0.2	0.1	-0.1	0.3	1.0	0.6	0.2	-0.1	-0.1	0.0	0.1	-0.2	-0.2	-0.3	-0.9	0.1	-0.1	0.0
N	-0.1	-0.3	0.2	-0.1	0.2	0.1	-0.4	0.0	0.1	0.1	0.0	0.2	0.6	1.0	0.0	0.0	0.1	0.1	0.1	0.0	-0.1	-0.2	-0.5	0.1	-0.4	0.0
O	0.1	0.0	0.2	-0.1	0.0	0.1	-0.1	-0.1	0.2	0.0	-0.1	0.0	0.2	0.0	1.0	-0.1	-0.5	-0.1	0.1	-0.4	-0.1	0.0	-0.2	0.1	0.1	-0.1
P	-0.1	-0.1	0.1	0.1	-0.1	0.0	0.0	0.1	-0.1	0.0	0.0	-0.1	-0.1	0.0	-0.1	1.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1	-0.1	-0.1	0.0
Q	-0.3	-0.1	-0.2	0.1	0.0	0.2	0.0	0.0	-0.1	0.0	0.1	0.0	-0.1	0.1	-0.5	0.1	1.0	0.1	0.1	0.4	0.0	0.0	0.1	-0.2	-0.1	0.1
R	-0.1	-0.2	0.1	0.1	0.1	0.1	-0.2	0.1	-0.3	0.0	0.1	0.1	0.0	0.1	-0.1	0.0	0.1	1.0	0.0	0.1	0.1	-0.2	0.0	0.1	0.0	0.0
S	-0.1	-0.3	0.5	0.0	0.1	0.1	-0.2	-0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.0	1.0	0.1	0.0	-0.2	-0.1	-0.1	0.0	0.0
T	-0.6	-0.1	0.0	0.0	0.0	0.5	0.1	0.0	-0.2	-0.4	-0.2	-0.1	-0.2	0.0	-0.4	0.0	0.4	0.1	0.1	1.0	0.0	0.1	0.2	0.0	-0.1	-0.2
U	-0.1	0.0	-0.1	0.5	-0.1	0.0	0.1	0.0	-0.1	-0.1	-0.1	-0.2	-0.2	-0.1	-0.1	0.1	0.0	0.1	0.0	0.0	1.0	-0.1	0.1	0.1	0.0	0.0
V	0.1	0.3	-0.1	0.0	-0.3	-0.1	0.5	0.2	0.0	0.0	0.1	-0.3	-0.3	-0.2	0.0	0.0	0.0	-0.2	-0.2	0.1	-0.1	1.0	0.4	-0.3	-0.2	0.1
W	-0.1	0.2	-0.3	0.1	-0.2	0.0	0.7	0.0	-0.3	0.0	0.1	-0.4	-0.9	-0.5	-0.2	0.1	0.1	0.0	-0.1	0.2	0.1	0.4	1.0	-0.1	-0.3	0.1
X	0.0	0.1	0.1	0.1	0.0	0.0	0.0	-0.2	-0.1	-0.1	-0.1	0.0	0.1	0.1	0.1	-0.1	-0.2	0.1	-0.1	0.0	0.1	-0.3	-0.1	1.0	0.1	-0.2
Y	0.2	0.3	0.1	0.0	0.1	-0.1	-0.3	-0.2	0.2	-0.1	-0.1	0.2	-0.1	-0.4	0.1	-0.1	-0.1	0.0	0.0	-0.1	0.0	-0.2	-0.3	0.1	1.0	0.0
Z	0.2	0.0	-0.3	-0.1	-0.1	0.0	0.1	0.0	0.0	0.7	0.7	0.1	0.0	0.0	-0.1	0.0	0.1	0.0	0.0	-0.2	0.0	0.1	0.1	-0.2	0.0	1.0

Table VI
FEATURE RANKING. w IS THE FRAME LENGTH IN SECONDS AND ORDER OF APPEARANCE IS THE ORDER OF THE FEATURES (TOP IS FIRST)

w=0.8s		w=1.0s		w=1.2s		w=1.4s		w=1.6s		w=1.8s		w=2.0s	
InfoGain	**SVMA**	**InfoGain**	**SVMA**	**InfoGain**	**SVMA**	**InfoGain**	**SVMA**	**InfoGain**	**SVMA**	**InfoGain**	**SVMA**	**InfoGain**	**SVMA**
D	Z	R	K	M	K	S	J	Q	J	N	K	A	K
F	J	I	M	X	J	F	K	G	K	I	Z	W	J
C	K	D	J	L	F	Z	B	V	L	X	J	Q	E
N	L	A	W	G	W	C	G	H	E	S	E	Y	L
E	I	B	E	P	I	H	E	X	W	M	L	N	G
M	G	S	L	N	A	W	L	M	N	H	Y	P	B
K	B	J	F	F	E	B	I	L	Y	J	W	E	X
Z	F	K	I	K	X	J	F	J	G	K	I	K	W
J	C	M	D	J	L	K	Z	K	V	Z	X	J	Q
G	D	L	R	A	M	L	S	N	Q	Y	N	B	A
B	E	F	B	E	P	I	H	Y	X	W	M	X	N
L	M	W	S	W	N	G	W	E	M	E	H	L	P
I	N	E	A	I	G	E	C	W	H	L	S	G	Y
R	O	H	P	Y	B	O	V	R	S	U	T	U	T
T	P	Z	N	Z	R	D	X	D	A	R	F	O	H
Q	W	Y	C	C	D	R	Y	T	F	V	Q	D	M
A	X	Q	G	U	Q	P	A	I	Z	B	O	Z	I
S	Y	X	U	O	V	T	N	C	O	G	P	R	S
H	V	T	O	S	T	Q	M	P	B	C	A	C	F
W	U	C	V	D	H	Y	U	F	U	Q	D	M	V
O	T	P	Z	B	Z	V	D	S	D	T	R	T	O
P	Q	N	Y	R	C	X	R	A	T	F	V	H	D
V	R	O	H	T	Y	M	O	B	R	A	U	F	U
U	S	V	X	H	O	U	T	U	C	D	G	V	R
X	H	G	T	Q	S	A	Q	Z	P	O	C	I	C
Y	A	U	Q	V	U	N	P	O	I	P	B	S	Z

rhythm in listener's arousal can be independent from the meaning, in language or logic, of sound.

V. CONCLUSIONS AND FUTURE WORK

In this work the impact of the rhythmic characteristics of generalized sound events (SEs) on the listener's arousal is evaluated as a starting point for the exploration of the relation between non-musical sound environments and the conveyed emotions. Towards this aim, SEs were used as a comprehensive form of sound whereas the focus on arousal and rhythm-related characteristics was performed according to their connection in music, as it is already stated in the literature. The IADS sound set was used, as an annotated SEs data base, along with the MIR Toolbox for feature extraction and the WEKA environment for feature evaluation and classification. The feature set considered consisted of 26 features, whereas two ranking algorithms, namely the "InfoGainAttributeEval" and the "SVMAttributeEval" were used.

The results obtained show a relatively high accuracy for the arousal recognition when solely rhythm related features

Table VII
PERFORMANCE RESULTS, w IS THE FRAME LENGTH IN SECONDS

	Algorithm	3-fold	8-fold	24-fold	LOO
$w = 0.8$	ANS	76.65%	77.84%	79.04%	80.84%
	LR	83.83%	85.03%	85.63%	85.63%
	KNN	82.05%	84.43%	85.03%	85.03%
$w = 1.0$	ANS	78.44%	79.04%	79.64%	81.02%
	LR	81.44%	86.23%	87.21%	88.37%
	KNN	83.23%	83.23%	83.23%	84.43%
$w = 1.2$	ANS	76.65%	77.96%	77.96%	80.02%
	LR	85.63%	85.63%	85.63%	85.72%
	KNN	84.43%	85.03%	85.63%	85.63%
$w = 1.4$	ANS	71.26%	76.05%	77.32%	78.25%
	LR	85.63%	85.63%	85.63%	85.63%
	KNN	84.43%	83.83%	83.83%	83.83%
$w = 1.6$	ANS	75.45%	77.84%	80.24%	81.32%
	LR	85.63%	85.63%	85.63%	85.63%
	KNN	85.63%	85.03%	85.63%	85.63%
$w = 1.8$	ANS	77.25%	80.24%	82.63%	84.57%
	LR	85.63%	85.63%	86.45%	87.71%
	KNN	85.03%	85.03%	85.03%	85.03%
$w = 2.0$	ANS	77.84%	76.65%	78.25%	79.32%
	LR	85.63%	85.63%	85.63%	86.04%
	KNN	83.83%	85.03%	85.03%	85.03%

are used. Moreover, signal's fluctuation was identified as the most informative feature regarding arousal recognition. The accuracy of the recognition process can be furthered examined with the utilization of different groups of features, as they are formed from the results of the present work. In addition, a further examination for the arousal recognition regarding SEs should be performed with the usage of additional timbre or energy-related features. Finally, since arousal represents a single component of affective modeling, valence recognition should be also attempted in order to provide an overall assessment of the mechanism that allows SEs to convey emotions.

REFERENCES

[1] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: different channels, same code?," Psychological Bulletin, vol. 129, pp. 770-814, September 2003.

[2] L. Lie, et al., "Automatic mood detection and tracking of music audio signals," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, pp. 5-18, January 2006.

[3] K. Drossos, et al., "Affective Acoustic Ecology: Towards Emotionally Enhanced Sound Events," presented at the Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound, Corfu, Greece, 2012.

[4] K. R. Scherer, "Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them?," Journal of New Music Research, vol. 33, pp. 239 - 251, 2004.

[5] C. Laurier, et al., "Exploring Relationships between Audio Features and Emotion in Music," presented at the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009) Jyväskylä, Finland 2009

[6] P. N. Juslin and D. Vastfjall, "Emotional responses to music: The need to consider underlying mechanisms," Behavioral and Brain Sciences, vol. 31, pp. 559-575, 2008.

[7] M. Marcell, et al., "Identifying, rating, and remembering environmental sound events," Behavior Research Methods, vol. 39, pp. 561-569, 2007.

[8] W. W. Gaver, "What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception," Ecological Psychology, vol. 5, pp. 1-29, 1993/03/01 1993.

[9] M. M. Bradley and P. J. Lang, "The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective Ratings of Sounds and Instruction Manual," NIMH Center for the Study of Emotion and Attention, Gainesville, Fl, Technical report B-3, 2007.

[10] B. Schuller, et al., "Automatic recognition of emotion evoked by general sound events," in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 2012, pp. 341-344.

[11] T. Li and M. Ogihara, "Detecting emotion in music," presented at the Proceedings of the International Symposium on Music Information Retrieval, 2003.

[12] C. Wai Ling and L. Guojun, "Music emotion annotation by machine learning," in Multimedia Signal Processing, 2008 IEEE 10th Workshop on, 2008, pp. 580-585.

[13] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," Speech and Audio Processing, IEEE Transactions on, vol. 10, pp. 293-302, Jul. 2002.

[14] O. Lartillot, et al., "A Matlab Toolbox for Music Information Retrieval," in Data Analysis, Machine Learning and Applications, C. Preisach, et al., Eds., ed: Springer Berlin Heidelberg, 2008, pp. 261-268.

[15] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," Journal of Behavior Therapy and Experimental Psychiatry, vol. 25, pp. 49-59, 1994.

[16] M. Hall, et al., "The WEKA data mining software: an update," SIGKDD Explor. Newsl., vol. 11, pp. 10-18, 2009.

[17] R. Kotsakis, G. Kalliris, C. Dimoulas, "Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification," Speech Communication, vol. 54, no. 6, pp. 743-762, 2012.

[18] C. Dimoulas, G. Papanikolaou, V. Petridis, "Pattern Classification and Audiovisual Content Management techniques using Hybrid Expert Systems: a video-assisted Bioacoustics Application in Abdominal Sounds Pattern Analysis," Expert Systems with Applications, vol. 38, no. 10, pp. 13082-13093, 2011.

[19] D. Hosmer & S. Lemeshow, "Applied logistic regression (2nd ed.)," New York: Wiley, 2000.

[20] T. Seidl & H. P. Kriegel, "Optimal multi-step k-nearest neighbor search," in proceeedings of ACM-SIGMOD international conference on management of data, pp. 154-165, 1998.