# Experimental Multimedia Systems for Interactivity and Strategic Innovation

Ioannis Deliyannis
*Ionian University, Greece*

Petros Kostagiolas
*Ionian University, Greece*

Christina Banou
*Ionian University, Greece*

A volume in the Advances in Multimedia and Interactive Technologies (AMIT) Book Series

Information Science REFERENCE
An Imprint of IGI Global

Managing Director: Lindsay Johnston
Managing Editor: Keith Greenberg
Director of Intellectual Property & Contracts: Jan Travers
Acquisitions Editor: Kayla Wolfe
Production Editor: Christina Henning
Development Editor: Caitlyn Martin
Typesetter: Kaitlyn Kulp; Tucker Knerr
Cover Design: Jason Mull

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com



Library of Congress Cataloging-in-Publication Data

Experimental multimedia systems for interactivity and strategic innovation / Ioannis Deliyannis, Petros Kostagiolas, and Christina Banou, editors.
pages cm
Includes bibliographical references and index.
ISBN 978-1-4666-8659-5 (hardcover) -- ISBN 978-1-4666-8660-1 (ebook) 1. Interactive multimedia--Design. I. Deliyannis, Ioannis, 1975- II. Kostagiolas, Petros. III. Banou, Christina, 1971-
QA76.76.I59E98 2015
006.7--dc23

2015015841

This book is published in the IGI Global book series Advances in Multimedia and Interactive Technologies (AMIT) (ISSN: 2327-929X; eISSN: 2327-9303)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

# Chapter 6
# Affective Audio Synthesis for Sound Experience Enhancement

**Konstantinos Drossos**
*Ionian University, Greece*

**Maximos Kaliakatsos-Papakostas**
*Aristotle University of Thessaloniki, Greece*

**Andreas Floros**
*Ionian University, Greece*

## ABSTRACT

*With the advances of technology, multimedia tend to be a recurring and prominent component in almost all forms of communication. Although their content spans in various categories, there are two protuberant channels that are used for information conveyance, i.e. audio and visual. The former can transfer numerous content, ranging from low-level characteristics (e.g. spatial location of source and type of sound producing mechanism) to high and contextual (e.g. emotion). Additionally, recent results of published works depict the possibility for* **automated synthesis** *of sounds, e.g. music and* **sound events***. Based on the above, in this chapter the authors propose the integration of* **emotion recognition from sound** *with* **automated synthesis** *techniques. Such a task will enhance, on one hand, the process of computer driven creation of sound content by adding an anthropocentric factor (i.e. emotion) and, on the other, the experience of the multimedia user by offering an extra constituent that will intensify the immersion and the overall user experience level.*

## INTRODUCTION

Modern communication and multimedia technologies are based on two prominent and vital elements: sound and image/video. Both are employed to transfer information, create virtual realms and enhance the immersion of the user. The latter is an important aspect that clearly enhances usage experience and is in general greatly aided by the elicitation of proper affective states to the user (Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). Emotion conveyance can be achieved from visual and auditory channels

DOI: 10.4018/978-1-4666-8659-5.ch006

(Chen, Tao, Huang, Miyasato, & Nakatsu, 1998). Focusing particularly on sound, one of its organized forms (music) was evolved as a means to enhance expressed emotions from another audio content type (speech) (Juslin & Laukka, 2003). But both aforementioned types are only a fraction of what actually occupies this perception channel (Drossos, Kotsakis, Kalliris, & Floros, 2013). There are non-musical and non-linguistic audio stimuli that originate from all possible sound sources, construct our audio environment, carry valuable information like the relation of their source and their receiver (e.g. movement of a source towards the receiver) and ultimately affect the listener's actions, reactions and emotions. These generalized audio stimuli are termed Sound Events (SEs) or general sounds (Drossos, Floros, & Kanellopoulos, 2012). They are apparent in all everyday life communication and multimedia applications, for example as sound effects or components of a virtual world depicting the results of user's actions (e.g. sound of a door opening or user's selection indication) (Drossos et al., 2012).

There are two main disciplines that examine the conveyance of emotion through music, namely the Music Emotion Recognition (MER) and Music Information Retrieval (MIR). Results presented from existing studies in these fields show **emotion recognition accuracy from musical data** of approximately 85% (Lu, Liu, & Zhang, 2006). Based on findings from MER and MIR there are some published works that are concerned with the synthesis of music that can elicit specific affective conditions to the listener (Casacuberta, 2004). But since music can be considered as an organized form of sound, the question if such practices can be applied to SEs was raised. Towards exploring this scientific area, recently, an ongoing evolution was initiated of a research field that focuses on **emotion recognition from SEs**. Although published works in that field are rather scarce (Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013), it has been shown by previous research conducted by the authors that **emotion recognition from SEs** is feasible with an accuracy reaching up to 88% regarding listener's arousal (Drossos et al., 2013). In addition, the authors have proposed and presented several aspects regarding systematic approaches to **automatic music composition** (Kaliakatsos-Papakostas, Floros, & Vrahatis, 2012c) and sound synthesis (Kaliakatsos-Papakostas, Epitropakis, Floros, & Vrahatis, 2012a), focusing on the generation of music and sound that adapts to certain specified characteristics (see also (Kaliakatsos-Papakostas, Floros, & Vrahatis, 2013c) for a review on such methodologies).

Thus, combining the aforementioned **automatic synthesis** methodologies with the findings from **SEs emotion recognition**, one can potentially synthesize SEs capable to elicit specific affective conditions to the listener. In this chapter proposal we intend to present novel findings and methodologies for affective enhanced SEs synthesis. According to authors' knowledge, there is no other similar published work. Such audio material can be used to enhance the immersion and the audio experience of users in multimedia by inflating the emotional conveyance from the application to the user. The rest of this chapter proposal is as follows. In the second section a brief overview is presented that concerns the state-of-the-art in **audio emotion recognition**, particularly focused on **emotion recognition from SEs**. The **automated music and sound synthesis** counterpart of the proposal is discussed in the third section whereas in the fourth section are some possible and proposed applications.

## AUDIO EMOTION RECOGNITION

In general, **audio emotion recognition** can be considered as a Machine Learning task (Drossos et al., 2013). It consists of two stages, i.e: i) Training, and ii) Testing. In the former, a classification algorithm is fed with the emotional annotations of the sounds and a group of extracted features and produces a

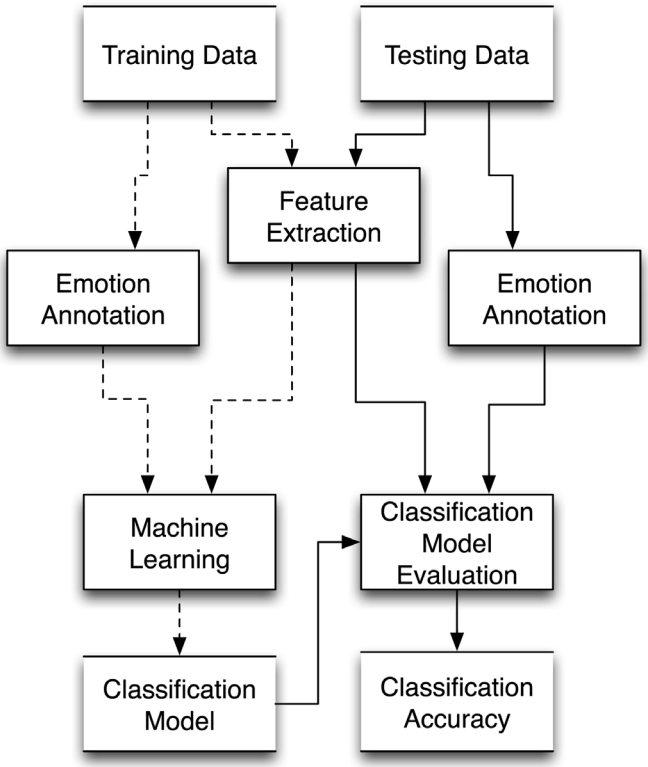
categorization model. The accuracy of the model, and thus of the whole process, is evaluated at the latter stage. This process is illustrated in Figure 1.

As can be seen in Figure 1, for both stages there is a set of components that are essential for the accomplishment of this task (Drossos, Floros, & Giannakoulopoulos, 2014) and these are:

1. **Emotionally annotated audio dataset**, which provide the training and testing data and their emotion annotations
2. Extracted features, resulting from the feature extraction process
3. Classification algorithm, employed in the machine learning phase
4. Classification model, resulting from the machine learning process

The determination of all the above components is relied on basic concepts of the **audio emotion recognition** process. The employed annotated dataset is determined by the emotions model that will be used in the recognition process and the extracted features are likely to focus on specific attributes of the audio signals. In addition, different algorithms exhibit various accuracy results in the process and some perform better when combined with specific models, features and emotions. Regarding the affective synthesis concept presented in this chapter, extracted features and classification algorithms are likely to have an impact both on the synthesis part (i.e. which features of signals will be processed) and to the classification of the produced sounds. In other words, in order to synthesize an **affective sound** the ef-

*Figure 1. Illustration of the audio emotion recognition process. With dotted line is the training and with solid the testing stage*



fect of specific features to the elicited emotion must be know and the resulting sound from the synthesis process should be evaluated in terms of the emotion that communicates to the listener.

In the following subsections will be presented the currently used emotional models along, a brief overview of annotation methods, the results and methods of the state-of-the-art in **emotion recognition from music** and **sound events** and the most common features used in such processes.

## EMOTIONAL MODELS AND ANNOTATION METHODS

The emotional models can be discriminated in two abstract categories; one using verbal description for referring to emotions (e.g. "Happiness", "Sadness") and another modeling emotion as a resultant of affective states. The former are referred to as discrete models whereas the latter as dimensional models.

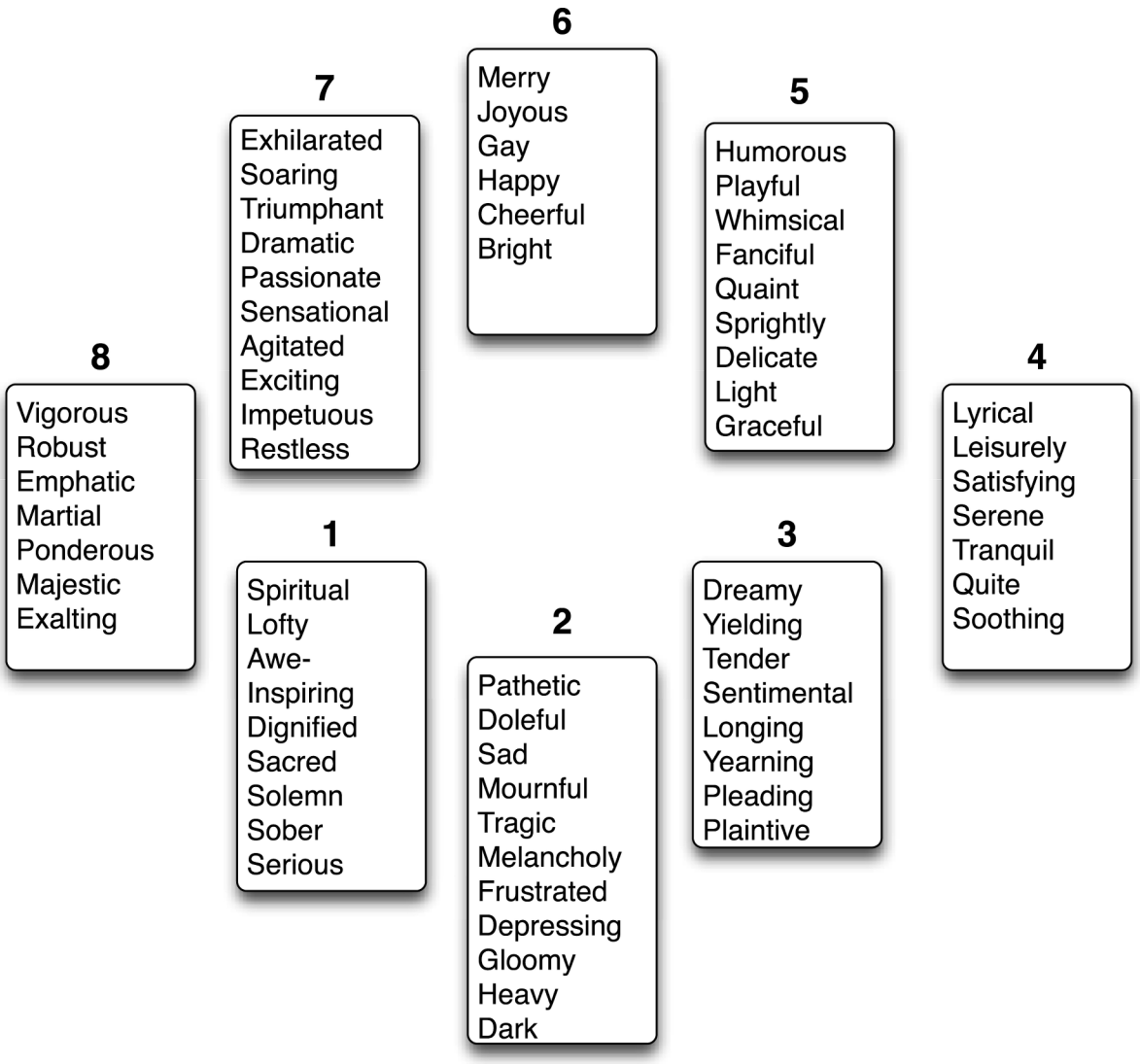
Discrete models emerge from the basic emotions model, which is dated back to the Darwing era and based on William James' theory of emotions (Cornelius, 2000). It consists of 4 emotions, termed as basic ("Happiness", "Sadness", "Fear" and "Anger"), and states that all other emotions can emerge from those. This model was questioned thoroughly in (Ortony & Turner, 1990) regarding the claim that a set of emotions can be used as a starting palette for all others. Probably due to immediate connection with human body components and functions, this model is used in modern published researches in the field of neuroscience (Drossos et al., 2012; Koelsch, 2010; Adolphs, 2002). Another frequently used discrete emotional model is the list of adjectives that contains several groups of synonym words (adjectives) (Hevner, 1936; Wieczorkowska, Synak, Lewis, & Ra, 2005). Originally consisted of 8 adjectives groups with each one group corresponding to specific emotional states (Hevner, 1936). An illustration of this model can be seen in Figure 2.

Some enhancements have been proposed to the adjectives list model, mostly regarding the increase of the adjectives groups, e.g. in (Li & Ogihara, 2003) there have been proposed 13 groups instead of the original 8. Although that this model can be considered as discrete (Drossos et al., 2012), the arrangement of the adjectives groups follows the concept of a 2D dimensional model, where each group is placed according to the increment in the affective states. For example, moving from group 8 to group 4 reflect the increase of the listener's arousal, whereas groups 6 and 2 exhibit opposite valence.

Dimensional emotional models represent emotion as a resultant of a set of components and emerged from the Circumplex model of affect (Russell, 1980). The latter render affective states, typically those of arousal and valence, in order to represent emotions. Thus, according to the employed affective states the model can be two-dimensional, three-dimensional etc. Every affective state is regarded as a component to the resulting space (2D, 3D etc.) and the final emotion is the resultant on that space. On a later stage, different verbal descriptions can be assigned to clusters of values in the aforementioned space and thus leading to the mapping of values to emotions (Drossos et al., 2014). Although that there are published works which portray such mapping, this is purely qualitatively since there is not a quantitative association of the values with verbal description of emotions. Typically, there are two dimensions utilized in dimensional models and these are 2: i) Arousal, and ii) Valence. There are proposals for the employment of extra dimensions, e.g. Dominance, but this is likely to result in an increased complexity of the model (Drossos et al., 2014).

The proposal and adoption of dimensional models also reduced a complication that existed regarding the cross-evaluation of emotional research. When a discrete model is utilized, the annotations of the data are performed by directly capturing the elicited emotion, most probably with a "question and answer"

*Figure 2. Illustration of the List of Adjectives. (Adapted from Hevner, 1936; Wieczorkowska et al.2005)*




scheme. Thus, the participants in the annotation process have to choose the specific verbal description of emotion that reflects their affective condition in order to perform an annotation. Their choices are limited to the used words by the conductors of the process, which words are also the emotions incorporated in the model being utilized. Although that some verbal descriptions are quite straightforward and most probably used to represent the exactly same emotion (e.g. “Anger” or “Fear”), there are employed words that is not clear what emotional condition represent, for example, “Joy” and “Enjoyment” or “Cheerful” and “Happiness” (Juslin & Laukka, 2003). This uncertainty obscures the cross-evaluation and usage of results between and among different works when discrete emotional models are employed (Juslin & Laukka, 2003).

Contradictory, when dimensional models are utilized the annotation is performed by the depiction of the emotional states employed in the model. Even if there are some published works where the an-

notation was performed by the direct outline of the resultant in the model's area (Yang, Lin, Cheng, & Chen, 2008), there is an annotation method which allows the representation of 3 emotional states and offers the capability for immediate association with a corresponding emotional model. This is the Self Assessment Manikin (SAM) and was presented in (Bradley & Lang, 1994). With the SAM, the participants in the annotation process choose a figure from a set of images. The chosen image directly reflects the value for the corresponding dimension in the model. An illustration of the available set of figures in SAM is in Figure 3. In Figure 3a are presented the SAM figures for arousal in descending sorting, i.e. from left to right the manikin figures portray a decrease to arousal. In a same arrangement are the SAM figures for valence in Figure 3b, where from left to right the valence is decreasing. In Figure 3c are the SAM figures for Dominance in an ascending fashion, i.e. from left to right the dominance is increasing.

The utilization of an emotional model in the affective synthesis can and will have an impact on the targeted emotions. A discrete one will likely uncomplicate the process of recognition and evaluation, since the targeted state of the listener will be incorporated in a single verbal description. But, this single word is likely to cluster a set of verbal description either not presented in the annotation and evaluation process or perceived from the participants (in both aforementioned stages) as synonyms to the employed emotion verbal description. A dimensional model although that can allow a more elaborate targeting

*Figure 3a. SAM figure for arousal*

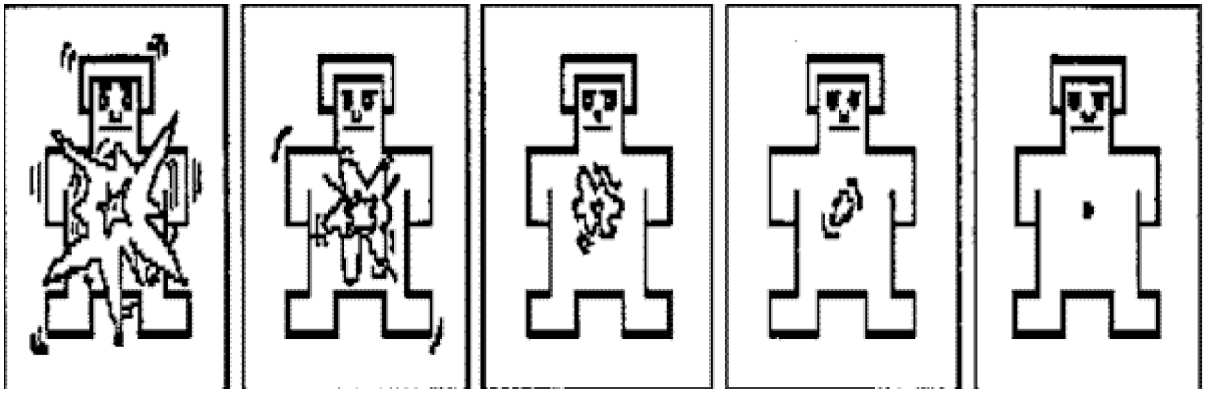


*Figure 3b. SAM figure for valence*

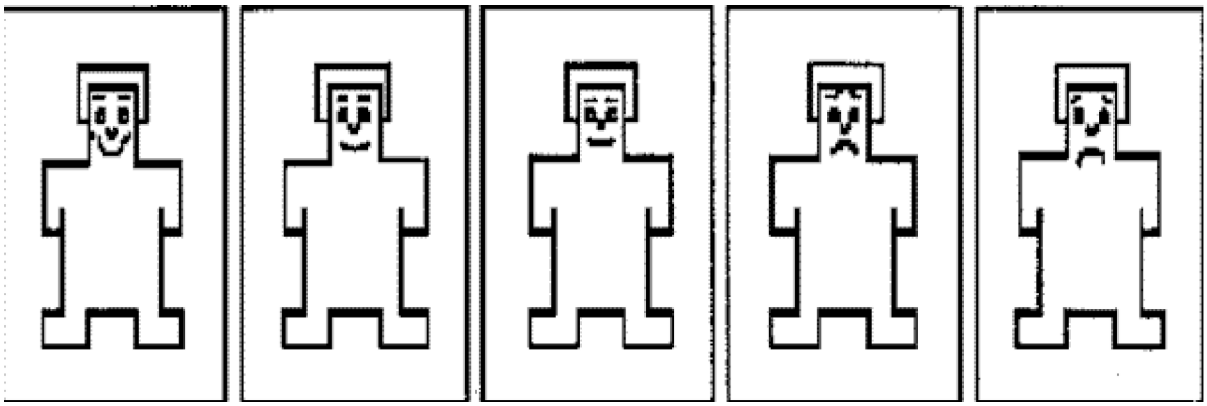
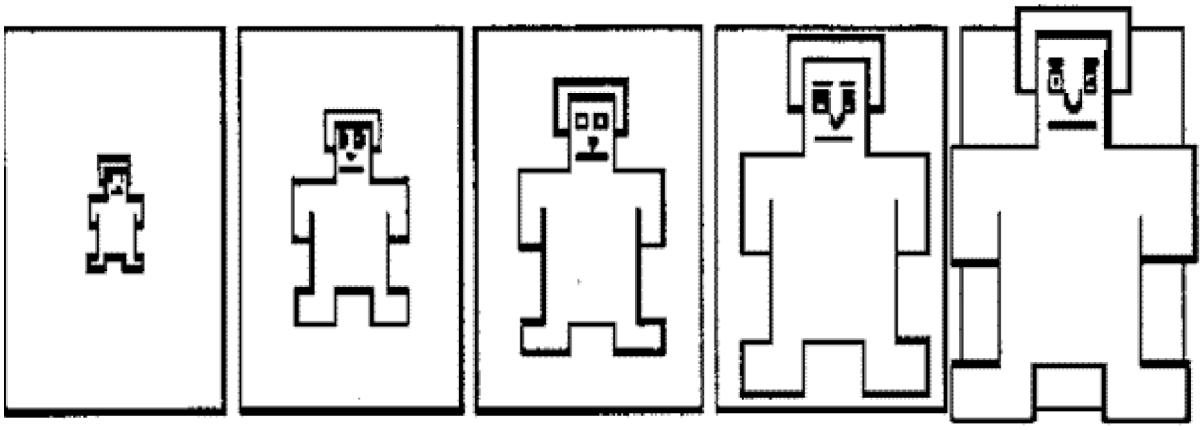


*Figure 3c. SAM figure for dominance*



of emotions, by not referring to verbal description but to affective states, will cluster values in the used space (e.g. arousal and valence) almost arbitrary due to the lack of exact quantitavely association of values with verbal descriptions of emotions. On the other hand, a dimensional model offers a more focused approach to the affective states, regardless of what words are used to describe possible combinations.

Taking into account possible applications of **automated synthesis**, e.g. **automated synthesis** of **emotional enhanced sound effects** for video games, the approach of a dimensional model seems to be more appropriate. This is due to the fact that in such artistic and/or application environments the main aim is to target the alteration of an affective state of the spectator/listener, e.g. to elicit more or reduce the arousal or pleasure than to actually induce a specific emotion.

## EMOTION RECOGNITION FROM MUSIC

The extraction of emotional information from music falls in the Music Information Retrieval discipline. One of the primary aims is the enhancement of music categorization by adding a content oriented parameter (emotion/mood) in addition to or instead of the legend "Artist / Year / Album" scheme (Zhang, Tian, Jiang, Huang, & Gao, 2008). The process is according to the one illustrated in Figure 1 and typical accuracy values can reach up to 89% (Xiao, Dellandrea, Dou, & Chen, 2008; Schmidt, Turnbull, & Kim, 2010).

In particular, as employed dataset are used annotated excerpts or complete musical pieces. The used emotional models vary. Both discrete and dimensional are employed but it seems that in recent works the latter are preferred (Kim et al., 2010). Typical extracted features are from various categories, i.e. timbre, dynamics, pitch and rhythm, e.g. attack time and slope, Mel-Frequency cepstral coefficients (MFCCs), spectral contrast, RMS energy, tempo, beat spectrum and histogram (Li & Ogihara, 2003; Kim et al., 2010; Yeh, Lin, & Chang, 2009). Usually their extraction is based on a short time scheme, e.g. consequent time frames of the musical information, in order also to evaluate the evolution of features as the musical stimuli develops through time (Schmidt et al., 2010; Lu et al., 2006). The categorization is performed with simple and meta-learning algorithms. Amongst the most used are Support Vector Machines, Tree-Based algorithms (e.g. C4.5), Gaussian Mixture Models and AdaBoost (Kim et al., 2010).

Results indicate that the correct classification of music regarding the elicited emotion is feasible, as indicated earlier in this section. But the question of whether the aforementioned results can be employed in a synthesis process is at interest in this chapter. According to the latter task, the employed classification algorithms seem to have an important role. For example, if a set of features, e.g. $F_i$, where F is the features set and *i* is each feature's index, there is the need for knowledge of the exact value that each *i* must have in order the synthesized audio material to elicit a specific emotional state or emotion. Thus, in an **automatic synthesis** procedure the results of the classification algorithms must be utilized in order to, one hand, evaluate the produced audio material and, on the other, to define criterions for the technical features' values that the synthesized sounds must have. For a most straightforward approach, Tree-Based algorithms seem to have an advantage, due to an uncomplicated interpretation of the features' values thresholds that determine the final categorization.

There are published works focused on the **affective music** synthesis task (Kejun & Shouqian, 2010). To authors' knowledge, most published works employ a dimensional model for emotions. Also, in order to produce a musical material that elicits specified emotional reactions to listeners, Genetic Algorithms (GAs) are utilized which evolve the material in order to obtained the needed results. For example, in (Zhu, S. Wang, & Z. Wang 2008) an approach for synthesizing **affective music** is presented by employing interactive GAs. The GAs control the performance parameters, based on the KTH rule system (Friberg, Bresin, & Sundberg, 2006), and the user interactively indicates his emotion. User's feedback is further employed in order to alter the performance and achieve the targeted emotion. Also, in (Sugimoto et al., 2008) is presented a work which deals with the design of emotional music through GAs by also employing an interaction from the Autonomic Nervous System (ANS) of human. In it, the final music material induces the targeted emotion by evolving musical characteristics according to the readings from the ANS and the utilization of GAs. The results show that the targeted emotions can be elicited to the listeners. Finally, in (Kejun & Shouqian, 2010) is presented a framework for emotional music design with GAs, which includes a proposed flow that will control the musical material's features alteration in order to perform the induction of the targeted emotion. The work concludes that there is the need for better GAs that will focus on the **automated music synthesis** targeting specific emotion elicitation. This aspect is covered in the **automated synthesis** part of this chapter.

## EMOTION RECOGNITION FROM SOUND EVENTS

The recognition of emotion from general sounds (i.e. **sound events**) is a rather new field with sparse published works (Weninger et al., 2013). The recognition process is also conforms to the one outlined in Figure 1 but, and in contrast to the **emotion recognition from music**, there are few public available datasets with emotion annotations (Drossos et al., 2014). These are: a) IADS dataset, consisted of 167 monophonic sounds events with annotations for arousal, valence and dominance (Bradley & Lang, 2007), b) the Emotional Sound Database (ESD), with 360 **emotionally annotated sound events**, retrieved from the findallmusic[1] web page, and annotated by 4 subjects and utilizing the evaluator weighter estimator (Schuller et al., 2012; Grimm, Kroschel, Mower, & Narayanan, 2007), and c) BEADS dataset, consisting of 32 binaural sounds, allocated at 0, 45, 90, 135 and 180 degrees and annotated for arousal and valence (Drossos et al., 2014). According to the published works in the mentioned field, the acoustic cues extracted are similar to the **emotion recognition from music** process (mentioned previously).

Due to the inherent connection of the **sound events** with the acoustic environment and hence to the acoustic ecology, in (Drossos et al., 2012) was proposed the enhancement of this concept. More specifically, it was proposed the **Affective Acoustic Ecology** where the emotional dimension is also regarded in the legend Acoustic Ecology concept. The primary component of the proposed framework is the **sound event** that is defined as a structure having discrete characteristics that can affect the elicited emotions. These characteristics are:

- A sound waveform
- Manifestation of source's instantaneous spatial position, relative to the receiver
- Duration (time)
- Indication of the sound's creation procedure/how the sound was created, e.g. impact, friction etc.
- Evidence of vibrating objects' nature state (i.e. solid, liquid, gas)
- Semantic content

Apart from the last component, i.e. the semantic content, the remaining could be used in the emotion recognition process. Although that there was proposed a framework for the nature of sound which also includes the elements of the creation procedure (impact, friction, etc.) and the nature of the vibrating object, until the time that this work is prepared only the first two elements have been used in emotion recognition from **sound events**.

Regarding the waveform, there are few published works that are concerned with extracted technical features and emotion recognition. In (Drossos et al., 2013) there is a work that evaluates the legacy concept that rhythm has an effect on the arousal. Several acoustic cues, all related to rhythm characteristics, have been utilized and the recognition process was performed using Artificial Neural Networks (ANN), Logistic Regression and K-th Nearest Neighbors. The presented accuracy results reached up to 88%. Also, in (Weninger et al., 2013) was attempted the emotion recognition from the ESD and a investigation on whether there are common features than can be used for **emotion recognition in speech, music and general sounds**. Results indicated that such a task is difficult but there are acoustic cues that are functioning in an opposite manner for these fields, i.e. some features have the opposite effect in speech, music and general sounds.

Summarizing the above, there are a variety of technical cues used for emotion recognition which covers seemingly all attributes of an audio signal as described in the **Affective Acoustic Ecology** concept. The employed models can reveal that there is an association of the affective states with the technical features. Although that there is not an analytic and quantitavely association of the features values with values in the dimensional model, there is a qualitatively connection. In addition, there is also an observed variation of the employed algorithms for the machine learning stage. From the aforementioned algorithms there are some that can be considered as more suitable for **automatic synthesis**, i.e. Tree based. Hence and combining all the above, the creation of a framework for **automatic synthesis** of **affective sounds** is feasible by integrating current findings in the field of audio emotion recognition and **automatic synthesis** techniques. In the next sections these techniques will be presented and discussed.

## AUTOMATIC SYNTHESIS

The task discussed in the proposed book chapter concerns the **automated music and sound synthesis** based on the guidelines that are provided by targeted audio-emotional features. To this end, the utilization of adaptable music and sound synthesis techniques will be discussed regarding their potential to compose music and **sound events** that adhere to specific audio and music features that convey emotional information. The systematic approaches that perform the desired task pertain to the category of the supervised synthesis techniques, as analysed in (Kaliakatsos-Papakostas et al., 2013b). Therein, the **intelligent music composition** systems have been categorized in three main groups:

1. Unsupervised **intelligent composition**: the algorithmic composition intelligence is expressed through simple rules that produce complex, unpredictable but yet structured output, a behavior that of- ten resembles natural phenomena.
2. Supervised **intelligent composition**: intelligent algorithms are utilized to modify the **automatic composition** system's parameters so that it may be able to compose music that meets some predefined criteria, not necessarily in real-time.
3. Interactive **intelligent composition**: the system is acknowledging the human preference in real-time and becomes adapted to it, by utilizing intelligent algorithms. Human preference is expressed either by a selection-rating scheme, or by performing tasks (like playing an instrument or adjusting target parameters in real-time).

According to the aforementioned categorization, supervised synthesis discusses the automated generation of music and audio according to the "supervision" provided by values describing a targeted output, e.g. by audio and music features. It is therefore studied whether the utilisation of audio and music features that can accurately describe the emotional content of SEs (as discussed in the second section), can indeed provide accurate compositional guidelines to supervised synthesis methodologies.

Supervised music and audio synthesis is performed effectively by evolutionary techniques, which allow the generation of output that adapts to better solutions towards solving a problem. The effectiveness of these algorithms derives from the immense force of population dynamics, which com- bines exploration and exploitation of the solutions' space, based on the sophisticated evolution of the large numbers of possible solutions. Several evolutionary approaches have been examined for the supervised generation of music and sound. Examples of such methodologies that have been employed for tasks related to the problem at hand are genetic programming (GP) (Koza, 1992) for block-based sound synthesis (Garcia, 2001a), genetic algorithms (GAs) (Holland, 1992) for the evolution of rhythms (Kaliakatsos-Papakostas, Floros, Vrahatis, & Kanellopoulos, 2012b) and differential evolution (DE) (Storn & Price, 1997) for the evolution of tones (Kaliakatsos-Papakostas, Epitropakis, Floros, & Vrahatis, 2013a). Thereby, this section of the chapter will describe the development of hybrid evolutionary systems that compose audio material – from the level of sound to the level of music. These systems could encompass specific emotional meaning, potentially harnessing multimedia systems with emotionally–driven content generation capabilities.

## Supervised Intelligent Music Composition Systems

The systems that pertain to the category of supervised **intelligent music composition** utilize intelligent algorithms to achieve music composition under some qualitative guidelines, which are often called "target features", or simply "features". The underlying model that these systems utilize to produce music, is either tuned or created from scratch with the utilization of intelligent algorithms, towards the directions dictated by the features that the music output has to satisfy. An advantage of the supervised IMC systems is their ability to produce music with certain aesthetic and stylistic orientation (Manaris et al., 2007), in contrast to unsupervised composition systems. The hypothesis in this chapter, is that these systems could be driven not only by features with aesthetic meaning, but also by features that convey emotional information, proving the composed music and sound an emotionally as well as aesthetically oriented texture. Therefore, the formulation of these systems incorporates the following challenges: a) apply an intelligent algorithm to optimally traverse search spaces, producing numerical output, b) create an interpretation of this output to musical entities and c) select a proper set of features that describe the desired aesthetic and emotional music characteristics.

The selection of proper features is of vital importance for the supervised systems' performance not only in terms of aesthetic quality, but also for accurate emotional orientation. However, while these features provide landmarks for the system to compose music with certain characteristics, it is also important that the algorithm should allow the introduction of a considerable amount of novelty to the composed music. The selection of proper features is thus crucial: on the one hand they should capture the aesthetic and emotional essence of the music to be composed and on the other hand they should not over-determine the compositions, a fact that would lead to "replicate" compositions, depriving form the system the ability to introduce novelty. Research on supervised **intelligent composition** mainly addresses the establishment of effective intelligent methodologies that tune the composition models, restraining from the fact that the quality of the targeted features that describe music, are equally important.

## Feature-Driven Composition Systems with Genetic Algorithms

The music composition algorithm may incorporate a set of parameters that define its compositional style and capabilities. It is thus crucial that a proper combination of these parameters is defined for the automatic creation of music, so that it may exhibit certain characteristics and aesthetic value of high quality. The utilization of Genetic algorithms (GA) provides proper values for these parameters, given a set of features that describe how the produced music should sound like. Thus, not only the formulation of a proper parametric model is required, but also, equally importantly, to the formalization of measures that accurately describe the target music texture. Regarding the perspective of this chapter, music texture is not only deteriorated in stylistic constraints, but it extents to the emotions conveyed by the audio/music output.

The Genetic Algorithms (GA) is a class of algorithms inspired by natural evolution, which iteratively produce better solutions to a problem. These algorithms belong to the wider class of "heuristics", meaning that they initially "guess" a set of random solutions and produce new ones grouped in generations, by utilizing information only from their current generation and their product candidate solutions. Specifically, the candidate solutions within a generation in a GA scheme are evolved using operators resembling natural genetic phenomena (like crossover and mutation) and a selection procedure that propagates the consecution of generations towards better solutions. A better solution means that the set

of the model's parameters that this solution describes, gives a more satisfactory result in regards with a qualitative measurement that is called "fitness function". Among the most popular genetic operators are the crossover and mutation. Crossover incorporates the combination of the parameter between two "parent" solutions for the creation of two "children" solutions, while mutation is the random reassignment of certain parameter values of a "parent" solution to produce a new "children solution". The progression from one generation of "parent" solutions to the next, is finally accomplished with a selection process that allows the "better" fitted solutions among parents and children to form the new generation.

The key-notion in the GA is the precise and informative description of what a "better" solution is. In the case of supervised IMC, precise denotes the correct demarcation of the target musical attributes that the automatically composed music should encompass. The term informative expresses the necessity to model the target musical attributes in an as "continuous" manner as possible, abstaining the hazard to create non-smooth and discretized error surfaces that abound in local optima.

Among the first works for supervised composition using "objective" musical criteria for the assignment of fitness evaluations, was the work of Papadopoulos and Wiggins (1998). In this work a system was presented which was composing jazz solos over a given chord progression. The solutions to the problem were the melodies themselves, specifically pitch-duration pairs, and after a random initialization, GA with some special genetic operators with "musical meaning" was applied, fostering new generations of possible solutions-melodies. The evaluation process of each candidate solution-melody was based on eight evaluation indicators, borrowed by music theory. The results were reported to be promising, however a thorougher statistical examination of the results was not realized, according also to the authors' opinion in the concluding section of their work. A similar system was introduced in (Ozcan & Eral, 2008), also providing a downloadable application called AMUSE. In this work a set of ten music features were used for fitness evaluation. Furthermore, experimental results of a questionary-based research on a group of 20 participants, indicated that these features are linked to human preference, at some extent. The utilization of a fitness function based on music theoretic criteria was also utilized for the creation of four-part melodies from scratch (Donnelly & Sheppard, 2011).

On the other hand, features not related to music theory have also been utilized. These features are related to informatics and statistics, measuring aspects of melodic information capacity either through compressibility, or through various characteristics of music that can be translated into discrete probabilities. In (Alfonseca, Cebrian, & Ortega, 2007) the Normalized Compression Distance (NCD) is used to allow a GA compose music pieces that resemble the pieces of a certain style. The sum of NCDs between each possible solution-melody and a set of target pieces in a certain style is computed, and a solution-melody with a better fitness is the one for which a smaller sum of NCDs is obtained. Systems that genetically evolve Cellular Automata (CA) (Lo, 2012) and FL-systems (Kaliakatsos-Papakostas et al., 2012b) for music and rhythm composition have also been presented, where again fitness is calculated with the utilization of probability measures like n-Grams (Markov transition tables), Shannon Information Entropy and Compression Rate among others.

## Feature-Driven Composition Genetic Programming

Genetic Programming (GP) works under the same evolutionary principle with the GA, that is evolving initially random possible solutions to new ones that are better fitted to the problem at hand. The difference between GP and GA is the problem's formulation. In GA, the optimization process targets at the model's parameters, while the utilization of GP allows the optimization of the model per se, since the popula-

tions of possible solutions incorporate entire programs that actually form the model. These programs are constituted of short sub-program segments, hierarchically structured in a tree-like representation. The genetic operators are similar to the ones used by the GA; however, they act on tree branches instead of string-like or numeric chromosomes. The crossover operator for example, exchanges sub-trees of the parents' trees, creating children that comprise of combined sub-program parts of their parents. Through a similar selection process as in the GA, new populations of possible solutions-programs are created which are better or equally fitted to the problem at hand.

## Sound Synthesis

Among the first works that examined **automated sound synthesis** through systems evolved with GP was the one presented in (Putnam, 1994). Therein, the author reports on some initial results yielded by a system that evolves simple mathematical functions through GP, with interactive fitness evaluation (i.e. evaluation provided by the user). These functions were producing values that constructed a waveform. Main aim of the system was to lead generations of such functions to ones that produced more pleasant sonic waveforms, since the evaluations incorporated the pleasantness itself, as perceived by the users of the system. However, as the author of the report (Putnam, 1994) notices, not all the noises that were produced by the system sounded either pleasant or interesting.

Although the methodology described in the previously discussed report did not yield the expected results, it is the first document that reveals some of the principles for evolution of sounds up to date. Two of the most important, among others, are:

1. Sound modeling plays an essential role in GP design. Thus, the power of GP is limited to the extent that the modeling methodology demarcates, i.e. very meticulous selection of non–terminal nodes is required.
2. In interactive evolution, it is important that the first few generations encompass interesting and non–disturbing individuals, since if the users lose focus in the first generations, then they are hardly able to drive evolution to a specific area of desired solutions.

The phenomenon discussed in point 2 above is widely known as user fatigue and is a matter of serious concern for all interactive evolutionary systems, not only the ones that produce artistic material. However, when evolution concerns artistic material there can hardly be a subjective target. This fact can intuitively be described as a process of wandering optimizer, since an indecisive user would force the algorithm to wander the space of possible artistic creations without preference to certain regions – further amplifying the fatigue on the user's side.

A different philosophy of GP algorithms for sound synthesis concerns the generation of sounds towards matching a targeted sound. Therefore, the evaluation of the produced output is not the subjective choices of users, but the objective similarity between a produced and the target sound. The distance that defines this similarity is computed through comparing some audio features extracted from both signals (generated and target). A general overview for such systems has been given in (Garcia, 2000) where the sound generation module includes some advanced signal processing units, instead of simple functions. These processing units can either be used as terminal nodes (e.g. sound generating oscillators) or as non–terminal nodes (e.g. with additive or frequency modulation or by combining/splitting signals). Additionally, this primitive work that expanded in subsequent publications discussed in the next paragraphs,

introduces an interesting fitness evaluation scheme that concerns the distance computation between the produced and the target sound not only by comparing their features, but also by having human judgment on the similarity. This aspect is interesting, since the similarity in audio features does not necessarily impose perceptual similarity, and vice versa. However, the paper under discussion does not provide any experimental evaluation on neither fitness estimation approach.

The approach discussed in the previous paragraph was materialized in two publications, (Garcia, 2001b) and (Garcia, 2001a), that focused on the generation of sounds according to targeted sounds provided by an acoustic instrument. The sound generating modules that were combined with GP included units categorized in three classes according to the input and output they were receiving and sending:

1. Source: 1 output (e.g. a float number indicating frequency or amplitude).
2. Render: 1 input (e.g. the block where an audio output device will be plugged).
3. Type A: 2 inputs and 1 output (e.g. signal addition and multiplication).
4. Type B: 1 input and 2 outputs (e.g. an oscillator or a filter).

Such blocks can be combined with GP and produce sounds with a great variety of characteristics. Since there is no human evaluation evolved, there are also no requirements for constraints during the initial steps of evolution. Therefore this approach takes advantage of the full potential of evolutionary processes, which lays in combining great numbers of randomly combined and altered individuals. The work presented in (Garcia, 2001a) describes a prototype system that implements the above-described methodology, named automatic generation of sound synthesizers (AGeSS). However, there are no indications about the accuracy of the produced results, except from images of spectrograms for target and produced sounds. Another approach that was inspired by the previously described work was presented in (Donahue, 2013). Therein, a similar formalization of sound producing–altering blocks was followed, but under a more sophisticated evolutionary scheme. These blocks can be divided in three categories, according to the number of children they are allowed to incorporate in the tree representation of the programs. These blocks and their parameters are evolved towards constructing synthesizers that "mimic" target sounds. The currently described thesis provides some very interesting insights about what audio similarity may concern regarding human perception; however for the purposes of the book at hand the focus remains on the evolutionary aspects of the work under discussion.

A considerable amount of work in the field of music acoustics has focused on the formulation of physical models for acoustic musical instruments (Smith, 2010). On the one hand, such an approach allows for further examination and understanding of the underlying dynamics that describe the vibrating parts of an instrument. On the other hand, the model's output potentially leads to the implementation of hardware and/or software sound generators, which produce sounds that are similar to the ones reproduced by the modeled instrument. The question addressed in the paper (Wilson, 2002) is the following: given a recorded note of a musical instrument, can the parameters of its physical model be adjusted so that the model produces a similar note event? The research in (Wilson, 2002) examined the case of the physical model of a violin, which descriptively encompasses some basic parameters that describe not only the frequency of the reproduced note, but also expressional characteristics like the bow velocity or bow force.

To approach the problem of physical modeling, a deep understanding of signal processing is required and therefore a detailed explanation of processes in the work under discussion is beyond the scope of this book chapter. A general description of the methodology concerns the construction of functions by

combining simple mathematical expressions with GP. These functions receive audio features as input and produce output that is fed to the parameters of the violin physical model. Therefore, the fitness evaluation is based on the differences between the model output and the target concerning magnitude frequency spectrum among others. The evaluation of the model was based on three different target samples: two recorded violin notes from Bach's solo partitas for violin and a third note recorded by the physical model itself. The presented results are depictions of target and produced sound spectrograms, while the author comment that the system's performance was not satisfactory. The author also proposes that a possible step towards improving the model's performance is to embody an additional model for room reverberation, since the frequencies introduced by room acoustics are not considered by the violin physical model.

## Symbolic Music Composition

Automated music composition is a special field of study, since a great portion of the evaluating process also address the factor of subjectivity in what sounds pleasant, or even musical. Systems that incorporate evolutionary mechanisms, as GP-based methodologies, are called to perform fitness evaluation either by interactive means, requiring from the user to provide fitness evaluation through rating simulations, or by developing fitness functions that effectively quantify qualitative characteristics of music. Both approaches have strong and weak points. Interactive evolution "models" the human preference directly, by letting the user decide for the pleasantness of the generated output. However, the implications of user fatigue do not allow a great number of evaluations, imposing restrictions to the number of individuals and generations of the evolutionary scheme, in addition to the potential misleading ratings from the user within the rating simulations – since user fatigue also affects the judgment of the user. The development of qualitative fitness functions solves the problems addressed in interactive evolution, however such functions are doubtfully effective since human subjectivity and creativity are yet imperfectly computationally modeled. The examined GP approaches are therefore presented in two categories. In the first category, advanced machine learning techniques are utilized to "learn" the aesthetic value of music pieces by themselves – thus considering no ground–truth aesthetic measures. The second category discusses interactive GP approaches. The discussion on works that examine aesthetic fitness criteria for evolutionary systems is addressed in the next section of this chapter.

The first documented approach to **automated music composition** using GP was reported in (Spector & Alpern, 1994). Music generation was based on already existing pieces, which are called "case bases". The core concept of this approach was to combine short "atom" programs (or music functions) that receive a melodic part comprising some note events, then alter these events and return as output the resulting melodic part. For example, the program named as REP takes a melody and returns another one that consists of four replications of the first melodies first bar. Another example is the INVERT program, which receives a melody and returns again a new one, which has the inverted intervals of the first. There are 13 such musical programs–functions that are utilized in the work under discussion. The altered melodies were evolved with fitness values provided by musical criteria inspired by the work presented in (Baker, 1988), which descriptively involve five indications of music and rhythmic novelty and balance. The results of this pioneering GP approach were produced by applying the system on a solo part of Charlie Parker. The authors commented on the results that although the fitness evaluation was found to be a good choice, the resulting music did not actually sound pleasant, a fact that indicates the importance and difficulty of describing aesthetic criteria for music.

The immense difficulty to extract music features that encompass aesthetically meaningful information is a thorough matter of discussion that is partly extended in the second section. The first approach in GP methodologies to tackle the problem of aesthetic definitions in music was performed by avoiding such definitions, through utilizing an artificial neural network (ANN) as an automatic fitness rater, an approach that has proven very popular to some GP techniques that followed. The first work to introduce ANN automatic fitness raters was presented in (Spector & Alpern, 1995), which continues the work discussed in the previous paragraph and also utilizes the GP programs–functions that were previously discussed. Therein, only 4 bar melodies are considered, with each bar including 48 equally spaced positions of potential note placements, i.e. each melody comprised 192 possible note values. The integer value of each of the 192 cells indicated the pitch height of the note, while the zero value indicating no note event. These 192 values were given as input to the ANN, which had two outputs, one indicating whether the presented melody is pleasant and one for the contrary case.

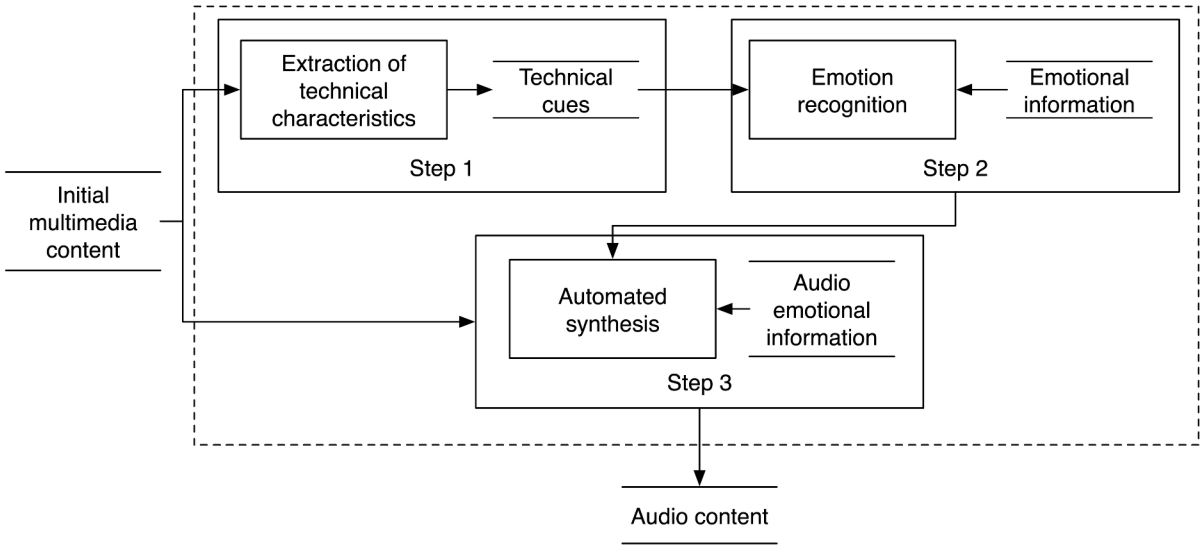
The ANN was trained using four sets of melodies, the first of which included the pleasant paradigms (Charlie Parker solos) and the remaining three included randomly altered versions of the pieces in the first category (constituting the non–pleasant paradigms for the ANN). The trained ANN exhibited the ability to recognize some sense of "musicality" since it judged as good music some Charlie Parker solos (not all of the presented ones) and also a solo by Jimi Hendrix, while it gave negative scores to some almost random sequences. The GP individuals were evolved according to the fitness indication provided by the output of the trained ANN, according to their pleasantness. The reported results were again poor generating "bizarre" rhythmic groupings, as described by the authors, and long interval jumps. According to the authors, this is primarily due to the small number of training samples for the ANN. The results were improved slightly by introducing additional explicit rules in the fitness evaluation scheme, which did penalized melodies which had a good ANN fitness but some obviously unpleasing characteristics. The interested reader is also referred to (Spector, Klein, & Harrington, 2005) for more elaborative comments of the systems described in (Spector & Alpern, 1994) and (Spector & Alpern, 1995).

Another approach to fitness evaluation through an "artificial expert" has been proposed in (Phon-Amnuaisuk, Law, & Kuan, 2007) where instead of an ANN, a self organizing map (SOM) (Kohonen, Schroeder, & Huang, 2001) is utilized as an automatic fitness rater. The GP framework is similar to the ones described hitherto in this section, where simple musical programs act as non–terminal nodes (e.g. which modulate the pitch of all notes in an input melody segment or alter its durations of notes). A typical evolutionary scheme is utilized, where monophonic melodies are represented in MIDI format. The melodies are transformed to pitch class and duration vectors and are imported to a trained SOM. The SOM is trained on a target melody, capturing some of its basic pitch and duration characteristics. Thereafter, each GP individual is mapped onto the trained SOM, thus providing fitness evaluations. The presented results concern evaluations with SOMs trained on small size melodies, while the statistical comments and the small score segments that are provided do not allow a deep understanding of the system's capabilities. For further study on artificial critics without aesthetic background, the interested reader is also referred to (Reddin, McDermott, & O'Neill, 2009) and (McDermott & O'Reilly, 2011).

## POSSIBLE AND PROPOSED APPLICATIONS

Automatic synthesis can be integrated in all aspects of a multimedia application. Although that is not a necessary supposition, a context connection between audio and other multimedia communication chan-

*Figure 4. An abstract illustration of the audio and multimedia content connection in the automated synthesis driven by emotion process*




nels will definitely increase the perceived, by the user, value of the presented information. Thus, in the follow proposed and possible applications the **automated synthesis** process is driven by a prior extraction of information by the actual and non-audio content. A general and abstract illustration of the latter process is in Figure 4. In this Figure, the illustrated process can be organized in three steps:

1. Extraction of technical characteristics from the non-audio multimedia content that serves as the main communication interface to the user. The nature of these cues is application and content specific. For example, if an audio track for news is to be generated then the extracted information will be the descriptors for emotion analysis from text (Calix, Mallepudi, Chen, & Knapp, 2010; Wu & Ren, 2010; Wu, Chuang, & Lin, 2006; KalaiSelvi, Kavitha, & Shunmuganathan, 2014)
2. Emotion recognition from the obtained technical cues by Step 1. According to each application, the extracted cues from previous step will be employed as input to an emotion recognition process that will generate the appropriate emotion description for the content
3. Having all the above information, the last step is the **automated synthesis**. Employing the description of the recognized emotions and the previous knowledge of what and how the technical features of audio can affect the conveyed emotion, the algorithms of **automated synthesis** will generate the final audio content for the multimedia application, as described in the previous sections of this chapter.

The choice of the proposed applications serves a two fold cause; a) the following approaches are believed that cover a wide range of everyday utilization of multimedia, and b) the employment of the proposed integration of **automated synthesis** techniques and emotion recognition will allow the faster creation of proper audio content for areas that, by empirical observation and evidence, provide a swift creation of new constituents. Thus, in the following first sub-section is presented a possible application for the automated generation of soundtrack for movies and videos that will serve as a tool for directors

and movie producers. In the second sub-section is proposed the automated generation of audio content for video games that will allow a real-time interaction between the player and the game machine. The latter will be capable to produce a user and story specific sound track according to the actions and choices that the player made in his game course. Finally, in the third sub-section is proposed an enhancement of the text-based information. The reader of news will be capable to hear an accompanied music or soundscape that will be based on the text native emotion, either recognized by automated procedures or indicated by the news writer.

## Audio Track for Videos and Movies

As written in this chapter, audio is one prominent element of multimedia applications. The other is video. Thus, as a first and possible application for **emotionally enhanced automated synthesis** is the creation of soundtracks for videos and movies. Although that the cinema is a vast industry, in nowadays video making is also possible in a home environment. The advances in digital image processing, embedded systems and electronic sensors have allowed the utilization of high quality hand-held video recorders. Internet platforms, e.g. the well-known youtube[2], exist that host videos from enthusiasts and professionals.

Even in semi-professional situations, a music composer is not always in the video creation team. In addition, pre-recorded musical creations of artists are not always fit in the emotional content of the movie. Thus with the proposed **emotionally enhanced automated synthesis**, a video creator could add musical content to his artistic creations in an automated fashion.

According to Figure 4, such an application would accept as input the video material and possible emotional annotations. The latter could be combined with recognized emotion, by the utilization of existing techniques for video emotion recognition, and serve as input to Step 3. Finally, the proper audio would be created and used as an audio track to the video material. The aforementioned process could be employed to various video contents and applications, e.g. video casts.

## Computer Game Audio Track

The application of **automated synthesis** of **emotionally enhanced music** could be beneficial for computer video games. The current, and most common, status of sound reproduction in video games is the playback of pre-recorded, pre-stored and pre-associated sounds with actions and key-points in the game. Although that these audio clips are connected with the story, plot, scenario and user actions and emotional experience, they are static. Thus, with the adaptation of the aforementioned automatic creation of audio clips, a video game could render in real time a proper audio content based on the user's actions and game's elements.

User's actions and game choices will serve as the technical cues that must be extracted from the multimedia content, according to Figure 4. Story, plot and other scenario based elements of the game and the affective pursuits of the user's progress will be the emotional information which, altogether with the technical cues, will result to the description of player's targeted emotion(s). Again, as the final Step, the **emotional enhanced automated synthesis** will produce the proper audio material in order to reflect the needs of respective situations. Such employment of the proposed framework is believed that will also intensify the immersion of the players by providing a personal and tailor made experience based on their individual game style.

## Enhancement of Text-Based Information

Internet and offline applications for providing text-based news are a common element in the pursuit for briefing of political, cultural, sport and other kind of news in modern culture. Although those applications do offer thoroughly detailed information, they are static and utilize only the visual channel for communication. By employing the proposed framework, an audio content with appropriate emotional matching could be also offered to the readers and thus enhance their overall experience. In an extent, the same application could also be utilized in educational content by providing means for emotional interaction and feedback in the process.

The textual information would serve as the technical cues, regarding Step 1 in Figure 4, and the targeted emotion according to the reader as the emotional information for Step 2. Finally, in Step 3, the proposed integration of **automated synthesis** and **emotion recognition** would provide the proper audio material in order to, on one hand, comply with the targeted emotions by the writer/journalist and, on the other hand, escalate user's involvement and experience.

## CONCLUSION

In this chapter was attempted a proposal for the enhancement of **automated sound and music synthesis** with an affective dimension. As can be seen, the existing level of technology and research in **automated synthesis** can provide a feature driven composition of audio material. In addition, the latter task can also be driven with an interactive scheme by employing GAs and having as targeted state an audio material that will present particular values for specific technical features. The exact values emerging from the audio emotion recognition process, which will also provide the association of those with the targeted listener's emotions.

Thus, a combination of the presented technologies, methods and results can lead to the implementation of an **automated affective sound synthesis**. Such audio material could be employed in various tasks, as depicted from the proposed applications, in order to enhance user experience by adding content or task-oriented dimension, i.e. the emotion.

## REFERENCES

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, *12*(2), 169–177. doi:10.1016/S0959-4388(02)00301-X PMID:12015233

Alfonseca, M., Cebrian, M., & Ortega, A. (2007). A simple genetic algorithm for music generation by means of algorithmic information theory. *Proceedings of IEEE Congress on Evolutionary Computation (CEC 2007)*. Singapore. IEEE doi:10.1109/CEC.2007.4424858

Baker, D. (1988). *David Baker's Jazz Improvisation: A Comprehensive Method for All Musicians*. U.S.A.: Alfred Music Publishing.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self- assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. doi:10.1016/0005-7916(94)90063-9 PMID:7962581

Bradley, M. M., & Lang, P. J. (2007). *The international affective digitized sounds (2nd edition; iads-2): Affective ratings of sounds and instruction manual* [Technical Report B–3]. Gainesville, Fl: NIMH Center for the Study of Emotion and Attention.

Calix, R., Mallepudi, S., Chen, B., & Knapp, G. (2010). Emotion recognition in text for 3-d facial expression rendering. *IEEE Transactions on Multimedia*, *12*(6), 544–551. doi:10.1109/TMM.2010.2052026

Casacuberta, D. (2004). Dj el nino: Expressing synthetic emotions with music. *AI & Society*, *18*(3), 257–263. doi:10.1007/s00146-003-0290-x

Chen, L., Tao, H., Huang, T., Miyasato, T., & Nakatsu, R. (1998). Emotion recognition from audiovisual information. *Proceedings of IEEE Second Workshop on Multimedia Signal Processing*. Redondo Beach, California, U.S.A. IEEE.

Cornelius, R. R. (2000). Theoretical approaches to emotion. In *Proceedings of International Speech Communication Association (ISCA) Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, N. Ireland, U.K: ISCA.

Donahue, C. (2013). *Applications of genetic programming to digital audio synthesis*. [Doctoral dissertation]. University of Texas at Austin, Texas, USA.

Donnelly, P., & Sheppard, J. (2011). Evolving four-part harmony using genetic algorithms. *Proceedings of the 2011 international conference on Applications of evolutionary computation - Volume Part II, EvoApplications'11*. Berlin, Heidelberg. Springer-Verlag.

Drossos, K., Floros, A., & Giannakoulopoulos, A. (2014). Beads: A dataset of binaural emotionally annotated digital sounds. *Proceedings of 5th International Conference on Information, Intelligence, Systems and Applications (IISA 2014)*. Chania, Crete, Greece. IEEE. doi:10.1109/IISA.2014.6878749

Drossos, K., Floros, A., & Kanellopoulos, N. G. (2012). Affective acoustic ecology: Towards emotionally enhanced sound events. *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*. Corfu, Greece. ACM. doi:10.1145/2371456.2371474

Drossos, K., Kotsakis, R., Kalliris, G., & Floros, A. (2013). Sound events and emotions: Investigating the relation of rhythmic characteristics and arousal. *Proceedings of Fourth International Conference on Information, Intelligence, Systems and Applications (IISA 2013)*. Piraeus, Greece. IEEE. doi:10.1109/IISA.2013.6623709

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the kth rule system for music performance. *Advances in Cognitive Psychology*, *2*(2), 145–161. doi:10.2478/v10053-008-0052-x

Garcia, R. A. (2000). Towards the automatic generation of sound synthesis techniques: Preparatory steps. *Proceedings of AES 109th Convention*. Los Angeles. AES.

Garcia, R. A. (2001a). Automating the design of sound syntheses techniques using evolutionary methods. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX- 01)*. Limerick, Ireland.

Garcia, R. A. (2001b). Growing sound synthesizers using evolutionary methods. In *Proceedings of European Conference in Artificial Life 2001 (ECAL2001) special workshop in Artificial Life Models for Musical Applications*. Prague, Chez Republic. Springer.

Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, *49*(10-11), 787–800. doi:10.1016/j.specom.2007.01.010

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, *48*(2), 246–268. doi:10.2307/1415746

Holland, J. H. (1992). *Adaptation in natural and artificial systems*. Cambridge, MA, USA: MIT Press.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *120*(5), 770–814. doi:10.1037/0033-2909.129.5.770 PMID:12956543

KalaiSelvi., R., Kavitha, P., & Shunmuganathan, K. (2014). Automatic emotion recognition in video. *Proceedings of International Conference on Green Computing Communication and Electrical Engineering (ICGC- CEE)*. Coimbatore, India. IEEE

Kaliakatsos-Papakostas, M. A., Epitropakis, M. G., Floros, A., & Vrahatis, M. N. (2012a). Controlling interactive evolution of 8-bit melodies with genetic programming. *Soft Computing*, *16*(12), 1997–2008. doi:10.1007/s00500-012-0872-y

Kaliakatsos-Papakostas, M. A., Epitropakis, M. G., Floros, A., & Vrahatis, M. N. (2013a). Chaos and music: From time series analysis to evolutionary composition. *International Journal of Bifurcation and Chaos (IJBC)*. 23, 1350181–1–1350181–19.

Kaliakatsos-Papakostas, M. A., Floros, A., & Vrahatis, M. N. (2012d). Intelligent real-time music accompaniment for constraint-free improvisation. In Proceedings of *24th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2012)*. Piraeus, Athens, Greece: IEEE. doi:10.1109/ICTAI.2012.67

Kaliakatsos-Papakostas, M. A., Floros, A., & Vrahatis, M. N. (2013b). Intelligent music composition. In X. S. Yang, Z. Cui, R. Xiao, A. H. Gandomi, & M. Karamanoglu (Eds.), *Swarm Intelligence and Bioinspired Computation* (pp. 239–256). Amsterdam, The Netherlands: Elsevier. doi:10.1016/B978-0-12-405163-8.00010-7

Kaliakatsos-Papakostas, M. A., Floros, A., Vrahatis, M. N., & Kanellopoulos, N. (2012b). Genetic evolution of L and FL–systems for the production of rhythmic sequences. *In Proceedings of 21st International Conference on Genetic Algorithms and the 17th Annual Genetic Programming Conference (GP) (GECCO 2012), 2nd Workshop in Evolutionary Music*. Philadelphia, USA. ACM.

Kejun, Z., & Shouqian, S. (2010). Music emotional design by evolutionary algorithms. *Proceedings of IEEE 11th International Conference on Computer-Aided Industrial Design Conceptual Design (CAIDCD)*. Yiwu, China. IEEE.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., & Scott, J., … Turnbull, D. (2010). Music emotion recognition: A state of the art review. Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR 2010). Utrecht, Netherlands.

Koelsch, S. (2010). Towards a neural basis of music-evoked emotions. *Trends in Cognitive Sciences*, *14*(3), 131–137. doi:10.1016/j.tics.2010.01.002 PMID:20153242

Kohonen, T., Schroeder, M. R., & Huang, T. S. (Eds.). (2001). *Self-Organizing Maps* (3rd ed.). Secaucus, NJ, USA: Springer-Verlag New York, Inc. doi:10.1007/978-3-642-56927-2

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: The MIT Press.

Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*. Boston, USA. ACM. doi:10.1145/1518701.1518813

Li, T., & Ogihara, M. (2003). Detecting emotion in music. In *Proceedings of 4th International Symposium on Music Information Retrieval*. U.S.A.

Lo, M. Y. (2012). Evolving Cellular Automata for Music Composition with Trainable Fitness Functions [Doctoral dissertation]. University of Essex, Essex.

Lu, L., Liu, D., & Zhang, H.-J. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 5–18. doi:10.1109/TSA.2005.860344

Manaris, B., Roos, P., Machado, P., Krehbiel, D., Pellicoro, L., & Romero, J. (2007). A corpus-based hybrid approach to music analysis and composition. *Proceedings of 22nd International conference on Artificial intelligence*. Vancouver, British Columbia, Canada. AAAI Press.

McDermott, J., & O'Reilly, U.-M. (2011). An executable graph representation for evolutionary generative music. *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, GECCO '11*. Dublin, Ireland. ACM. doi:10.1145/2001576.2001632

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, *97*(3), 315–331. doi:10.1037/0033-295X.97.3.315 PMID:1669960

Ozcan, E., & Eral, T. (2008). A genetic algorithm for generating improvised music (pp. 266–277). In Monmarch, N., Talbi, E.-G., Collet, P., Schoenauer, M., & Lutton, E. (Eds.). Artificial Evolution, volume 4926 of Lecture Notes in Computer Science. Berlin/Heidelberg: Springer.

Papadopoulos, G., & Wiggins, G. (1998). A Genetic Algorithm for the Generation of Jazz Melodies. *Proceedings of Finnish Conference on Artificial Intelligence (STeP)*. Jyväskylä.

Phon-Amnuaisuk, S., Law, E., & Kuan, H. (2007). Evolving music generation with som-fitness genetic programming. In M. Giacobini (Ed.), Lecture Notes in Computer Science: Vol. 4448. *Applications of Evolutionary Computing* (pp. 557–566). Berlin, Germany: Springer. doi:10.1007/978-3-540-71805-5_61

Putnam, J. B. (1994). Genetic programming of music. Retrieved from ftp://ftp.cs.bham.ac.uk/pub/tech-reports/1997/CSRP-97-07.ps.gz

Reddin, J., McDermott, J., & O'Neill, M. (2009). Elevated pitch: Automated grammatical evolution of short compositions. In M. Giacobini, I. De Falco, & M. Ebner (Eds.), *Applications of Evolutionary Computing, EvoWorkshops2009* (Vol. 5484, pp. 579–584). Tubingen, Germany. Springer Verlag. doi:10.1007/978-3-642-01129-0_65

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. doi:10.1037/h0077714

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression. *Proceedings of International Conference on Multimedia Information Retrieval, MIR '10*. Philadelphia, PA, USA. ACM. doi:10.1145/1743384.1743431

Schuller, B., Hantke, S., Weninger, F., Han, W., Zhang, Z., & Narayanan, S. (2012). Automatic recognition of emotion evoked by general sound events. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan. IEEE doi:10.1109/ICASSP.2012.6287886

Smith, J. O. (2010). *Physical audio signal processing: for virtual musical instruments and audio effects*. Stanford: W3K Publishing.

Spector, L., & Alpern, A. (1994). Criticism, culture, and the automatic generation of artworks. In *Proceedings of the twelfth national conference on Artificial intelligence (*vol. 1*), AAAI '94*. Seattle, Washington, USA. American Association for Artificial Intelligence.

Spector, L., & Alpern, A. (1995). Induction and recapitulation of deep musical structure. *Proceedings of the IFCAI-95 Workshop on Artificial Intelligence and Music*. Montreal, Quebec, Canada.

Spector, L., Klein, J., & Harrington, K. (2005). Selection songs: Evolutionary music computation. *YLEM Journal*, *25*(6 & 8), 24–26.

Storn, R., & Price, K. (1997). Differential evolution – a simple and efficient adaptive scheme for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359. doi:10.1023/A:1008202821328

Sugimoto, T., Legaspi, R., Ota, A., Moriyama, K., Kurihara, S., & Numao, M. (2008). Modeling affective-based music compositional intelligence with the aid of ANS analyses. *Knowledge-Based Systems*, *21*(3), 200–208. doi:10.1016/j.knosys.2007.11.010

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., & Scherer, K. R. (2013). On the acoustics of emotion in audio: What speech, music and sound have in common. *Frontiers in Psychology*, 4. PMID:23750144

Wieczorkowska, A., Synak, P., Lewis, R., & Ra, W. Z. (2005). Extracting emotions from music data. In Hacid, M.-S., Murray, N., Ra, Z., & Tsumoto, S. (Eds.), Foundations of Intelligent Systems (LNCS) (Vol. 3488, pp. 456–465). Berlin Heidelberg: Springer.

Wilson, R. S. (2002). First steps towards violin performance extraction using genetic programming. In J. R. Koza (Ed.), *Genetic Algorithms and Genetic Programming at Stanford 2002*. Stanford, California, USA: Stanford Bookstore.

Wu, C.-H., Chuang, Z.-J., & Lin, Y.-C. (2006). Emotion recognition from text using semantic labels and separable mixture models. [TALIP]. *ACM Transactions on Asian Language Information Processing*, *5*(2), 165–183. doi:10.1145/1165255.1165259

Wu, Y., & Ren, F. (2010). Improving emotion recognition from text with fractionation training. *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*. Beijing, China. IEEE. doi:10.1109/NLPKE.2010.5587800

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2008). What is the best segment duration for music mood analysis? *Proceedings of* International Workshop on *Content-Based Multimedia Indexing (CBMI 2008)*. London, U.K. IEEE.

Yang, Y. H., Lin, Y. C., Cheng, H. T., & Chen, H. H. (2008). Mr. emo: music retrieval in the emotion plane. *Proceedings of 16th ACM international conference on Multimedia*. Vancouver, Canada: ACM. doi:10.1145/1459359.1459550

Yeh, C. H., Lin, H. H., & Chang, H. T. (2009). An efficient emotion detection scheme for popular music. *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS 2009)*. Taipei, Taiwan. IEEE. doi:10.1109/ISCAS.2009.5118126

Zhang, S., Tian, Q., Jiang, S., Huang, Q., & Gao, W. (2008). Affective mtv analysis based on arousal and valence features. *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2008)*. Hanover, Germany. IEEE. doi:10.1109/ICME.2008.4607698

Zhu, H., Wang, S., & Wang, Z. (2008). Emotional music generation using interactive genetic algorithm. *Proceedings of International Conference on Computer Science and Software Engineering* (volume 1). Wuhan, Hubei, China. IEEE. doi:10.1109/CSSE.2008.1203

## KEY TERMS AND DEFINITIONS

**Affective Acoustic Ecology:** An enhanced concept of the Acoustic Ecology that contains also the emotional interaction between the listeners and the audio environment.

**Affective Sound/Music:** Sound or music that is capable to elicit emotions to listener.

**Automated Music Synthesis/Composition:** The automated synthesis/composition of music by utilizing proper algorithms.

**Emotion Recognition From Sound/Music/Audio/Sound Events:** The process of automatically identifying the emotion that the listeners of a sound, music sound event or in general audio stimulus will feel.

**Emotionally Annotated:** Something that also contains emotion annotations (labels) for it.

**Emotionally Enhanced:** Something that is created or altered in order to elicit specific emotion(s) to its recipients/receivers.

**Intelligent Music/Sound Synthesis:** The automated synthesis of music/sound that also automatically adapts to specific pre-requisites.

**Sound Event (SE):** A generalized audio stimulus, non-linguistic and non-musical.

## ENDNOTES

1 http://www.findsounds.com

2 http://www.youtube.com