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