Evaluating the Impact of Sound Events' Rhythm Characteristics to Listener's Valence

KONSTANTINOS DROSSOS,[†][*] *AES Student Member,* **ANDREAS FLOROS,**[†] *AES Member,* **AND KATIA-LIDA KERMANIDIS**[‡]

[†]*Lab of Audiovisual Signal Processing, Dept. of Audiovisual Arts, Faculty of Music & Audiovisual Arts, Ionian University, Corfu, Greece*

[‡]*Lab of Informatics in Humanistic and Social Sciences, Dept. of Informatics Faculty of Information Science & Informatics, Ionian University, Corfu, Greece*

Sound is a prominent element in the communication of humans with their environment. It can be either in an organized form, e.g., speech or music, or in a non–organized form, i.e., Sound Events (SEs). There are many studies focusing on emotion recognition from music content, but the research on emotion recognition from SEs is a rather new field. This works falls within this field and concludes that the rhythmic characteristics of an SE have a major impact on the listener's arousal. This result verifies the empirical knowledge that the rhythm of sound affects the listener's activation state. Moreover, we also investigate whether the above characteristics also affect the pleasure of the listener. Toward this aim, we have utilized a well known data set of emotionally annotated SEs, extracted various rhythm-related technical cues, and conducted a series of machine learning experiments. The overall results indicate a relation of the rhythm and listener's valence with accuracy results reaching up to 63%.

1 INTRODUCTION

Humans communicate with their environment through various sense channels with the most prominent being the visual and auditory ones. The latter does not require an obstacle–free path between the source and the receiver. We also keep receiving information through the auditory channel even when we are asleep. Stimuli transmitted over this channel can be either organized in music or speech forms, or not [6]. Sounds that do not demonstrate organized patterns are usually termed as Sound Events (SEs) or general sounds. These represent the majority of audio stimuli received by humans [7].

SEs construct our acoustic environment [8]. They emanate from all surrounding sound sources and communicate various attributes and parameters of the sound source itself, for example its spatial position in respect to the receiver and the nature of the particular sound producing mechanism. Such information helps listeners to perceive close environs and particularly their relation to the sound source. Consequently, it has a direct effect on their actions. This source–listener relation is extensively employed and exploited in various applications, including virtual and augmented environments, movies, video games, and auditory interfaces [7].

Being a significant component of the above perceptual process, emotion communication and elicitation through the auditory channel is not a recent concept. The notion that music conveys emotion is likely to be well spread and employed in the early usage of music, i.e., the enhancement of emotions elicited from speech [9]. Emotion recognition from musical data is a recent and emerging research field with applications mainly in content-based categorization and retrieval [10, 11]. Different emotional models are employed in the aforementioned process with the most prominent one being the arousal-valence (AV) space [6]. However, according to the authors' current knowledge, most works employing the AV space consider emotions as areas defined within this space. They do not address the affective state recognition problem, i.e., the recognition of arousal and valence as distinct dimensions. Instead, they cluster AV values under verbal descriptions of emotions without a direct and quantitative relation and/or mapping process. Although this approach yields accuracy that may reach up to 90% [12, 10], it is restricted by the employed verbal descriptions of emotions and potentially reduces the added value of dimensional emotion models. For example, if four different emotions are used in the recognition process, then the obtained results account for these emotions alone and not different combinations of the arousal and valence states that correspond to these emotions. Moreover, a similar work with synonym or similar verbal descriptions cannot be

[*] Corresponding author kdrosos@ionio.gr

directly and freely compared to the aforementioned results. This is a well known problem in research employing verbal descriptions of emotions [13].

Since SEs are also elements of the audio communication channel, the research question of whether the general sounds evoke emotions is intuitively risen and studied in recent publications. Although such work is scarce and loosely connected [14] there are reports of emotion recognition results from SEs that reach up to 89% accuracy regarding the arousal of the listener [6]. Additionally, a recent work published by the authors proposes an extension of the Acoustic Ecology term in order to include the affective reactions of the listener [7]. Nevertheless and according to the authors' best of knowledge, there are no published investigations that regard solely the recognition of valence from SEs. Combined with arousal recognition, valence identification can offer a complete emotional recognition from SEs and thus possibly allow affective-driven synthesis of generalized SEs. Such a synthesis approach can be utilized in various applications (including sound design for gaming environments) in order to enhance the user experience by selectively eliciting specific affective states to the human listener.

The work at hand represents an extension of previously published research that originally outlined the potential impact of SEs rhythm related characteristics on specific components of the human listener affective state [6]. In particular, this consideration was based on the fact that rhythm in music is well and impulsively connected to the arousal of the listener. As a consequence, the research scope of the aforementioned work was limited to arousal only, concluding a similar arousal and rhythmic characteristics relation trend in the case of generalized SEs. On the other hand, in the current work, we aim to additionally address the question whether rhythm also affects the valence affective dimension when the human listener is exposed to generalized SEs. Obviously, this complementary research consideration is necessary toward a complete and systematic exploration of the potential relation between the SEs' rhythmic characteristics and the elicited listener's emotions. Under the above perspective, we hereby focus exclusively on the valence dimension, probe the connection of rhythm with listener's valence, and deduce results on whether rhythm characteristics of general sounds affect (or not) the elicited *pleasure* of the listener. Clearly, a future combination of the arousal and valence–related results is expected to significantly enhance the state-of-the-art on emotion recognition from SEs by providing the necessary foundations on the relation of rhythm with emotion.

Toward the above aim a series of machine learning tests was conducted, considering as ground truth the affective annotated International Affective Digitized Sounds (IADS) [1] data set. The rhythm characteristics are solely acoustic cues extracted from raw digital audio signals. Additionally, taking into account the previously reported interdependence between listener arousal and valence that defines that it is rather uncommon for a listener not to like (low valence) a sound and not to feel aroused (low arousal) by it [1], we also performed a set of experiments having as a feature the arousal ratings of SEs in the employed sound corpus. The obtained results on one hand indicate a clear enhancement of accuracy ratings when arousal is also employed and, on the other hand demonstrate that rhythm characteristics can indeed have an effect on valence (although not a major one).

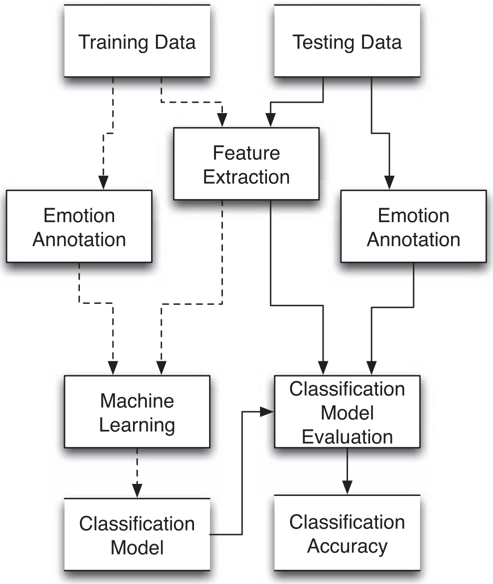


Fig. 1. An illustration of a typical emotion recognition from audio data process.

The rest of the paper is organized as follows. In Sec. 2 a brief overview of the existing literature on emotion recognition from SEs is included. Sec. 3 describes the employed experimental procedure, followed by the obtained results that are analytically presented in Sec. 4. A detailed discussion of the results is performed in Sec. 5 and the paper is concluded in Sec. 6.

2 EMOTION RECOGNITION FROM GENERAL SOUND

Emotion recognition from SEs is mainly a machine learning task [15]: from a ground truth data corpus a set of characteristics is initially extracted. Then, based on a selected affective model, an emotional annotation task is performed. Finally, a recognition model is built. This process is illustrated in Fig. 1. Clearly, the key–requirements for recognizing emotions from the SEs process is the availability of affectively annotated SEs data sets, the definition of specific emotional classes employed (which are directly connected with the emotional model employed), and the selection of the appropriate features that will be extracted from the sound corpus.

Next, we will present a brief overview of the existing and available data sets with emotionally annotated SEs, the affective models that are used mainly in sound and music emotion recognition, and summarize the outcomes of published works in the field of emotion recognition from SEs.

2.1 Emotionally Annotated SE Data Sets

In the literature one can find only three freely available data sets with emotionally annotated SEs. In brief, two of them consist of single channel (monophonic) audio data, while the third one incorporates binaurally processed SEs.

More specifically, the IADS data set [1] represents the first available data set with affective annotated general sounds. It consists of 167 SEs having the same time–length (6 seconds), with various sampling frequencies ranging from 8 to 44.1 kHz. The contained audio waveforms have diverse semantic content and are created using a variety of sound sources. For each of the incorporated 167 SEs, three affective state annotations are provided (for the arousal, valence, and dominance affective components respectively). Each of them is represented by its mean and the standard deviation value. Every SE in the IADS data set has been annotated by approximately 100 human subjects [1]. IADS is the only available SE data set that provides such vast annotations per SE.

Based on these subjective annotations, it was reported that low valence values are not likely to be combined with low arousal. This is due to the fact that one is unlikely to feel unpleasant (low valence) and at the same time tranquil (low arousal) when this condition is elicited by a sound stimulus [1]. The above fact can be depicted from the scatter plot of the arousal and valence values of the IADS sound corpus shown in Fig. 2.

Recently a new data set was also presented consisting of 360 monophonic SEs, the Emotional Sound Database (ESD) [16]. All audio data were retrieved by the online database of FindAllSounds[1], having semantic content from the categories of animals, musical instruments, nature, noisemakers, people, sports, tools, and vehicles. The data were annotated by four persons and the annotations were processed using the evaluation weighter estimator [17], in order to increase their robustness due to the small amount of annotators employed. The utilized affective states were arousal and valence. SEs in the ESD data set have variable time–lengths and sampling frequencies.

In [18] the authors presented a binaural data set with emotionally annotated SEs. This is the only data set with binaural emotionally annotated SEs and was based on the IADS data set, from which 32 monophonic general sounds were included. Each one was binaural rendered and spatially positioned in five different angles on the horizontal plane (0, 45, 90, 135, and 180 degrees). This process resulted in a data set with 160 SEs (i.e., 32 SEs for each of the aforementioned five angles). Due to the immediate relation to the IADS data set, all SEs in the BEADS sound corpus have the same technical characteristics as the ones in the IADS data set.

2.2 Models of Emotions

Affective models can be classified into two abstract categories: (a) discrete and (b) continuous [19]. The former includes models that use discrete labels/values for each emotion they incorporate. The most common way for implementing a discrete model is the utilization of verbal descriptions for emotions. Such models are the basic emotions set and the list of adjectives [2].

The basic emotions set was based on the concept that there is a set of emotions that can be accounted as the primary emotions from which all other emotions can be constructed. These emotions are "Fear," "Anger," "Happiness," and "Sadness." The basic hypothesis of this model was thoroughly questioned by [20], and it seems that this model is not frequently employed in audio emotion recognition works but is usually utilized by neurological works on emotion recognition, due to the immediate relation of parts and areas of the brain with emotion in this model [21, 22]. The list of adjectives, on the other hand, employs a number of synonym sets in order to describe various emotions. It was proposed by [2] where the number of synonym sets was eight. There are variations of this model in which a different number of sets is utilized, e.g., 13 [3, 23]. Fig. 3 is an illustration of the list of adjectives model with eight groups of synonyms.

Discrete emotions models can be regarded as the first employed ones, since the basic emotions model can be traced back from the Darwin era [24, 25]. However, due to the utilization of verbal descriptions of emotions, there is a tendency for inconsistency between different published works. For example, there are a few words that can be used interchangeably for declaring that someone is feeling happy but they can also mean a different emotional state like "Cheerful/Happiness," "Joy/Enjoyment." This fact introduces an inconsistency and noise when someone wants to compare results from different research [13].

The second category of emotions models (i.e., the dimensional ones) can be considered as an answer to this hurdle. In this category the models do not represent emotions as words. Instead, they utilize a set of components of emotions, i.e., emotional states, and a geometrical space whose number of dimensions is equal to the amount of affective components used. The resultant of the emotional states in this space is the targeted emotion and can be later described by a word corresponding to the appropriate emotion.

Such models usually employ a basic set of two components: (a) the activation, termed as arousal and (b) the pleasure, termed as valence. Although there are more dimensions proposed in the literature, such as the dominance, the 2D space with arousal and valence are preferred and are considered to be the most universal [26]. Fig. 4 is an illustration of a 2D space with arousal and valence. Modeled emotions correspond to particular areas of this space. Clearly, the correspondence with the verbal descriptions of emotions is performed by clustering the values in the aforementioned space. This approach is widely used in work focused on audio/music emotion recognition [7].

2.3 Emotion Recognition from Sound Events

There are few published works concerned with emotion recognition from general sounds. Lately, the Affective Acoustic Ecology concept was introduced as an expansion

[1] www.findallsounds.com

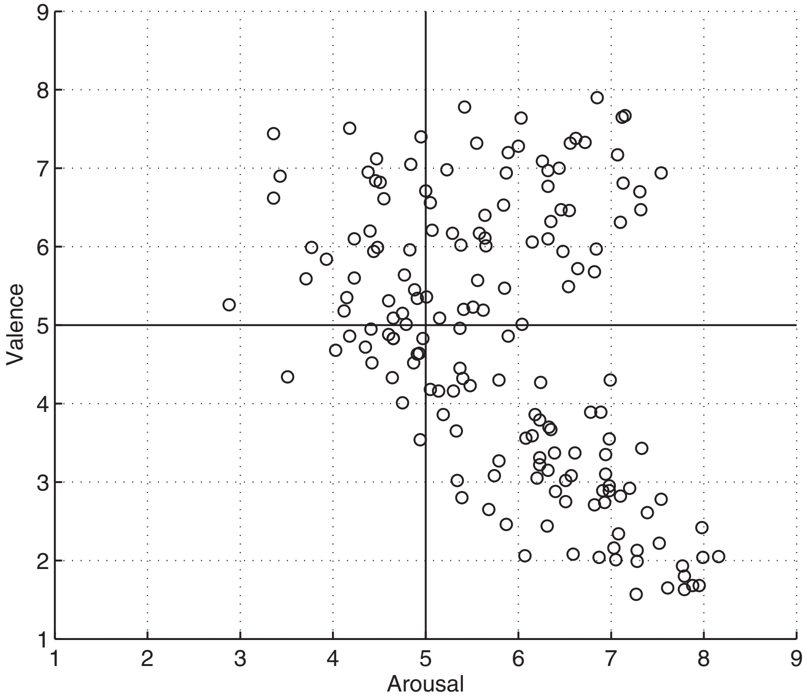


Fig. 2. Scatter plot of IADS arousal and valence values, after [1].

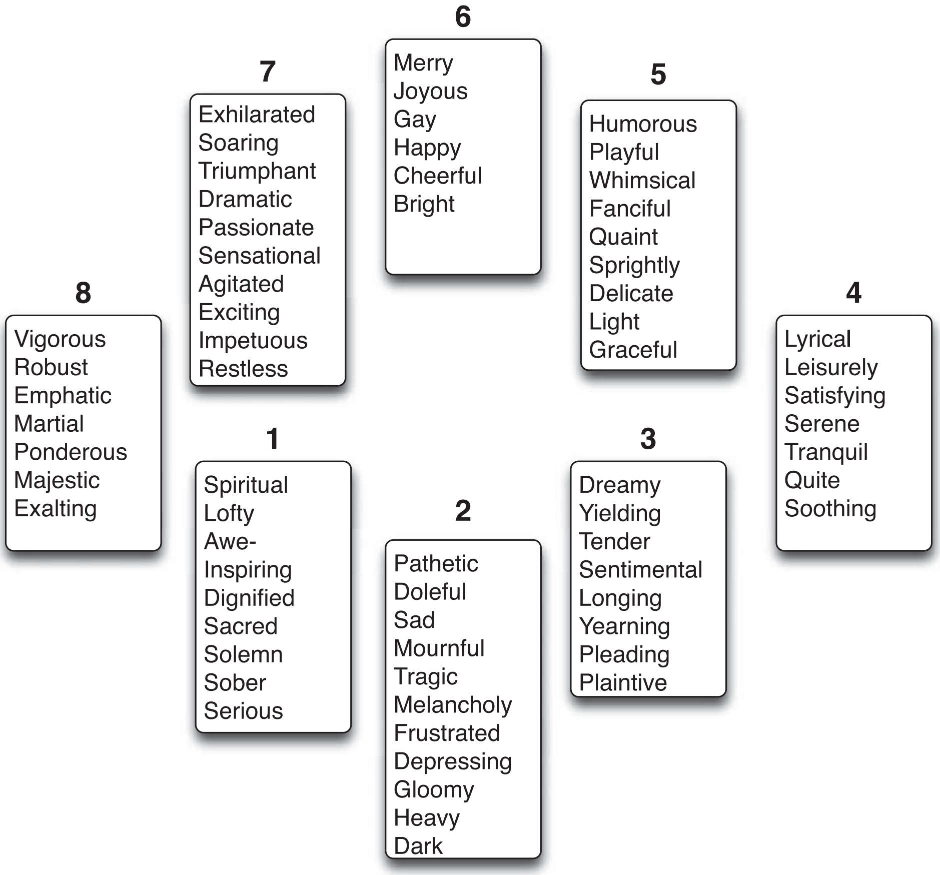


Fig. 3. The list of adjectives with eight groups of synonyms words, after [2, 3].

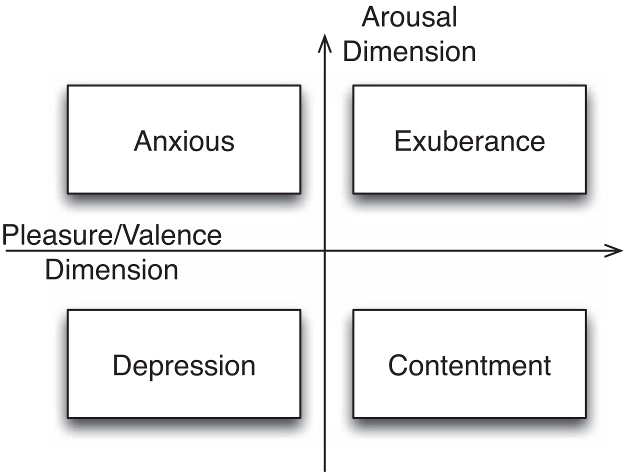


Fig. 4. The AV plane with four clustered emotions, according to [4, 5].

of the legacy Acoustic Ecology, which aims to provide the foundations for efficiently describing the inherent relation between sound and human listener emotions. Its key element is the SE itself, defined as a complex data structure consisting of the sound waveform, and multiple other attributes related to the SE, such as the spatial position of the sound source and the SE duration among others [7].

Additionally, in [6] an investigation was performed on the relation between rhythm characteristics of SEs and the induced arousal with accuracy results up to 89%. The IADS data set was utilized and several low level acoustic cues were selected for calculating their variation as the SEs were reproduced. Window lengths for signal segmentation ranged from 0.8 to 2 seconds with 0.2% overlap. Machine learning algorithms included Artificial Neural Networks and Logistic Regression. Furthermore, in [16] emotion recognition from general sounds yielded to correlation coefficient values of approximately 0.6% for arousal and 0.5 for valence. The data set employed for this research is the ESD described above, while several low level audio features were utilized. The machine learning algorithms used were the well-known REPTree [27] and Random Forests [28].

From the above literature overview it is evident that the potential relation of SEs rhythmic characteristics and the elicited listener's valence is not yet considered. An investigation toward this relation would provide concluding findings for the overall connection of rhythm and emotion, regarding SEs. In this paper we aim to provide a basis for exploring and outlining this relation by extending our previous work that correlates rhythm–related characteristics and arousal [6]. We particularly try to address the question whether the rhythm characteristics of an SE have an impact on the listener's pleasure. We conform with the majority of the published works related to emotion recognition from audio in general by employing the AV space and popular machine learning algorithms. To this aim, we followed the same procedure as in [6] and focused on valence recognition instead of arousal. The results of the present work may allow and support any further analysis of arousal and valence recognition in order to provide the fundamental framework for delivering efficient SE synthesis engines driven by target emotions.

3 EXPERIMENTAL PROCEDURE

In order to investigate the relation between the rhythm characteristics of an SE and the listener's valence we conducted a series of machine learning experiments. The experimental process followed can be divided in three basic tasks: (a) data pre–processing, (b) sound technical features extraction, and (c) machine learning tests. During the pre–processing stage, the utilized data sets were constructed from the original IADS data set. Consequently, the feature extraction process was performed, producing several rhythm-related features for each of the constructed data sets. Then, for the training process the extracted features along with the emotional annotations from the IADS data set were passed as inputs to the used machine learning algorithms and several models were produced. In the evaluation/testing sequence, these models with the extracted features and the emotional annotations were utilized and the classification results for each model were provided. A graphical representation of the followed experimental process is illustrated in Fig. 5. In the following subsections each of these steps are analytically presented.

3.1 IADS Data Set Pre-Processing

As it was mentioned previously, the IADS data set was used as the original sound corpus. It consists of 167 emotionally annotated SEs, with different dB Full–Scale (dBFS) values. The semantic content of the data set varies from everyday human activities, e.g., walking, whistling, chatting, etc., to emergency human reactions, animal sounds, mechanical sounds, and explosions. The emotional annotation of the data set was performed using the Self Assessment Manikin (SAM) method including also intermediate states between the original SAM's figures. The available annotation values, including these intermediate states, are 9, with 1 representing the lowest and 9 the highest rating. Figs. 6 and 7 illustrate the original SAM figures (i.e., without the intermediate states) for arousal and valence respectively [1].

In order to utilize a homogeneous data corpus in terms of energy, we normalized all SE waveforms to 0 dBFS. This choice was made in order to eliminate potential dependencies between valence and the SE energy/loudness. Moreover, the sampling frequencies of the audio data were not altered in order not to introduce artificially produced samples in the data set. Furthermore, in order to investigate the fluctuation in time of the rhythm characteristics, each SE in the utilized data set was segmented and all technical characteristics were extracted from the resulting frames of each SE. Seven different time lengths for frames, $t_{wl}[i]$, were employed with values $t_{wl}[i] = [0.8, 2]$ seconds, $i \in [1, 7]$, and with and increment step of 0.2 seconds. This process led to seven different data sets, one for each frames'

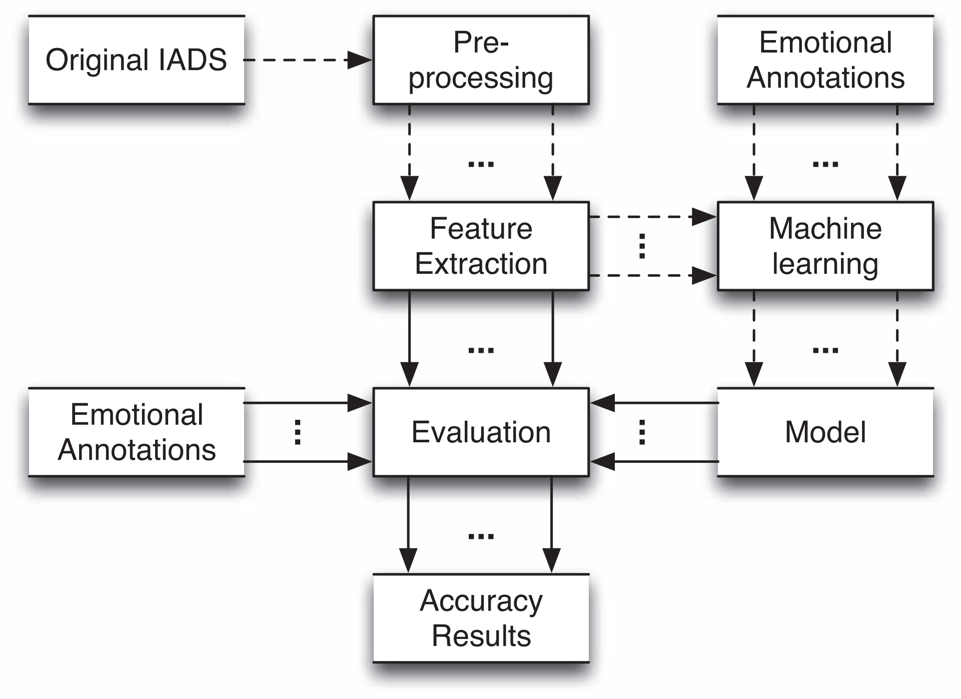


Fig. 5. A graphical representation of the followed experimental procedure. With dashed line is the training stage while with solid line is the testing phase.

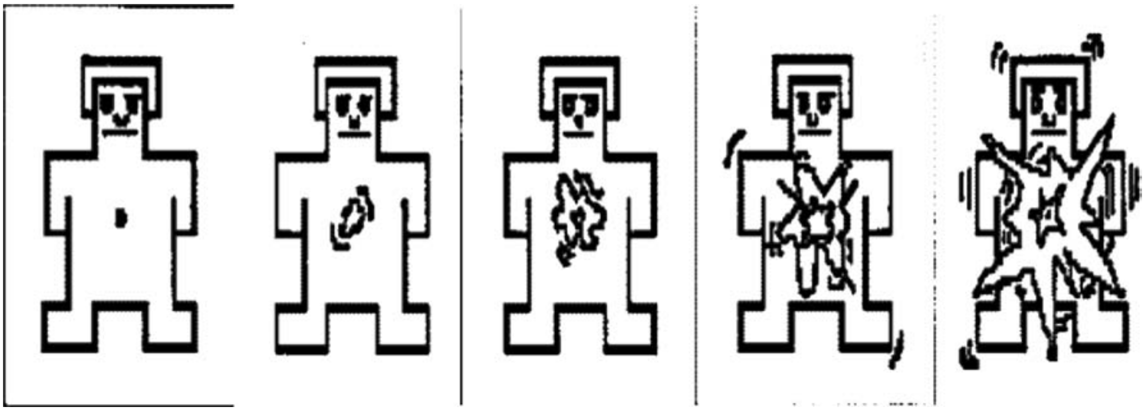


Fig. 6. SAM figures used for annotation of arousal in the IADS data set.

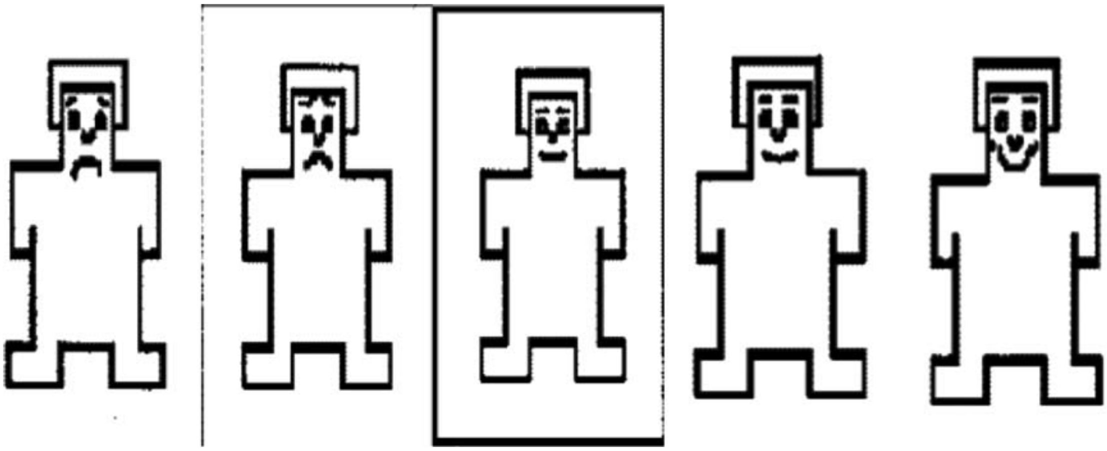


Fig. 7. SAM figures used for annotation of valence in the IADS data set.

Table 1. $S_{i,x'}$ with their index *i* and their corresponding frames' time length

Dataset index (*i*)	Frames' time length (secs)
1	0.8
2	1.0
3	1.2
4	1.4
5	1.6
6	1.8
7	2.0

Table 2. Extracted features

Feature index (i')	Extracted Feature
1	Beat spectrum
2	Onsets
3	Tempo
4	Fluctuation
5	Event density
6	Pulse clarity

time length. More specifically, if the initial IADS data set is defined as $S_x[n]$, $x \in [1,\ 167]$ and *n* the samples for each SE, then the normalization process was:

$$S_{x'}[n] = f(S_x[n]),\ x \in [1,\ 167] \tag{1}$$

where *f*(*y*) is the normalization function applied for each $S_x[n]$, $S_{x'}[n]$ is the normalized version of the original IADS data set, and *n* is the samples of each SE. Consequently, each $S_{x'}[n]$ was segmented in frames with time length equal to $t_{wl}[i]$, $i \in [1,\ 7]$, and overlap of 20%. Thus, for each $S_{x'}[n]$ the set of segments $S_{i,x'}$ was created, where *i* is the index of the segments' length according to $t_{wl}[i]$ and x' is the index of the signal $S_{x'}[n]$. At each segment in $S_{i,x'}$ the hamming window function was applied. The resulting sets of segments with their index (*i*) and their corresponding frames' time lengths are shown in Table 1.

For the final data set, valence and arousal values for each SE (denoted here as $E[x']$—see below), were clustered in two classes, C_1 and C_2: one for denoting valence/arousal values above the mean state (C_2), and one for denoting values below the mean state (C_1). The value of 5 was considered as the mean state, taking into account the method employed for the emotional annotation of the IADS data set [1]. The final SE data set components $S'_{i,x'}$ were constructed from the $S_{i,x'}$ and the corresponding IADS emotional annotations $E[x']$ as:

$$S'_{i,x'} = \{S_{i,x'},\ E[x']\},\ i \in [1,\ 7],\ x' \in [1,\ 167] \tag{2}$$

where $E[x'] = \{A[x'],\ V[x']\}$ and $A[x']$ and $V[x']$ are the arousal and valence class values respectively for the x'–index SE. The distribution of SEs in arousal and valence classes is summarized in Fig. 8.

3.2 Feature Extraction

For each segment of $S'_{i,x'}$ a set of features, $F[i']$ was extracted. These features are listed in Table 2 and are all considered to be rhythm–related ones [6].

Table 3. Statistical measures employed with their corresponding indices

Measure index (*z*)	Statistical Measure
1	Mean (M)
2	Standard deviation (STD)
3	Gradient mean (GM)
4	Gradient STD (GSTD)
5	Skewness (SKW)
6	Kurtosis (KURT)

Table 4. The features in each $R_{i,x'}$

Num.	**i′**	**z**	Num.	**i′**	**z**
1	1	1	18	4	5
2	1	2	19	4	6
3	1	3	20	5	1
4	1	4	21	5	2
5	1	5	22	5	3
6	1	6	23	5	4
7	2	1	24	5	5
8	2	2	25	5	6
9	2	3	26	6	1
10	2	4	27	6	2
11	2	5	28	6	3
12	2	6	29	6	4
13	3	1	30	6	5
14	4	1	31	6	6
15	4	2	32	Arousal	
16	4	3	33	Valence	
17	4	4			

All $F_{i'}$ were extracted using the MIR Tollbox [29]. Thus, for each segment in $S'_{i,x'}$ a curve showing the fluctuation of each $F[i']$ could be drawn if the values for each $F[i']$ were plotted against the segments in $S'_{i,x'}$. In order to (a) describe this curve as closely as possible, (b) not to loose the information about the fluctuation of each $F[i']$ with the advance of segments, and (c) obtain one value and, consequently, reduce the amount of dimensions for the latter machine learning process, for each $F[i']$, a set of statistical measures was used. These measures, $M[z]$, are shown in Table 3.

In particular, for each $S'_{i,x'}$ the set of features $F[i']$ was extracted from $S_{i,x'}$. This process yields a set of results according to $t_{wl}[i]$, where the latter affected the amount of segments for each x'. Thus, for each *i*, a different set of resulting values was obtained for each $F[i']$, denoted here as $R_{i,x',i'}$, $i \in [1,\ 7]$, $x' \in [1,\ 167]$ and $i' \in [1,\ 6]$. For each i' and for each $F[i']$, and where applicable, the statistical measure $M[z]$ was calculated resulting in the final data set of $R_{i,x'}$. Finally, each of $R_{i,x'}$ contained the features shown in Table 4.

From the $R_{i,x'}$ two sets were constructed in order to be used in the following machine learning process. One without the feature with number 32 in $R_{i,x'}$ and one with it, named henceforth $R'_{i,x'}$ and $R''_{i,x'}$ respectively. The latter set was utilized in order to employ the connection between valence and arousal values in SE perception, as presented in the IADS report, where it was stated that a listener is more likely to experience high activation, i.e., high arousal,

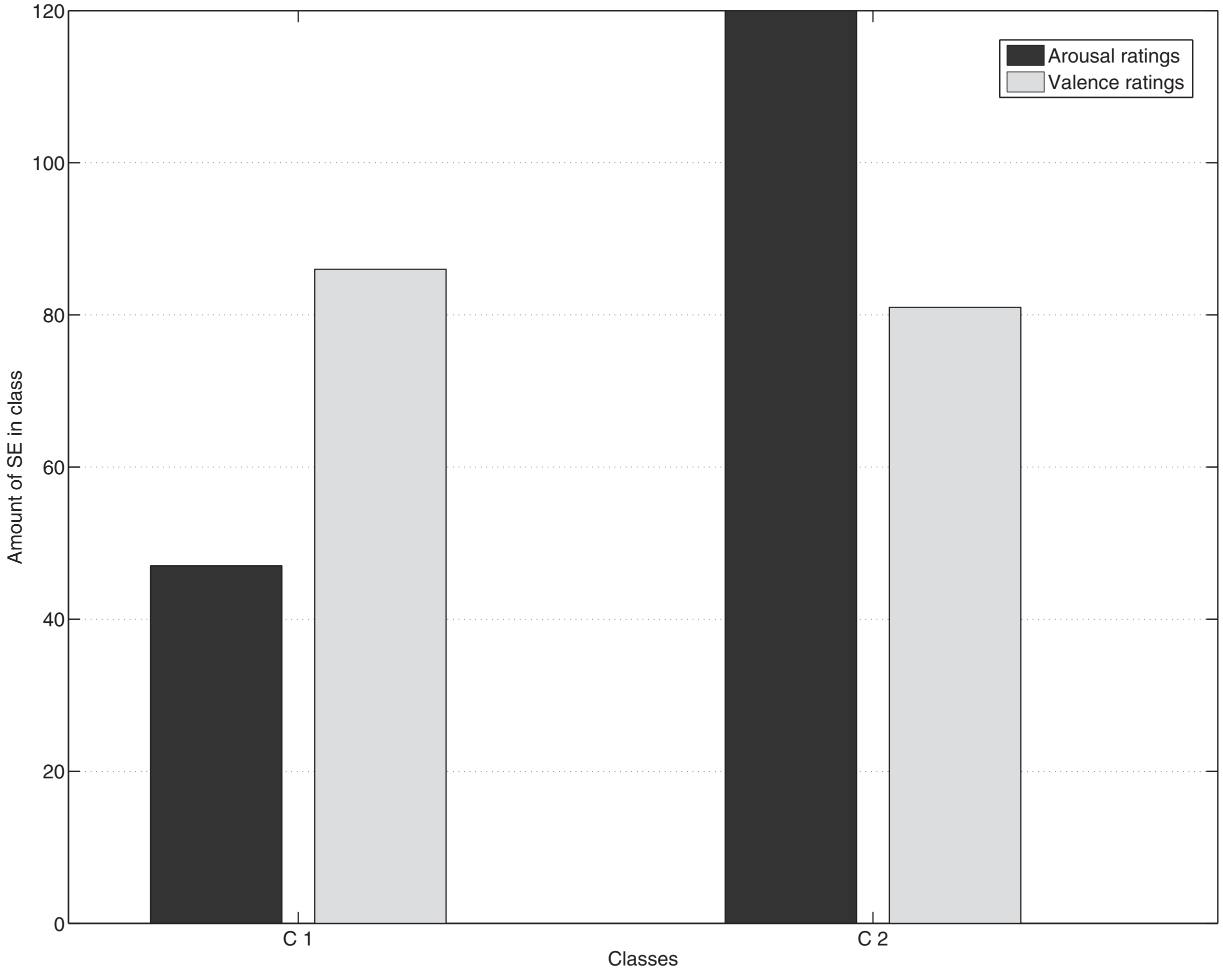


Fig. 8. The distribution of the SEs in arousal and valence classes.

when listening to something that he does not like, i.e., low valence [1]. Thus, this investigation allows for the inclusion of previous findings from arousal recognition from rhythm related characteristics [6] in order to enhance valence recognition.

3.3 Machine Learning Tests

During the machine learning tests, a number of well–known classification algorithms was employed for valence prediction. For comparative purposes, and motivated by the lack of related work on predicting valence provoked by this specific type of audio input, these algorithms were chosen to cover a wide range of learning approaches, i.e., statistical, tree-based, and geometrical. All of the tests were conducted using the functionality of the WEKA environment.[2]

In particular, the employed algorithms acted as standalone classifiers that varied from simple decision stumps, fast decision trees (RepTree), the pruned $C4.5$ decision tree induction algorithm [30] with a confidence factor of 0.25 and logistic regression, to more sophisticated ones like Support Vector Machines [31] (with a first degree polynomial kernel function and the Sequential Minimal Optimization (SMO) algorithm [32] for training the classifier), as well as ensemble learning schemata. The latter include Bagging (with a 100% bag size and the SMO chosen as base classifier—set with the aforementioned parameters) and AdaBoost [33], again with SMO as the base classifier. Both $R'_{i,x'}$ and $R''_{i,x'}$ sets were used with the employed algorithms and the models built were validated using 10-fold stratified cross validation.

4 RESULTS

This section presents the results obtained by the previously presented machine learning tests. For each algorithm employed, the recall and precision values per class are listed along with the accuracy and root mean square error. All values are for both $R'_{i,x'}$ and $R''_{i,x'}$ data sets. Due to the nature of the obtained results, the Tables included in this section were chosen to portray the obtained values in order also to provide the full results list. Thus, Tables 5 and 6 show the recall and precision values per class, for all the utilized algorithms and for the data sets $R'_{i,x'}$ and $R''_{i,x'}$ respectively. Highlighted columns in these Tables are the ones with the highest value of precision and recall for each i and each class. In addition, the same information is illustrated in Fig. 9a to 9d for reader's convenience and faster comparison of values.

[2] www.cs.waikato.ac.nz/ml/weka

Table 5 . Recall (R) and precision (P) values per class(C_1 and C_2) for the $R'_{i,x'}$ data set and for all employed algorithms. Grey cells are the ones with the highest values per class and "D. Stump" stands for "Decision Stump."

	Ada Boost		Bagging		D. Stump		**C4.5**		SVM		RepTree		SMO	
	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$
i = 1 (window length = 0.8 seconds)														
R	0.558	0.568	0.616	0.556	0.512	0.556	0.512	0.494	0.535	0.519	0.709	0.284	0.523	0.593
P	0.578	0.548	0.596	0.577	0.550	0.517	0.518	0.488	0.541	0.512	0.513	0.479	0.577	0.539
i = 2 (window length = 1.0 seconds)														
R	0.640	0.543	0.616	0.469	0.593	0.420	0.802	0.235	0.628	0.556	0.791	0.321	0.640	0.519
P	0.598	0.587	0.552	0.535	0.520	0.493	0.527	0.528	0.600	0.584	0.553	0.591	0.585	0.575
i = 3 (window length = 1.2 seconds)														
R	0.628	0.457	0.628	0.457	0.605	0.494	0.488	0.531	0.616	0.506	0.651	0.395	0.640	0.469
P	0.551	0.536	0.551	0.536	0.559	0.541	0.525	0.494	0.570	0.554	0.533	0.516	0.561	0.551
i = 4 (window length = 1.4 seconds)														
R	0.581	0.494	0.547	0.519	0.512	0.444	0.674	0.395	0.535	0.506	0.709	0.358	0.523	0.519
P	0.549	0.526	0.547	0.519	0.494	0.462	0.542	0.533	0.535	0.506	0.540	0.537	0.536	0.506
i = 5 (window length = 1.6 seconds)														
R	0.593	0.617	0.616	0.556	0.570	0.741	0.547	0.654	0.581	0.531	0.686	0.556	0.605	0.580
P	0.622	0.588	0.596	0.577	0.700	0.619	0.627	0.576	0.568	0.544	0.621	0.625	0.605	0.580
i = 6 (window length = 1.8 seconds)														
R	0.628	0.617	0.558	0.531	0.488	0.802	0.430	0.667	0.593	0.556	0.535	0.691	0.628	0.617
P	0.635	0.610	0.558	0.531	0.724	0.596	0.578	0.524	0.586	0.563	0.648	0.583	0.635	0.610
i = 7 (window length = 2.0 seconds)														
R	0.593	0.543	0.616	0.531	0.279	0.765	0.523	0.519	0.581	0.568	0.640	0.444	0.558	0.568
P	0.580	0.557	0.582	0.566	0.558	0.500	0.536	0.506	0.588	0.561	0.550	0.537	0.578	0.548

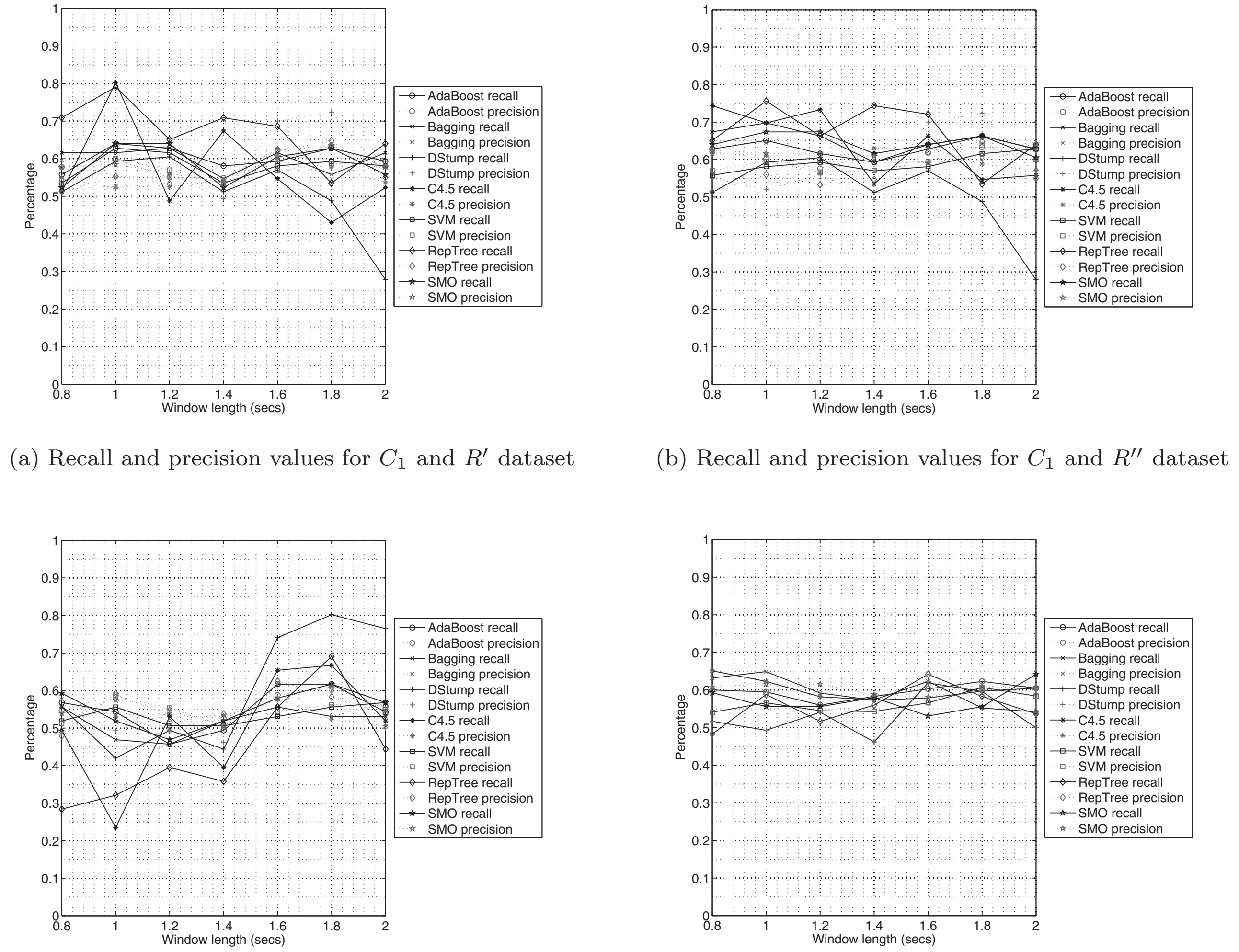
Additionally, Figs. 10a to 10b illustrate the accuracy and root mean square error for all algorithms and data sets utilized.

5 DISCUSSION

The following discussion is organized based on two discrete conceptual layers. The first one is obvious and corresponds to the global particular aim of this work, i.e., the exploration of the potential relation between the rhythmic characteristics of SEs on the listener's valence. The consideration of the second discussion layer originates from the need to validate the results obtained using different machine learning strategies. This approach is widely accepted between the members of the audio emotion recognition community, aiming to provide evaluation strategy indepen-

Table 6 . Recall (R) and precision (P) values per class(C_1 and C_2) for the $R''_{i,x'}$ data set and for all employed algorithms. Grey cells are the ones with the highest values per class and "D. Stump" stands for "Decision Stump."

	Ada Boost		Bagging		D. Stump		**C4.5**		SVM		RepTree		SMO	
	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1$	$\mathbf{C}_2$
i = 1 (window length = 0.8 seconds)														
R	0.628	0.593	0.674	0.593	0.512	0.556	0.744	0.506	0.558	0.556	0.651	0.346	0.640	0.593
P	0.621	0.600	0.637	0.632	0.550	0.517	0.615	0.651	0.571	0.542	0.514	0.483	0.6250	0.608
i = 2 (window length = 1.0 seconds)														
R	0.651	0.543	0.698	0.593	0.593	0.420	0.698	0.531	0.581	0.580	0.756	0.370	0.674	0.556
P	0.602	0.595	0.645	0.649	0.520	0.493	0.612	0.623	0.595	0.566	0.560	0.588	0.617	0.616
i = 3 (window length = 1.2 seconds)														
R	0.616	0.519	0.663	0.519	0.605	0.494	0.733	0.395	0.593	0.519	0.663	0.383	0.674	0.556
P	0.576	0.560	0.594	0.592	0.559	0.541	0.563	0.582	0.567	0.545	0.533	0.517	0.617	0.616
i = 4 (window length = 1.4 seconds)														
R	0.593	0.605	0.593	0.580	0.512	0.444	0.535	0.667	0.570	0.543	0.744	0.346	0.616	0.580
P	0.614	0.583	0.600	0.573	0.494	0.462	0.630	0.574	0.570	0.543	0.547	0.560	0.609	0.588
i = 5 (window length = 1.6 seconds)														
R	0.640	0.580	0.628	0.543	0.570	0.741	0.663	0.593	0.581	0.580	0.721	0.531	0.640	0.531
P	0.618	0.603	0.593	0.579	0.700	0.619	0.633	0.623	0.595	0.566	0.620	0.642	0.591	0.581
i = 6 (window length = 1.8 seconds)														
R	0.663	0.593	0.663	0.531	0.488	0.802	0.547	0.593	0.616	0.630	0.535	0.691	0.663	0.556
P	0.633	0.623	0.600	0.597	0.724	0.596	0.588	0.552	0.639	0.607	0.648	0.583	0.613	0.608
i = 7 (window length = 2.0 seconds)														
R	0.628	0.605	0.628	0.605	0.279	0.765	0.558	0.556	0.628	0.556	0.640	0.444	0.605	0.642
P	0.628	0.605	0.628	0.605	0.558	0.500	0.571	0.542	0.600	0.584	0.550	0.537	0.642	0.605



(a) Recall and precision values for C_1 and R' dataset

(b) Recall and precision values for C_1 and R'' dataset

(c) Recall and precision values for C_2 and R' dataset

(d) Recall and precision values for C_2 and R'' dataset

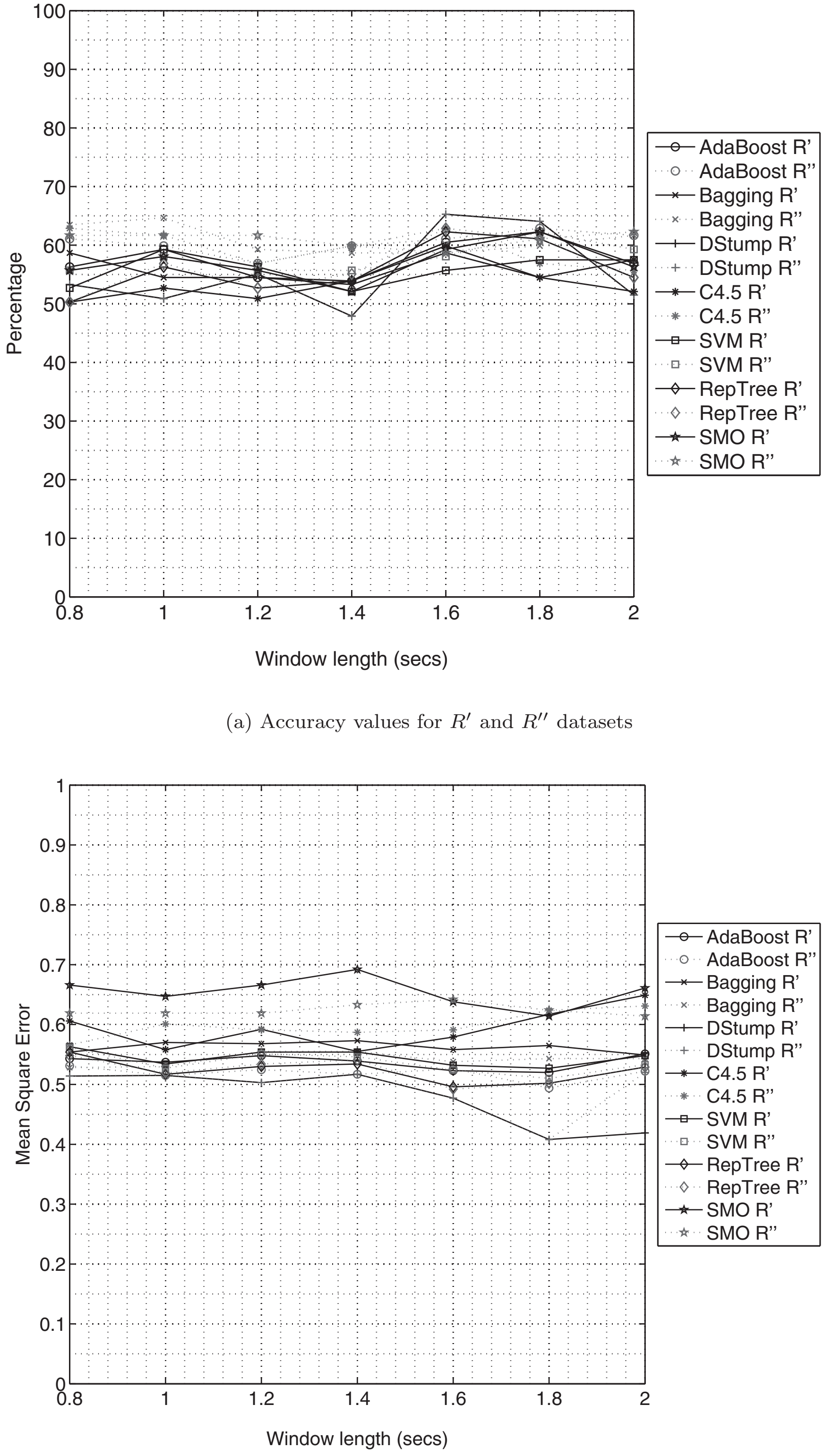
Fig. 9. Recall and precision values for all data sets (R' and R''), classes (C_1 and C_2), and algorithms employed.

dent results. Therefore, the discussion here inevitably includes the assessment of the performance achieved by each employed machine learning algorithm in the recognition procedure. Due to the former's dependency on the latter, the algorithms' capabilities are discussed first.

Looking at the results obtained by the employed classification algorithms, the superiority of the ensemble learning schemata, when compared to the stand-alone classifiers, is evident. The improved results of AdaBoost for the majority of the data sets compared to the most sophisticated single learner, i.e., SMO, are clear in Table 5. Among the single classifiers, SMO and RepTree address the prediction of both valence classes more accurately. The optimal window size proves to be the 1.6 sec window length. This size that is slightly smaller than the one-third of the entire signal duration seems to capture optimally all acoustic features related to the signal. It is also noteworthy that in the majority of data set, algorithm, and window size combinations, high (positive) valence (C_2) is harder to predict than negative valence (C_1). This is in accordance to previous work on music and audio emotion recognition [34].

Moving to Table 6, for the majority of the conducted experiments, the prediction model benefits significantly from the inclusion of arousal in the feature sets, proving the strong interdependence between arousal and valence. Results with AdaBoost exceed 65% for recall and precision for negative valence and 60% for positive valence (63% for precision). The fact that the simple decision stump algorithm achieves even higher results for the 1.6 sec window size is attributed to circumstantial, data-specific characteristics, with very low generalizable prospects. In other words, these results were most likely achieved by chance, and it is very unlikely that the same performance can be reached again in new test data.

In particular, a first look at Tables 5 to 6 and Figs. 10a to 10b can reveal that on one hand the Decision Stump algorithm seems to achieve the maximum accuracy and lowest error values and, on the other, all algorithms have better results around 1.6 to 1.8 seconds. Regarding the nature of the aforementioned algorithm, such a performance that out-scores the rest is rather strange. Also, there is a clear indication that the obtained scores are better in the case of $R''_{i,x'}$, for both accuracy and error and precision and recall.



(a) Accuracy values for R' and R'' datasets

(b) Error values for R' and R'' datasets

Fig. 10. Accuracy and mean square error values for all data sets (R' and R'') and algorithms employed.

The maximum accuracy and minimum error values obtained by the Decision Stump algorithm are: (a) 65.27 and (b) 0.477 for accuracy and mean square error, respectively, and for both data sets. This algorithm classifies the provided data by only one rule that it chooses, a fact that justifies the constant accuracy results when the arousal values were also included in the feature set. Thus, one outcome could be that valence classification regarding rhythm characteristics is dependent on only one acoustic cue that is not the arousal feature. But, due to the results of the other and more

Table 7. Precision and recall values for RepTree algorithm for $i = 5$

	Class $\mathbf{C_1}$	Class $\mathbf{C_2}$
	Dataset $\mathbf{R'_{i,x'}}$	
Recall	0.686	0.556
Precision	0.621	0.625
	Dataset $\mathbf{R''_{i,x'}}$	
Recall	0.721	0.531
Precision	0.620	0.642

Table 8. Precision and recall values for AdaBoost and SMO algorithms for $i = 6$

	Class $\mathbf{C_1}$		Class $\mathbf{C_2}$	
	AdaBoost	SMO	AdaBoost	SMO
		Dataset $\mathbf{R'_{i,x'}}$		
Recall	0.628	0.628	0.617	0.617
Precision	0.635	0.635	0.610	0.610
		Dataset $\mathbf{R''_{i,x'}}$		
Recall	0.663	0.663	0.593	0.556
Precision	0.633	0.613	0.623	0.608

sophisticated algorithms, it is most possible that these results are obtained by dependencies of the specific feature set and SEs and thus are not likely to be repeated at the future with a different data set.

Focusing to the remaining algorithms, from Figs. 10a and 10b it can be seen that the maximum accuracy results are obtained at the same window lengths where the Decision Stump presented its maximum values, i.e., $i = 5$ and $i = 6$, where for $i = 7$ there is a drop at the performance of all algorithms. Hence it is clear that the optimal time length, $t_{optimal}$, for valence recognition from rhythm related characteristics is in the space of [1.4, 2] seconds. For the $t_{optimal}$ values of time length can be seen that RepTree and AdaBoost algorithms' performance surpass all the others.

More specifically, and for $i = 5$, RepTree depicts the greater accuracy and lowest error values for both data sets and among all algorithms except Decision Stump. These values are: (a) 62.28 and 62, 87 and (b) 0.496 and 0.492 for accuracy and mean square error, respectively, and both data sets. Also, the effect of arousal's inclusion seems to be negligible since the increase of accuracy was only 0.59% and the decrease of error was 0.007. These findings are clearly in contrast with the claim that when a listener hears a sound that (s)he does not like, (s)he is likely to feel more aroused and thus the arousal should have an important role in valence recognition. But a closer examination at Tables 5 and 6 reveal that both precision and recall are increased when the arousal was used as a feature. Namely, the exact values of RepTree algorithm for $i = 5$ and precision and recall for both data sets are presented in Table 7.

Examining the values of Table 7 and focusing on the $R'_{i,x'}$ data set, it can be observed that the percentage of SEs in C_2 that are recognized as not part of that class is greater than the percentage of SEs in C_1 that are recognized as not part of their class. However, for both classes in the $R'_{i,x'}$ data set, there is almost the same percentage of SEs (with a difference of 0.004 in precision) that the RepTree algorithm assigned to the correct class.

Regarding the utilization of the $R''_{i,x'}$ data set, it can be seen that the employment of arousal as a class does decrease the assignment of SEs in C_1 to C_2 (increase of recall for C_1) but also increases the assignment of C_2 SEs to C_1 (decrease in recall for C_2). Also, the arousal feature slightly decreases (0.017) the precision for C_2. Hence, for $i = 5$, the arousal feature does increase the correct recognition of low valence class but reduces the one of high valence. The above fact is also supported by Fig. 2, where there are SEs that have high valence and low arousal (2$^{\text{nd}}$ quadrant of the Arousal-Valence space). Thus, and regarding the window length of 1.6 seconds, there is a strong indication that the rhythm related acoustic cues can result in a recognition of SEs assigned in the low valence class (C_1) and the utilization of the arousal feature can increase significantly the recall of such SEs. Nevertheless, SEs assigned to C_1 exhibit difficulties in correct recognition with sole usage of rhythm-related features, and the integration of arousal as a feature seems not to increase the performance of classification algorithms.

A similar behavior can be observed for $i = 6$ but with respect to AdBoost and SMO algorithms. For a window length of 1.8 seconds these two algorithms portrayed the best results in terms of accuracy, precision, and recall among all, excluding Decision Stump. Recall and precision values for these two algorithms are in Table 8.

Both algorithms show a recall and precision score over 60% for both classes in the $R'_{i,x'}$ data set. In particular, the AdaBoost meta-learner achieves the highest recall values, i.e., most instances of a class are classified into the correct class. The usage of arousal as a feature again leads to an increase of recall for C_1 but to a decrease for C_2, strengthening the belief that SEs in the 2$^{\text{nd}}$of Arousal-Valence space can have a negative impact in the performance of the recognition process. Also, neither AdaBoost nor SMO (for $i = 6$) surpass the performance of RepTree (for $i = 5$). However, and taking into account Figs. 10a and 10b, AdaBoost's accuracy is identical to RepTree but the latter performs better when it comes to error values. Thus, there is an indication that tree–like classification schemes tend to perform better than function–based algorithms and meta–learners with these algorithms as base learners.

Recapitulating all the above and focusing on the valence and rhythm dependency, it is evident that, on one hand, SEs valence recognition from rhythm-related characteristics is feasible up to an extent and thus the valence of the listener is also affected by the rhythm attributes of SEs. On the other hand, the window length for which the classification algorithms perform better equals to 1.6 seconds. These facts are more prominent considering recall, precision, accuracy, and error measures from the employed algorithms. As can be seen from the preceding analysis, rhythm characteristics of SEs seem to have a rather limited impact on the elicited valence that is reflected by the obtained low metrics (i.e., precision, recall, and accuracy) values. This outcome is strengthened due to previous conducted research [6] that utilized same methods and fewer algorithms but presented higher accuracy results.

Additionally, the employment of the arousal as a feature increases the recognition's performance but only for the low valence class whereas for high valence has the opposite effect. This fact, compared with the aforementioned ones, implies that valence recognition could significantly benefit by a hierarchical approach and the employment of other aspects of the SE. Also, despite the relatively low recognition scores reported here and taking in parallel into account prior works and results, the outcome of the current investigation seems to provide useful evidences that the relatively–relaxed but well–analyzed relation of valence and the rhythmic SE characteristics can be proved beneficial for increasing the respective arousal recognition scores, thus indirectly strengthening the rhythm impact on the observed listener affective state. Moreover, there is a noticeable difficulty in the recognition of the high valence class. One possible reason could be the binary definition of the presented classification problem (i.e., two classes). Maybe more valence cases (e.g., three—one for negative, one for tranquil, and one for high), could provide better results. Finally, regarding the improvement of the performance for the negative valence class, a hierarchical approach to valence recognition could potentially improve the obtained results.

6 CONCLUSIONS

In the presented work valence recognition from SEs has been conducted with the utilization of solely rhythm-related acoustic cues. The outcome of this work could relate the rhythm of a generalized SE not only with the arousal but also with the valence of the listener. To this aim, a well known data set with emotionally annotated SEs was employed, various and widely employed rhythm-related characteristics were extracted, and several machine learning experiments were conducted. For the latter, a variety of algorithms was used, including meta–learners.

Results reached up to an accuracy of 63% and also portrayed that the rhythm of an SE can affect the listener's valence up to an extent, which implies that there are also other aspects of an SE that can and do affect the receiver's valence. To this outcome contributes the fact that high valence class exhibited the lowest precision and recall values, indicating that a pleasurable condition caused by audio stimuli is strongly affected by other sound parameters. Additionally, the optimal time length of window for valence recognition from rhythm related characteristics is 1.6 seconds. Last, there are clear indications that more valence classes and/or a hierarchical approach could be beneficial to the recognition process.

Concluding, the results of the present work combined with the outcomes of previous one, i.e., [6], could lead to a comprehensive recognition of the rhythm's effect on the emotional state of the listener. This integration is very possible to effectuate the synthesis of SEs with proper rhythm characteristics in order elicit specific emotional reactions to the receiver.

7 REFERENCES

[1] 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, Tech. Rep. B–3 (2007).

[2] K. Hevner, "Experimental Studies of the Elements of Expression in Music," *Am. J. Psych.*, vol. 48, no. 2, pp. 246–268 (1936).

[3] A. Wieczorkowska, P. Synak, R. Lewis, and Z. W. Ra, "Extracting Emotions from Music Data," in *Foundations of Intelligent Systems*, ser. Lecture Notes in Computer Science, M.-S. Hacid, N. Murray, Z. Ra, and S. Tsumoto, Eds. (Springer Berlin Heidelberg, 2005), vol. 3488, pp. 456–465. [Online]. Available: http://dx.doi.org/10.1007/11425274_47.

[4] D. Liu, L. Lu, and H. Zhang, "Automatic Mood Detection from Acoustic Music Data," *4th International Conference on Music Information Retrieval (ISMIR)* (2003 Oct.).

[5] L. Lu, D. Liu, and H.-J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 5–18 (2006 Jan.).

[6] K. Drossos, R. Kotsakis, G. Kalliris, and A. Floros, "Sound Events and Emotions: Investigating the Relation of Rhythmic Characteristics and Arousal," Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on, pp. 1–6 (2013 Jul.).

[7] K. Drossos, A. Floros, and N.-G. Kanellopoulos, "Affective Acoustic Ecology: Towards Emotionally Enhanced Sound Events," Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound, ser. AM '12 (New York, NY, USA: ACM, 2012), pp. 109–116.

[8] W. W. Gaver, "What in the World Do We Hear? An Ecological Approach to Auditory Event Perception," *Ecological Psych.*, vol. 5, pp. 1–29 (1993).

[9] P. N. Juslin and P. Laukka, "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?" *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814 (2003 Sep.).

[10] S. Pouyanfar and H. Sameti, "Music Emotion Recognition Using Two Level Classification," Intelligent Systems (ICIS), 2014 Iranian Conference on, pp. 1–6 (2014 Feb.).

[11] B. Kostek and M. Plewa, "Parametrisation and Correlation Analysis Applied to Music Mood Classification," *Int. J. of Computational Intelligence Studies*, vol. 2, no. 1, pp. 4–25 (2013).

[12] L. Lu, D. Liu, and H.-J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 5–18 (2006 Jan.).

[13] P. N. Juslin and P. Laukka, "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?" *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814 (2003).

[14] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of

Emotion in Audio: What Speech, Music and Sound Have in Common," *Frontiers in Psychology*, vol. 4 (2013 May).

[15] R. Reisenzein, E. Hudlicka, M. Dastani, J. Gratch, K. Hindriks, E. Lorini, and J.-J. Meyer, "Computational Modeling of Emotion: Toward Improving the Inter- and Intradisciplinary Exchange," *Affective Computing, IEEE Transactions on*, vol. 4, no. 3, pp. 246–266 (2013 Jul.).

[16] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic Recognition of Emotion Evoked by General Sound Events," *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 341–344 (2012 Mar.).

[17] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-Based Evaluation and Estimation of Emotions in Speech," *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800 (2007 Oct.). [Online]. Available: http://dx.doi.org/10.1016/j.specom.2007.01.010.

[18] K. Drossos, A. Floros, and A. Giannakoulopoulos, "Beads: A Dataset of Binaural Emotionally Annotated Digital Sounds," Information, Intelligence, Systems and Applications (IISA), 2014 Fifth International Conference on (2014 Jul.).

[19] K. Sun, J. Yu, Y. Huang, and X. Hu, "An Improved Valence-Arousal Emotion Space for Video Affective Content Representation and Recognition," Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on, pp. 566–569 (2009 Jun.).

[20] A. Ortony and T. J. Turner, "What's Basic about Basic Emotions," *Psychological Rev.*, vol. 97, no. 3, pp. 315–331 (1990).

[21] S. Koelsch, T. Fritz, D. Y. v. Cramon, K. Mller, and A. D. Friederici, "Investigating Emotion with Music: An FMRI Study," *Human Brain Mapping*, vol. 27, no. 3, pp. 239–250 (2006). [Online]. Available: http://dx.doi.org/10.1002/hbm.20180.

[22] R. Adolphs, "Neural Systems for Recognizing Emotion," *Current Opinion in Neurobiology*, vol. 12, no. 2, pp. 169–177 (2002).

[23] T. Li and M. Ogihara, "Detecting Emotion in Music," 4th International Conference on Music Information Retrieval (ISMIR) (2003 Oct.).

[24] R. R. Cornelius, "Theoretical Approaches to Emotion," International Speech Communication Association (ISCA) Tutorial and Research Workshop (ITRW) on Speech and Emotion, pp. 3–10 (2000 Sep.).

[25] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200 (1992).

[26] Y.-H. Yang and H. Chen, "Music Emotion Ranking," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 1657–1660 (2009 Apr.).

[27] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. (Elsevier, 2011).

[28] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32 (2001).

[29] O. Lartillot, P. Toiviainen, and T. Eerola, "A Matlab Toolbox for Music Information Retrieval," *Data Analysis, Machine Learning and Applications*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. (Springer Berlin Heidelberg, 2008), pp. 261–268. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-78246-9_31.

[30] J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993).

[31] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297 (1995 Sep.). [Online]. Available: http://dx.doi.org/10.1023/A:1022627411411.

[32] J. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization" (1998). [Online]. Available: http://research.microsoft.com/∼jplatt/smo.html.

[33] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An Efficient Boosting Algorithm for Combining Preferences," *J. Mach. Learn. Res.*, vol. 4, pp. 933–969 (2003 Dec.). [Online]. Available: http://dl.acm.org/citation.cfm?id=945365.964285.

[34] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition* (CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2011).

THE AUTHORS

Konstantinos Drossos

Andreas Floros

Katia L. Kermanidis

Konstantinos Drossos holds a B.Eng. equivalent on sound and musical instruments technology with distinction from the Technological Educational Institute of Ionian Islands. In 2007 he received his M.Sc. from ISVR, University of Southampton. Currently he is a Ph.D. candidate at the Dept. of Audiovisual Arts of Ionian University. His main research interests are emotion recognition from sound events, affective acoustic ecology, audio processing, audio perception, and audio interfaces. He has also worked as an acoustic consultant and as an Adjunct Lecturer at the Dept. of Sound and Musical Instruments Technology of the Technological Educational Institute of Ionian Islands. Mr. Drossos is a student member of the Audio Engineering Society and the Hellenic Institute of Acoustics.

●

Andreas Floros holds an engineering and Ph.D. degree in the area of digital audio technology from the department of electrical and computer engineering, University of Patras. In 2001, he joined the semiconductors industry, leading projects in the area of digital audio delivery over PANs and WLANs, Quality-of-Service, mesh networking, wireless VoIP technologies, and, lately, with audio encoding and compression in embedded processors. During 2003–2005 he was a member of IEEE Tasks Groups 802.11e, .11k, and .11s. For a period of three years (2005–2008), he was an adjunct professor at the department of Informatics, Ionian University, teaching also at the postgraduate (M.Sc.) program Arts and Technologies of Sound organized by the department of Music Studies. Since 2009 he is an Assistant Professor at the department of audiovisual arts, Ionian University. His current research interests focus on analysis, processing and conversion of digital audio signals, digital audio coding and distribution techniques, digital audio technologies for multimedia and networking applications, audio systems for consumer and professional applications, wireless technologies for multimedia applications (with emphasis on Quality of Service for Wireless LANs) and high-quality audio streaming over packet networks. Dr. Floros is a member of ACM, the Audio Engineering Society (currently serving as the Secretary of the AES Greek Section), and the treasurer of the Hellenic Institute of Acoustics.

●

Katia L. Kermanidis was born in Aachen, Germany, in 1975. She received her first degree from the department of Electrical and Computer Engineering at the University of Patras, Greece, in 1999, and her Ph.D. degree in natural language processing from the same department in 2005. She has participated as junior researcher in the Artificial Intelligence Group of the same department in national and European R&D projects in the areas of speech and language processing, data mining, and artificial intelligence. She has been a lecturer at the department of Informatics at the Ionian University, Corfu, Greece, since 2009. She has published 17 journal articles and more than 30 conference papers in the areas of humanistic data mining, machine learning, and natural language processing. Dr. Kermanidis is a member of the Hellenic Artificial Intelligence Society (EETN) and the Technical Chamber of Greece (TEE).