Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines

Cyril Laurier, Perfecto Herrera
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
{cyril.laurier,perfecto.herrera}@upf.edu

Abstract
Creating emotionally sensitive machines will significantly enhance the interaction between humans and machines. In this chapter we focus on enabling this ability for music. Music is extremely powerful to induce emotions. If machines can somehow apprehend emotions in music, it gives them a relevant competence to communicate with humans. In this chapter we review the theories of music and emotions. We detail different representations of musical emotions from the literature, together with related musical features. Then, we focus on techniques to detect the emotion in music from audio content. As a proof of concept, we detail a machine learning method to build such a system. We also review the current state of the art results, provide evaluations and give some insights into the possible applications and future trends of these techniques.

Introduction
Why do people enjoy music? One of the main factors is that music easily induces emotions and affects the listener. Can machines enjoy music as people do? Or, surely easier and less philosophically debatable, can we develop systems that are capable of detecting emotions in music and use this ability to improve human-machine interaction?

Stating that music and emotions have a close relationship is no revelation. One significant motivation of composers is to express their sentiments, performers to induce feelings, and listeners to feel emotions. There are also some findings that show a direct link between audio processing and emotion in the brain, which is not so clear for other sensory modalities. Moreover music has a noticeable social role and is ubiquitous in everyday life. To communicate with humans using musical emotions, machines should be able to detect and predict them. Enabling this ability will enhance the communication between the machine and the environment. First they can sense the acoustic scene with a microphone. The surrounding music can be understood in terms of emotions and the machine can react accordingly. The face of a robot can give an explicit feedback of the emotions it detects. Moreover robots with musical abilities can select, play and even compose music conveying targeted emotions. The technology we detail in the remainder of this chapter enables machines to detect emotions from raw audio material, which is directly from the digital signal. In this chapter we expose the main findings about music and emotions, together with techniques in artificial intelligence and more explicitly in machine learning to create emotionally sensitive machines.

This chapter is structured in four parts. In the first section we comment on the relationship between emotion and music and review theories from different expertise. In section 2 we define the machine learning techniques that can be used to create emotion aware machines; we detail also the methodology and give evaluation results from state of the art research in this area. Then, in section 3, we develop some ideas around emotion-based music assistants. Finally, in the last part, we present some general observations and give future perspectives.

Section 1. Music and emotions: emotion in music & emotions from music

To study the relationship between music and emotion, we have to consider the literature from many fields. Indeed, relevant scientific publications about this topic can be found in psychology, sociology, neuroscience, cognitive science, biology, musicology, machine learning and philosophy. We focus here on works aiming to understand the emotional process in music, and to represent and model the emotional space. We also detail the main results regarding the pertinent musical features and how they can be used to describe and convey emotions.

Why does music convey emotion?

Emotion and expressive properties of musical elements have been studied since the time of ancient Greece (Juslin and Laukka, 2004). The fact that music induces emotions is evident for everyone. However we do not intuitively apprehend why. Emotions are mostly said to be complex and to involve a complicated combination of cognition, positive or negative feeling changes, appraisal, motivation, autonomic arousal, and bodily action tendency or change in action readiness.

One of the first things to clarify is the definition of an emotion and the difference between emotions and moods. The concept of emotion is not simple to define: "Everyone knows what an emotion is, until asked to give a definition" (Fehr and Russell, 1984, p. 464). It could be defined as an intense mental state arousing the nervous system and invoking physiological responses. According to Damasio (1994), emotions are a series of body state changes that are connected to mental images that have activated a given brain subsystem (e.g., the music processing subsystem). So emotions involve physiological reactions but also they are object-oriented and provoke a categorization of their object: “if the emotion is one of fear its object must be viewed as harmful” (Davies, 2001, p. 26). Emotions also induce an attitude towards the object. Moods could be considered as lasting emotional states. They are not object oriented and take into account quite general feelings. Moods and emotions can be very similar concepts in some cases, for instance happiness, sadness and anger can be seen as both moods and emotions. However some emotions can only be considered as transient, such as surprise.

Understanding how music conveys emotion is not trivial. Kivy (1989) gives two such hypotheses. The first might be a “hearing resemblance between the music and the natural expression of the emotion”. Some musical cues can induce emotions because of their similarity to speech. One example is “anger” where the loudness and the spectral dissonance (derived from frequency ratios and harmonic coincidence in the sound spectrum and based on psychoacoustic tests) are two components we can find in both an angry voice and music. However it might not always be that simple. The second hypothesis Kivy gives is the “accumulated connotations a certain musical phenomena acquire in a culture”. In that case, we learn in our culture which musical cues correspond to which feeling. Most probably, both hypotheses are valid. Frijda (1987, pp. 469) argues for a notion of emotions as action tendencies where “various emotions humans or animals can have - the various action readiness modes they may experience or show - depends upon what action programs, behavior systems, and activation or deactivation mechanisms the organism has at its disposal.”. As pointed out by Nussbaum (2007), this correlates with results in neuroscience from scientists such as Damasio (1994).

Grewe et al. (2007) demonstrated that the intensity of the emotion induced by music could vary depending on personal experience and musical background. If a musician knows and has studied the piece for a performance, he/she is more likely to rate the intensity of the emotion higher. This is an auto-reinforcement by training. We can also imagine that listening to a musical piece too many times can create the opposite behavior. Almost everyone has experienced the fact of being bored, or less and less sensitive to a musical piece they used to love. Besides, it is important to notice that emotions in

music are not restricted to adults or musically trained people. The emotional processing of music starts at an early age. Four-months-old children have a preference for consonant (pleasant) over dissonant (unpleasant) music (Trainor, Tsang and Cheung, 2002). At five years old, they can distinguish between happy and sad music using the tempo (sad = slow, happy = fast), but at six, they use information from the mode (sad = minor, happy = major) such as adults do (Dalla Bella et al., 2001).

Studies in neuroscience, exploiting the current techniques of brain imaging also give a hint about the emotional processing of music, with some schemas of the brain functions involved (Koelsch et al., 2006). Gosselin et al. (2005) demonstrated that the amygdala, well established to have an important role in the recognition of fear, is determinant in the recognition of scary music. Blood and Zatorre (2001) revealed that music creating highly pleasurable experience like "shivers-down-the-spine" or "chills" activate regions in the brain involved in reward and motivation. It is worth noticing that these areas are also active in response to other euphoria-inducing stimuli like food, sex and drugs. Huron (2006) simply states that music making and listening are primarily motivated by pleasure and that the contrary is biologically implausible (p. 373). Meyer (1956) describes the importance of expectation as a tool for the composer to create emotions. This work has been continued and formalized as the ITPRA[1] theory by Huron (2006). One important way to control the pleasure in a musical piece is to play with this feature by delaying expected outcomes and fulfilling our expectation.

Additional research (Menon and Levitin, 2005) seems to have also found the physical connections between music and mood alteration by means of antidepressants: the latter act on the dopaminergic system which has one of its main centers in the so-called *nucleus accumbens*, a brain structure that also receives a dramatic degree of activation when listening to music. These results are coherent with Lazarus (1991), when he argues that emotions are evolutionary adaptations, to evoke behaviors that improve chances for survival and procreation, and with Tomkins' (1980) view that emotions can be understood as "motivational amplifiers". It links music with survival related stimuli. Often, damages to emotional controls limiting the normal functionability of the emotional behavior are disastrous for people (Damasio, 1994). Moreover people who did not develop social emotions seem incapable of appreciating music (Sacks and Freeman, 1994). However, this evolutionary adaptation theory can be balanced by the fact that most emotional responses to music are neither used to achieve goals, nor practically related to survival issues. This argument is used by researchers who assume that music cannot induce basic survival emotions, but more "music-specific emotions" (Scherer and Zentner, 2001, p. 381). Nonetheless, other notable researchers affirm about music that it is "remarkable that any medium could so readily evoke all the basic emotions of our brain" (Panksepp and Bernatzky, 2002). This is one of the multiple contradictions we can observe in current research on music and emotions. As pointed out by Juslin and Västfjäll (2008), the literature presents a confusing picture with conflicting views. Nevertheless there is no doubt that music induces emotion because of the related context. It evokes emotions from past events because it is associated in our memory to emotional events.

When talking about emotion and music, one important distinction to make is the difference between induced and perceived emotions (Juslin and Laukka, 2004). That is what we define as "emotion in music" and "emotion from music". The former represents the intended emotion and the latter the emotion felt while listening to a musical piece. A typical example of differentiation between both is the expression of anger. When someone is angry, people might perceive anger but feel scared or defensive. The induced emotion is radically different from the perceived one. Different factors can influence both

[1] ITPRA stands for : Imagination response, Tension response, Prediction response, Reaction response, Appraisal response (Huron, 2006, pp. 357-365)

types, for instance the symbolic aspect or the social context of a song will influence more the induced emotion (like for a national anthem). As noticed by Bigand et al. (2005) both aspects are not strictly independent and there will always be an influence of the induced emotion on someone asked to judge the perceived one. Nevertheless it should be observed that people tend to agree more on the perceived emotion than on the induced emotion (Juslin and Laukka, 2004).

It is worth noticing that a relevant part of the emotion in songs comes from the lyrics. Psychological studies have shown that part of the semantic information of songs resides exclusively in the lyrics (Besson et al., 1998). This means that lyrics can contain relevant information to express emotions that is not included in the audio. Indeed, Juslin and Laukka (2004) reported that 29% people mentioned the lyrics as a factor of how music expresses emotions.

Although there is an increase in research about the causal links between music and emotion, there still remain many open questions (Patel, 2007). In addition to the biological substrate, there are important links related to the musical features that are present or absent when perceiving or feeling a given music-related emotion. In section 2, we give some results about these musical features, but first we will discuss the different representations of musical emotions that arise from psychological studies.

Emotional Representations

One main issue in making machines emotionally sensitive is to find models of human representation of emotion in music. From the literature in music psychology, there exist two main paradigms to represent emotions. This distinction is quite general, it is not only about musical emotions, but studies were designed specifically to test and refine these models for music. The first one is the categorical representation that distinguishes among several emotion classes. The other one is the dimensional representation defining an emotional space. We detail here the main theories using both approaches and we make explicit the special case of musically-related emotional representations.

Categorical representation

The categorical representation aims to divide emotions in categories, where each emotion is labeled with one or several adjectives. The most canonical model is the concept of basic emotions where several distinct categories are the basis of all possible emotions. This concept is illustrated by Ekman's basic emotion theory distinguishing between anger, fear, sadness, happiness and disgust (Ekman, 1992). Nevertheless other categorical approaches are possible. Indeed a lot of psychologists propose that their emotion adjective set is applicable to music. One of the most relevant works in this domain is the study by Hevner (1936) and her adjective circle shown in figure 1. Hevner's adjective list is composed of 67 words arranged into eight clusters. From this study each cluster includes adjectives that have a close relationship. This similarity between words of the same cluster enables one to work at the cluster level reducing the taxonomy to eight categories. Farnsworth (1954) modified Hevner's list into ten clusters. These categories were defined by conducting listening tests and subjective answers. Moreover, we should note that most of these studies were conducted using classical music from the western culture and mainly of the baroque and romantic periods. We can imagine that the emotions evoked by popular music are different. A problem of the categorical approach is that classifying a musical piece into one or several categories is rather difficult sometimes, as pointed out by Hevner (1936). For instance in one of her studies, based on a musical piece called "Reflections on the water" by Debussy was rated to belong to all the clusters unless a continuous measure was considered. Although it was argued that a word list couldn't describe the variety of possible emotions in music, using a reduced set helps to achieve an agreement between people (even if it gives less meaning) and offers the possibility for automatic systems to model the general consensus of musical pieces.

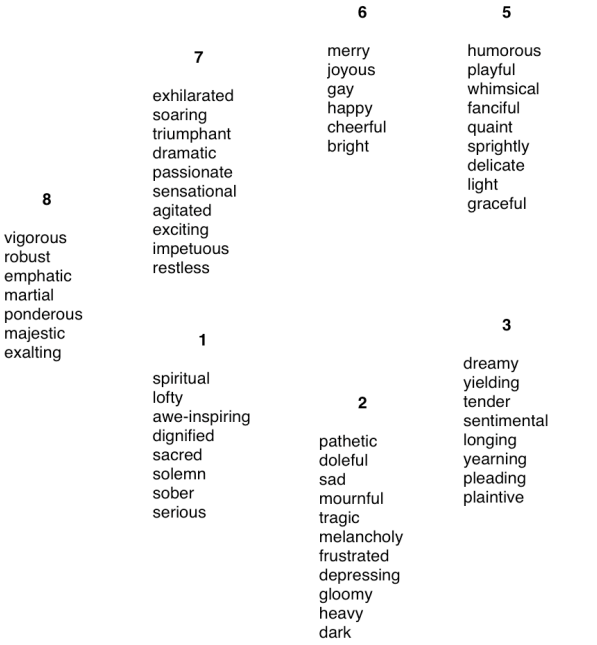

Figure 1. Adjectives and clusters, adapted from Hevner (1936)

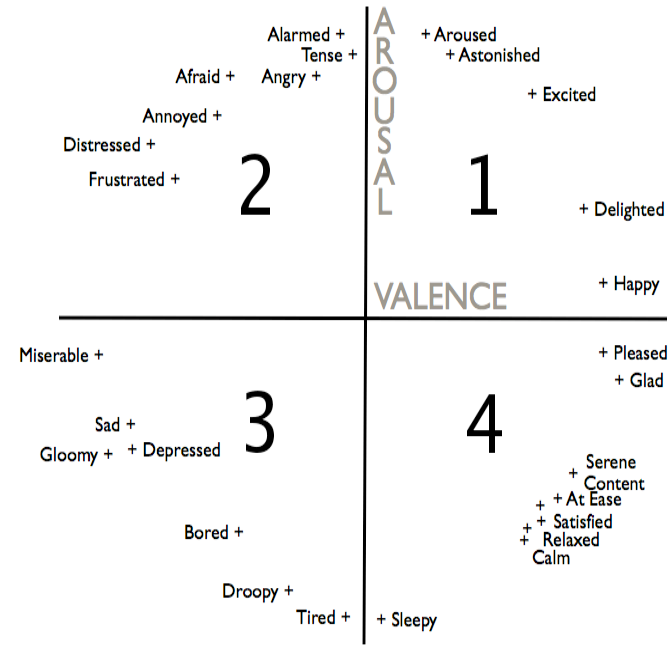

Figure 2. "Circumplex model of affect", adapted from Russel (1980)

Dimensional representation

In a dimensional representation, the emotions are classified along axes. Most of the proposed representations in the literature are inspired by the Russell (1980) "circumplex model of affect", using a two-dimensional space spanned by arousal (activity, excitation of the emotion) and valence (positivity or negativity of the emotion). In figure 2, we represent this bipolar model with the different adjectives placed in this emotional space. In this two-dimensional space, a point at the upper-right corner has high valence and arousal, which means happy with a high activity such as "excited". Opposite to this one, the lower-left part is negative with low activity like "bored" or "depressed". Several researchers such as Thayer (1989) applied this dimensional approach and developed the idea of an energy-stress model. Other studies propose other dimensional representations. However they all somehow relate to the models previously presented, as in the case of Schubert's (1999) two-dimensional emotion space (called 2DES), with valence on the x-axis and arousal on the y-axis with a mapping of adjectives from different psychological references. The main advantage of representing emotion in a dimensional form is that any emotion can then be mapped in that space. It allows a model where any emotion can be represented, within the limitation of these dimensions. One common criticism of this approach is that very different emotions in terms of semantic meaning (but also in terms of psychological and cognitive mechanisms involved) can be close in the emotional space. For instance, looking at the "circumplex model of affect" in figure 2, we observe that the distance between "angry" and "afraid" is small although these two emotions are quite different.

Nevertheless, if both categorical and dimensional approaches are criticized and not perfect, both are used and can be considered as valid, as partial evidence for and against each one can be found in the available experimental literature.

Musical features and emotion

Several studies investigated musical features and their relations to particular emotions. However most

of the available research is centered on the western musical culture and mainly from classical music. Note that both composers and performers use these musical features. In table 1, we report the main mapping between musical features and emotion categories found in the literature (Juslin and Laukka, 2004). Each independent feature is probably not sufficient to conclude about one emotion; on the contrary this may require a rich set of musical descriptors. It is interesting to notice that these features correlate with research made on speech by Scherer (1991, p. 206). Of course the comparison is limited to only a small set of attributes useful for speech like the pitch, the loudness and the tempo.

From the list shown in table 1, we observe that some features can be automatically extracted from polyphonic audio content (like commercial CD tracks or mp3 files) with existing technology[2]. These features are marked with an asterisk. For instance the tempo can be estimated by locating the beats. Of course it would work better on music with evident tempo and prominent percussion on beats (rock or techno for example). The results are less reliable for music with a smooth and subtle rhythm (some classical music, for instance). From audio content the reliability of these features is not always optimum but still it makes sense to use them, as they are informative. The key and the mode can also be extracted with a satisfying correctness (Gómez, 2006) by analyzing frequency distributions and comparing with tonal profiles. Other attributes are more difficult to extract from a complex mix of instruments and would be reliable only on monophonic tracks (one instrument). They are marked with two asterisks in table 1. For example the vibrato or the singer formant changes can be detected if we work on audio information containing just the singer's voice, but it becomes too complex on a mix containing all the instruments. From these results, can we seriously think about automatically predicting the emotion from music? Can machines have an emotional understanding close to ours? Depending on the information an automatic system can get from the environment the answer may vary. It is clear that an audio signal taken from a microphone and a musical score give very different information. In the recent years, research in machine learning and signal processing has allowed one to extract relevant and robust high-level musical features with techniques we will detail in the next section.

[2] For a review on automatic extraction of audio features see Herrera et al. (2005) and Gouyon et al. (2008).

Musical Features	Happiness (1)	Sadness (3)	Anger (2)	Fear (2)	Tenderness (4)
Tempo*	Fast, small variability	Slow	Fast, small variability	Fast, large variability	Slow
Mode*	Major	Minor	Minor	Minor	Major
Harmony*	simple and consonant	dissonant	atonality, dissonant	dissonant	consonant
Loudness*	medium-high, small variability	low, moderate variability	high, small variability	low, large level variability, rapid changes	medium-low, small variability
Pitch**	high, much variability, wide range, ascending	low, narrow range, descending	high, small variability, ascending	high, ascending, wide range, large contrasts	low, fairly narrow range
Intonation**	rising	flat, falling	accent on tonally unstable notes	-	-
Singer's formant**	raised	lowered	raised	-	lowered
Intervals**	perfect 4th and 5th	small (minor 2nd)	major 7th and augmented 4th	-	-
Articulation**	staccato, large variability	legato, small variability	staccato, moderate variability	staccato,large variability	legato, small variability
Rhythm*	smooth and fluent	ritardando	complex, sudden changes, accelerando	jerky	-
Timbre*	bright	dull	sharp	soft	soft
Tone attacks**	fast	slow	fast	soft	slow
Timing variability*	small	large (rubato)	small	very large	moderate
Vibrato**	medium-fast rate, medium extent	slow, small extent	medium-fast rate, large extent	fast rate, small extent	medium fast, small extent
Contrast between long and short notes**	sharp	soft	sharp	-	soft
Micro-structure*	regularities	irregularities	irregularities	irregularities	regularities
Others		pauses	spectral noise	pauses	accents on tonally stable notes

Table 1. *The most frequent musical features mapped with the emotion categories based on Juslin and Laukka (2004). An asterisk (*) means that some information can be extracted from polyphonic audio content; two asterisks (**) means that it can be extracted only from monophonic audio content (one instrument), in both cases using state-of-the-art technology. In parenthesis is the quadrant number in Russell's dimensional space (see figure 1).*

Section 2. Music Information Retrieval: Building automatic detectors of music emotions

Several studies have demonstrated that musical emotions are not too subjective or too variable to deserve a mathematical modeling approach (Bigand et al., 2005; Juslin and Laukka, 2004; Krumhansl, 1997; Peretz, Gagnon and Bouchard, 1998). Indeed, within a common culture, the emotional responses to music can be highly consistent within and between listeners, but also accurate, quite immediate and precocious (Vieillard et al., 2008). This stated, it opens the door to reproduce this consistent behavior with machines.

In this section we give a technical explanation of how to build a system to automatically detect musical emotions from audio. To achieve this goal, we use machine-learning techniques and more specifically supervised learning methods. The overall idea of supervised learning is to learn by example. It requires that the system is presented with enough examples of a given emotional category. We focus here on the categorical representation because it seems easier for people to categorize using simple emotions rather than to give a value for each dimension (arousal, valence). An important part of the work is to gather a substantial amount of reliably labeled examples (called ground truth). Then we extract acoustical and musical information (called features) from the audio of each example file, and finally we learn the mapping between the features and the labels (emotions in our case). This mapping is validated using cross-validation[3] methods or an independent test database. These methods ensure that our system can build general models of the emotional classes (i.e., that the model is not overfitting to the training data). Using this procedure, along with standard automatic classifiers, we can build a system able to reliably and consistently predict the emotion in music to a certain extent.
This type of methodology is part of the research conducted by the Music Information Retrieval (MIR) community. The mostly studied problem in this field is genre classification (Tzanetakis and Cook, 2002; Guaus and Herrera, 2006). However recent trends focus on emotion or mood detection. We review and compare the existing systems to our approach at the end of this section.

If we can work on a symbolic representation (like the musical score, a MIDI file or other), we can use accurate representations of the melody, chords, rhythm and other musical dimensions. It allows generating new versions of the music modifying the emotional content in a more flexible and efficient way than from audio content. Indeed one can operate directly on the relevant musical aspects like Fridberg, Bresin and Sundberg (2006). In our system, we want to deal directly with the audio signal, as we cannot always have access to symbolic information. On one hand we loose the precision in notes and measure mentioned before but on the other we can process the vast amount of musical data available in a digital audio format. Although it seems more complicated, it corresponds to a realistic usage. The machine can then analyze any kind of music from audio files but also from the sonic environment using a microphone.

Methodology

To detect emotions in music, we are using statistical classification. Classification algorithms need a large amount of labeled examples, but also a good and rich musical description of each example, in order to learn how to classify it properly. The information gathered from the examples are numerical data called features (or descriptors). They are computed directly from the audio signal and can describe

[3] In order to assess the ability that the system has to predict a label for new and unseen music files, the training of the system uses only a portion of all the available data, and its testing is done using the remaining data. In n-fold cross-validation, the data is split into n portions, n-1 folds are used for training, and the remaining fold is used for testing. This is done n times, each one using one of the n folds for testing and the remaining folds for training; finally an average of the n tests is used to estimate the mean error of the classification system.

different aspects like for instance timbre, rhythm or tonality[4]. With this information and enough realistic data, the classifier can learn from simple rules to complex functions to predict the emotional label of any new music. We specify here each step of this approach and summarize it in figure 3.

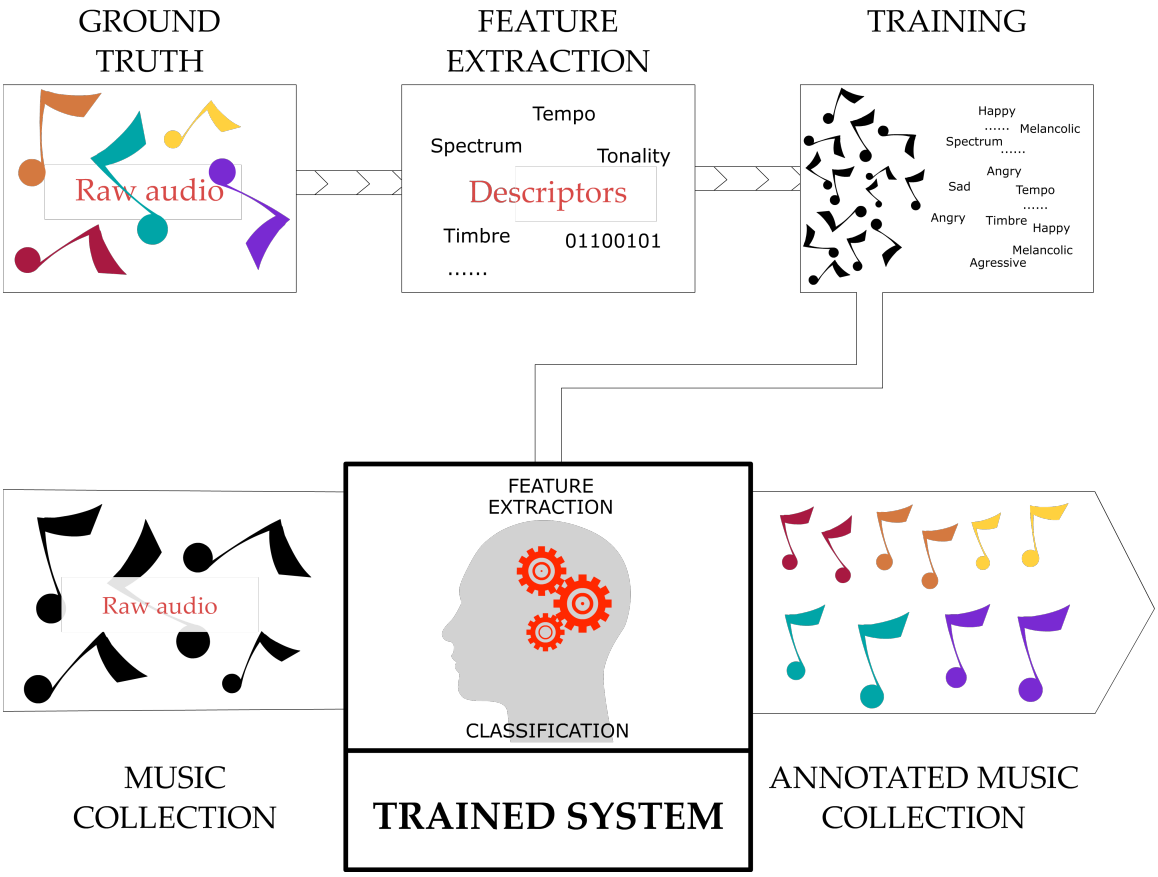

Figure 3. Schema of the supervised learning approach. From the manually annotated ground truth, features are extracted to train a classifier. This trained system can then annotate automatically any new music collection.

Ground Truth

The first step is to create the database of examples. In the case of emotion, the representation chosen will determine the rest of the process. From music and emotion theories, and from psychological studies mentioned in section 1, we can understand the pros and cons of each representation for this purpose. In the MIR field, the representation most often chosen is the categorical approach based on basic emotion theories. Each emotion is considered as independent from the others and all existing emotions would be a combination of these basic emotions. In that case the categories are considered mutually exclusive. This categorization fits particularly well in the automatic classification way of thinking, were we have several classes and one element can belong to only one class. Other studies consider this approach too restrictive as emotions are more complex and because one piece of music can evoke both happiness and sadness at the same time.

For our experiments, we decided to use a categorical approach to ease the process of annotating the

[4] Even though it plays a crucial role in any music cognition aspect, melodic information is still out of scope of the current state-of-the-art automatic music content description. It can only be addressed very roughly or unreliably when polyphonic music files are analysed. Even with this limitation, the current audio descriptors can deal with many practical applications, such as the one we describe here.

data, making it clearer for the people involved in that process. However choosing one type of representation does not totally solve the problem. Deciding categories is not trivial. Each theory of basic emotion in music gives a different set of emotions. To decide on which taxonomy to use, and to study the overlap between categories, we have conducted a small experiment. We asked 16 people to annotate 100 musical excerpts choosing one or several adjectives in a set. The results showed that already on a simple set and a few people, it was not easy to come to an agreement. Moreover some categories are difficult to take into account separately. From the literature and our preliminary experiments, we decided to use a simple approach based on emotion categories well distinguished by people (Laurier and Herrera, 2007). This allows one to have the best agreement between people when labeling, and to make the system as general as possible. We also decided to have a binary approach. Each category is considered to be boolean, for instance a song is "happy" or "not happy". With this approach we have multiple binary classifiers, one for each emotion (instead of a single multi-class classifier). This avoids the strict separation of so-called basic emotions as if they would be mutually exclusive. In fact, this approach is closer to the theory considering that each emotion can be a combination of basic emotions. Therefore we consider that we have an expert for each basic emotion which will estimate the amount of this particular emotion in a given music file. This allows for a more detailed description using an ensemble of multiple boolean experts.

Once we have chosen a proper representation, we need to build the database of examples. In our case the examples are musical pieces labeled with emotions. This step is very time consuming, because people have to listen and manually annotate music. Moreover we want to have as many annotations as possible and on a large amount of musical examples. There are several ways to gather this data. The main method used is a questionnaire, either web-based or in laboratory settings to have more control on the factors that can influence the annotation. The effect of using web-based experiments instead of laboratory settings for musical perception studies is discussed in Honing and Ladinig (2008). Another way is to conduct games to gather this data. Kim, Schmidt and Emelle (2008) created a flash-based game using the dimensional paradigm called *MoodSwings*. In the arousal-valence plane the users are marking the perceived emotion in the music and get points if they agree. Mandel and Ellis (2007) invented a web-based game using the category paradigm. This game is not limited to mood but open to any music labeling. When different users use the same tag to define a musical piece, they get points. These online games are useful to gather much more data than asking people to annotate with no special motivation.

In all cases, several issues have to be addressed. Many different factors can have an impact on the annotation reliability. On one hand, in laboratory settings it is easier to control these factors than when using web-based interfaces. On the other hand in a laboratory environment one might not react as if he was in everyday conditions. Beyond these considerations, in the case of emotion in music, several factors also have to be controlled. Indeed, the emotion in the music depends on many different elements, such as the cultural background, the social context, the lyrics, the temporal evolution of the music, or the personal preferences. The cultural background could mean the experience one has with music. Considering mainstream popular music from the western culture, we can limit the cultural impact so that it would work for many people (but maybe not with people not exposed to western popular music). Information about the social context is by definition not included in the music itself but relies on the context of the music. This is particularly difficult to control as one might have a very personal relationship with one musical piece. That is also one reason to focus our system on detecting emotions in the music and not from the music. Indeed the induced emotion can be quite different from the one perceived, especially because of the social context or the personal history of the listener with that particular music. Moreover, in the annotation process, we can limit this influence by checking if the annotator knows the music. Finally, as to the effect of lyrics, one possibility is to use instrumental

music; we can also reduce the song to a short excerpt so that the whole meaning of the lyrics can not influence the annotation process. Although all these factors are important in the way a ground truth is constructed, they are almost never mentioned in the current MIR literature.

In our case, we have built a ground truth of popular music, with four categories: “happy”, “sad”, “angry” and “relaxed”. We have chosen these emotions because they are related to basic emotions from psychological theories and also because they cover the four parts of the 2D valence/arousal representation. But as we also do not want to restrict to exclusive categories, we consider the problem as a binary classification for each term. One song can be “happy” or “not happy”, but also independently “angry” or “not angry” and so on.

Our collection is made of popular music pre-selected from a large online community (Last.fm[5]), which is active in associating labels (tags) with the music they listen to. We looked for the songs mostly tagged with our categories and synonyms and we asked a small group of listeners in our lab to validate this selection. We included this manual confirmation in order to exclude songs that could have been wrongly tagged, to express something else, or because of a "following the majority" type of effect. The annotators were asked to listen to 30 seconds of the songs, first to avoid as much as possible changes in the emotion, then to reduce the influence of the lyrics and finally to speed up the annotation process. In total 17 different evaluators participated and the final database is composed of 1000 songs divided between 4 categories of interest plus their complementary categories (“not happy”, “not sad”, “not angry” and “not relaxed”).

Feature Extraction

If early MIR systems were able to process only symbolic data like MIDI (symbolic musical standard, which provides a score-like music representation), the evolution of Digital Signal Processing (DSP) techniques have provided new tools to extract audio features. DSP techniques combined with perceptual and musical knowledge allow us to compute descriptors about timbre, rhythm, harmony, loudness or pitch.

An audio file or stream is digitally represented as a waveform, basically a succession of values between -1 and 1 with a rate of several thousand values per second. Typically (as with the Compact Disc format) we consider 44100 values per second for psychoacoustic reasons. In the last decade MIR researchers have been very active in extracting meaningful information from this raw data. Several levels of abstraction can be addressed, from low level (close to the signal) to high level (semantic level, like musical concepts). Taking advantage of expertise in signal processing, psychoacoustic, musicology, statistics, machine learning and information retrieval many descriptors have been proposed (Herrera et al., 2005; Gouyon et al., 2008 pp. 83-160) Each descriptor can be computed as a series of values for a time window and summarized for the entire music file using statistical measures like the mean or the variance. It can also be directly computed as a global value corresponding to a song (like the estimation of the key and mode, e.g. C major, for instance).

Some widely used descriptors are the Mel-Frequency Cepstral Coefficients (MFCCs) (Logan, 2000), because they are very informative about the timbre of the acoustic signal. This type of spectral descriptor is useful to classify music by genre and many other tasks. Another example is the Harmonic Pitch Class Profiles (HPCP) from Gómez (2006) or chroma features for tonality. They describe how the energy in the audio is spread over the notes. It allows estimations of the chord, the key and, with an appropriate algorithm they can be used to detect different versions of the same song (Serrà et al., 2008).

[5] http://www.last.fm

For the mood detection, many features are relevant. It is important to keep in mind that we use these techniques to extract information of a different kind: timbral (for instance MFCCs, spectral centroid), rhythmic (for example tempo), tonal (like HPCP) and temporal descriptors. Among others we have also an estimation of the dissonance, the mode, the onset rate and the loudness. Not all the musical features detailed previously in section 1 can be accurately retrieved from audio content only. Nevertheless, these audio descriptors studied and developed by MIR researchers are sufficient to model many aspects of music. Other kinds of information can be gathered, such as text from the lyrics (we will present some results about this later), reviews, blogs or symbolic musical data like the score or a MIDI file. However we restrict our starting point to the raw audio data.

Classification

Statistical classification algorithms use the features extracted from examples and try to derive a mathematical or predictive relationship between each of them and its label (an emotion in our case). In a supervised learning approach, the descriptors from each example of the database are used to train a classifier that learns a statistical mapping and models the problem. For instance it may automatically learn from many examples that happy music is more likely to be in a major mode and sad in a minor mode.

To achieve the classification task, we use well-known methods for statistical classification like k-Nearest-Neighbors (k-NN) or Support Vector Machines (SVM). Most of the standard algorithms are included in the WEKA software (Witten and Frank, 1999), no particular classifier is to be preferred by default. Several approaches should to be tested. However in machine learning in general and in music information retrieval in particular, SVM seem to be one of the best options. They are known to be efficient, to perform relatively well and to be reliable in many cases. In the emotion classification literature, the main differences are in the representation chosen, the methodology to get a ground truth and to evaluate the results. The classification stage is largely standardized using SVM and sometimes other classifiers, but with no dramatic improvement in the classification results.

Results

In this part we present evaluation results from different experiments and relevant empirical studies found in the literature. If predicting the emotion from audio is feasible, it is quite arduous to compare all the different approaches because they use different representations, databases and evaluation schemas. The Music Information Retrieval Evaluation eXchange (MIREX) attempts to make this comparison possible (Downie, 2006). The MIREX provides evaluation frameworks and metrics with which researchers could scientifically compare their approaches and algorithms. In 2007 a first evaluation in Audio Music Mood Classification was organized. The representation chosen for this contest was a categorical approach with mood clusters, where the clusters were mutually exclusive (one instance could only belong to one mood cluster). There were five categories, or mood clusters shown in table 2 and the best results achieved were around 60% of accuracy (Laurier and Herrera, 2007). It means that the best systems were able to classify correctly 60 % of the music given to test. This percentage is a mean made using a 3-fold cross-validation. Almost all the systems submitted to this evaluation were using SVM to classify and different sets of descriptors (Hu et al., 2008).

Clusters	Mood Adjectives	Accuracy in percentage
Cluster 1	passionate, rousing, confident, boisterous, rowdy	45.8 %
Cluster 2	rollicking, cheerful, fun, sweet, amiable/good natured	50.0 %
Cluster 3	literate, poignant, wistful, bittersweet, autumnal, brooding	82.5 %
Cluster 4	humorous, silly, campy, quirky, whimsical, witty, wry	53.3 %
Cluster 5	aggressive, fiery, tense/anxious, intense, volatile, visceral	70.8 %

Table 2. *Clusters of adjectives used for the MIREX 2007 mood evaluation task and mean accuracy of our classifier.*

In the literature other results are available and can be of interest, especially if the approach is different. Basically almost every scientific contribution differs in at least one key aspect. Several consider the category representation based on basic emotions (Laurier and Herrera, 2007; Sordo, Laurier and Celma, 2007; Shi et al., 2006; Lu, Liu and Zhang, 2006), while others treat the categories in a multi-labeling approach like Wieczorkowska et al. (2005). The basic emotion approach gives simple but relatively satisfying results with accuracies around 80-90% depending on the data and the number of categories. The lower accuracies for the MIREX approach mentioned before might be due to an overlap in the concepts included in the class labels (Hu et al., 2008). It could also be due to a stricter evaluation on more data than the other mentioned works. The latter (multi-labeling) suffer from a difficult evaluation in general, as the annotated data needed should be much larger. Indeed if we want to use precision and recall[6] in an appropriate way we need to annotate all the data we evaluate with all categories (presence or absence), otherwise we might consider wrong results that are actually correct. There are also similar approaches to ours, such as the work by Li and Ogigara (2003), where they extracted timbre, pitch and rhythm features and trained Support Vector Machines. They used 13 categories, 11 from Farnsworth (1954) and 2 additional ones. However the results were not satisfying (it was one of the very first studies of mood classification), with low precision (around 0.32) and recall (around 0.54). This might be due to the small dataset labeled by only one person, and to the large adjective set. Another similar work should be mentioned; Skowronek et al. (2007) used spectral, tempo rhythm, tonal and percussive detection features together with a quadratic discriminant analysis to model emotions. He achieved a mood predictor with 12 categories considered binary with an average accuracy around 85%.

Other studies concentrated on the dimensional representation. Lu, Liu and Zhang (2006) used Thayer's (1996) model based on the energy and stress dimensions and modeled the four parts of the space: contentment, depression exuberance and anxious. They modeled the different parts of the space using Gaussian Mixture Models. The system was trained with 800 excerpts of classical music and the system achieved around 85% accuracy (trained with three fourths and tested on the remaining fourth of the data). Although it was based on a dimensional system the prediction was made on the four quadrants as exclusive categories. However another relevant study (Yang et al., 2008a) used Thayer's arousal-valence emotion plane, but with a regression approach, to model each of the two dimensions. They used mainly spectral and tonal descriptors together with loudness features. With these tools, they modeled arousal and valence using annotated data and regression functions (Support Vector

[6] Precision and recall are two typical measures in Information Retrieval. Precision is a measure of exactness (ratio of correct instances in the retrieve set) and Recall a measure of completeness (amount of correct instances retrieved over the whole set of correct instances)

Regression). The overall results were very encouraging and demonstrated that a dimensional approach is also feasible (see figure 4 for an application of this research).

In another work worth to be mentioned here, Mandel, Poliner and Ellis (2006) designed a system using MFCCs and SVM. The interesting aspect of this work is the application of an active learning approach. The system learns according to the feedback given by the user. Moreover the algorithm chooses the examples to be labeled in a smart manner, hence reducing the amount of data needed to build a model achieving a similar accuracy with a standard method.

Our ground truth is based on songs already tagged by hundreds of people (Last.fm users). We added a manual validation step to ensure the quality and reliability of our data. The evaluation was conducted on the 1000 annotated examples mentioned previously. We extracted audio features and performed classification with a SVM. Four categories were considered: “happy”, “sad”, “relaxed” and “angry”, each one approached as a binary problem. Either an instance belongs to the category or not. It means that each category is a boolean problem with a random baseline of 50 % of accuracy (i.e., a classifier just based on random choice between both categories would give an average accuracy of 50%). In table 3, we report the results of our evaluation using SVM and 10-fold cross-validation. The evaluation data were obtained after 10 runs of the same experimental setup (i.e., a random seed changed the allocation of files to folds for each run).

Category	**Accuracy in percentage (standard deviation)**
Angry	98.1 % (3.8)
Happy	81.5 % (11.5)
Sad	87.7 % (11.0)
Relaxed	91.4 % (7.3)

Table 3. *Accuracy of our classifiers on the different categories. Each category implies a binary decision (for instance ”angry” vs. “not angry”). This was made using SVM and 10 runs of 10-fold cross-validation.*

The performances we obtained using audio-based classifiers are quite satisfying and even exceptional when looking at the “angry” category with 98 %. It is difficult to directly compare this with the results from the MIREX evaluation, because we use here different categories and each one is considered binary. All four categories reached accuracies above 80%, and two categories (“angry” and “relaxed”) above 90%. Even though these results can seem surprisingly high, this is coherent with similar studies (Skowronek et al., 2007). Moreover as we deal with binary comparisons on a balanced dataset, the random baseline is 50%. Also, the examples are selected and validated only when they clearly belong to the category or its complementary. This can bias the database towards very clear differences. We should also notice that these models might work only for popular music (there was no classical music in our database), so it can generalize only to a certain extent. We conducted an experiment using the lyric information and combined the two classifiers: one for audio and one for lyrics. For the lyrics we used a text information retrieval method to detect the words that discriminate best between categories (Laurier, Grivolla and Herrera, 2008). We obtained the results presented in table 4.

Category	Accuracy in percentage (standard deviation)	Difference adding lyrics information to audio
Angry	99.1 % (2.2)	+ 1 %
Happy	86.8 % (10.6)	+ 5.3 %
Sad	92.8 % (8.7)	+ 5.1 %
Relaxed	91.7 % (7.1)	+ 0.3 %

Table 4. *Accuracy of a multimodal system using audio and lyrics.*

Results presented in table 4 show that lyrics contribute positively in correctly classifying emotions, especially for the "happy" and "sad" categories. It may be because lyrics are more informative about the valence. But we should also notice that the highest improvement occurs when there was more room for improvement. In a nutshell, detecting emotion in music is feasible if we consider simple categories or dimensions. The available results are encouraging continuing along this line and perhaps addressing more complex representations and models of emotions.

Conclusions

Even if we can predict some aspects of the emotion in a musical piece, the level of analysis can be made more precise. In addition, there are some important aspects that should be taken into account like the effect of the singer's voice, which theoretically contains much emotional information that is not considered by the existing techniques. Moreover the degree of emotional extent is limited to simple categories or to a few dimensions. Finally, we do not examine the time development of the emotions but we average musical features over the entire piece, which is certainly a simplification of the rich emotional tapestry that certain musical pieces can weave. Even though our initial results have been encouraging, there is room for many improvements. As explained previously the current state-of-the-art in automatic detection is quite limited to a simplistic view. Some effort should be made towards designing systems with a better music understanding and to allow a process of user modeling. Currently we average the perceived emotion among people to have a general prediction and the predictive models are "universal" (i.e., the same for all the users), but we should also seek to yield predictions at the user level. This would make possible the development of personal music assistants.

Section 3. From Music Information Retrieval to personalized emotion-based music assistants

People voluntary use music as a mood regulator, trying to induce emotional reactions (Sloboda, 1999). For instance, after heartbreak or sad event, someone may prefer to listen to sad songs either to give some solemnity to this moment or to find solace and consolation (Sacks, 2007). On the contrary one may want to feel better by playing happy songs. Music can be employed to emphasize the current mood or to decrease the intensity of certain emotions (Levitin, 2007). Someone feeling nervous could relax by listening to calm music. There is also evidence that experiencing musical emotions leads to physiological and cognitive changes. People can intentionally play with this phenomenon to influence their own state but also to communicate to others persons. A typical example would be a teenager listening to loud heavy metal or hardcore techno music (or any aggressive alternative) to express his anger and rebellion against his parents. It might make less sense to listen to this music if his parents are not around to receive the message (North, Hargreaves and O'Neill, 2000).

Nowadays personal electronic devices are ubiquitous. Almost everyone has at least a cell-phone or a music player and now these kinds of mobile devices have huge capabilities. They can already play more music that one has time to listen to, they can store hours of video and thousand of pictures, they are capable of taking pictures and can be used as a notebook and agenda. They enable one to trace

listening habits, to geographically locate the place where users are, to detect subtle movements with accelerometers, and soon they could use all this information and additional physiological data to contextualize any listening experience.

The aim of an emotion-based music assistant would be to exploit these types of devices and the techniques mentioned in the previous section to automatically and intelligently recommend musical pieces. Based on one's feeling or a targeted mood, the machine could choose the appropriate music. MIR techniques help to extract information from musical content and in our case they can automatically detect emotions from audio content. The technical issues have been explained previously in section 2. Basically learning from examples, an automatic system is able to retrieve songs with the similar mood or emotion in a large collection. It means that this trained system is able to detect if a song is sad or happy and can even estimate its degree of happiness.

Possible applications of this technology are numerous[7]. For instance, a device can play music according to one's mood and make him feel better (or intentionally worse). By manually selecting the current mood or a targeted one, the machine can choose or even create music accordingly. Many different factors are important to detect the emotion induced by music, but in our case, although we can grasp social data from the user, we mostly concentrate on the audio level. Even though is does not cover all the processes it is already enough to use it in many applications. The system can provide the music corresponding to one's demand in terms of mood. Skowronek et al. (2007), or Laurier and Herrera (2008) demonstrated prototypes that can extract the emotion from the audio content to visualize the prediction of the automatic classifiers, and the "intensity" of the predicted category. For instance in *Mood Cloud*, one can see the estimated amount of happiness or sadness of a song evolving while it is being played (see figure 5).

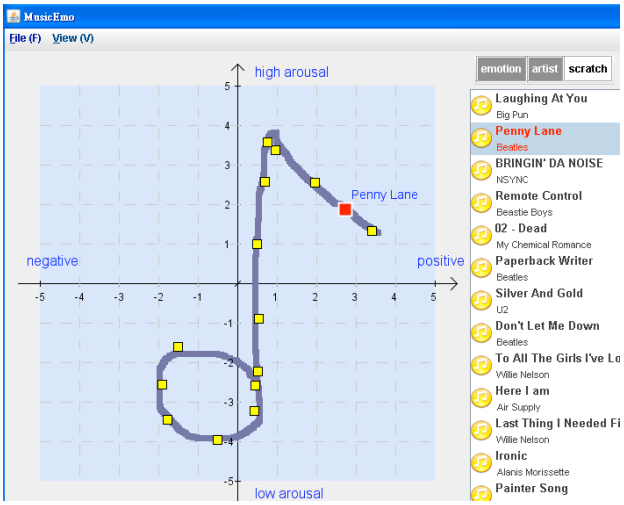

Figure 4. Mr. Emo (Yang et al., 2008b), reprinted with the permission of the authors

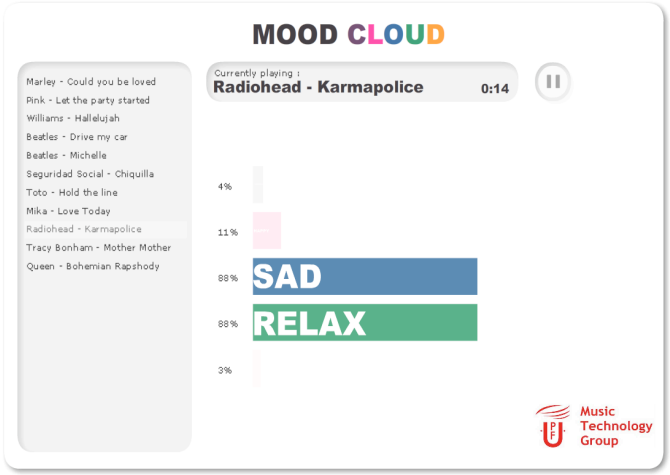

Figure 5. Mood Cloud interface (Laurier and Herrera, 2008)

This information can be directly used to provide to the listener songs containing the targeted emotion. Moreover using the probability of each emotion given by the automatic classifier, we can estimate how clearly an emotion is present in the music. This also allows defining mood trajectories by practically creating musical playlists. For instance one can select to start with relaxing music and progressively move to happy music (see figure 4). This can also be used to produce music or to help a composer in

[7] See, for example, http://www.bmat.com

his artistic work. An example would be a DJ picking out music to influence the emotional level of people in the audience. In an experiment, Casacuberta (2004) have created a virtual DJ called *DJ el Niño* that was playing music according to the emotion contained in the musical loops. This process could be automated and adapted depending on the audience reaction, which could be sensed using electronic devices.

If such a system can be employed at a personal level to browse and find music by emotion, this information can also be exploited in a professional environment for selecting soundtracks or for context-based adaptive music selection in video games. Plenty of applications are possible, as it would clearly ease the tedious process of manually choosing each musical piece. Advertisers can also take advantage of this technology. They use emotional content widely to associate feelings with a product or a message. Indeed, the effect on the customer has been demonstrated and people tend to have more positive evaluations of advertisements when exposed to happy music (North and Hargreaves, 1997). In general and for every user, the selection of music could be simplified using an intelligent assistant that knows the sensitivity of the user. It would make possible a more dynamic interaction where music could be chosen not by its name but with reference to another musical piece according to one's point of view. For instance if someone particularly likes one song or album for the emotion it contains, he can ask the system for something similar. This can be called the "more like this" feature and in that perspective it would be tuned and biased to fit to the needs of the user.

More interaction with human feelings can be achieved if the system can sense the emotional state by physiological features, like skin conductivity response, heart rate or blood pressure. Several studies have shown a correlation between these and other measurements and the emotion triggered (Grewe et al., 2007). Then the system would match the music and try to help regulate the emotional state. More complex techniques and sensors could help to detect stronger or subtle responses like chills, shivers down the spine, tears or goose bumps elicited by the music.

At this point, other applications become obvious. If such a system can provoke emotions and so change the physiological and psychological states, one step further is to use this for therapy. Scientific findings in music therapy have shown that it could help patients with psycho-physiological problems. Some analysis tools could also be developed liked in Luck et al. (2006). Information on how a subject can detect emotions compared to an automatic system can give the therapist additional information. One main application can be music recommendation for a medical purpose. This would consist of providing music according to the emotional sensibility of the subject and its needs. Depression treatment can be paired with listening to music that triggers positive emotions. As human memory is of an associative kind, the emotional links of music can also trigger or recover personal experiences and skills that have apparently been missed or forgotten as a consequence of strokes and other brain injuries (Sacks, 2007). Of course, any emotion-based assistant like the one we aim to achieve has to be studied deeply in a medical context to provide a substantial help to therapists. Using "positive" emotions in music can help to achieve tasks with a higher rate of success. Being exposed to "positive" music rather than "negative" music improves psycho-motor abilities like the writing speed (Pignatiello, Camp and Rasar, 1986) or the count time (Clark and Teasdale, 1985), but also the motivation to participate in social activities (Wood et al., 1990), or some information processing such as the time to produce associations to words (Kenealy, 1988). In any case, defining what is positive music is a debatable issue.

An emotion-based assistant can help in the context of therapy such as described by Bonny (2002). She developed a method called Guided Imagery and Music (GIM), where music takes an important part in the therapy process. She considers music as a "co-therapist" in this method where the music should be carefully selected to relax the patient and to be relevant in provoking certain emotions. The use of this

technique and the positive effect on physical health and mental state has been demonstrated in several studies (McKinney et al., 1997). Most probably these effects are partly due to the emotions that the music can induce. Automatic and intelligent systems can learn the patient profile to help in this context of music therapy. Once trained to one's concept, taste and expectations, an emotion-based musical assistant can be used as "Prozac", a mood regulator and more generally can create highly enjoyable musical experiences.

Future Trends

Emotionally sensitive devices are being studied and will soon appear on the commercial market. There already exist mobile devices including a bi-dimensional space to browse music (Sony Ericsson mobile phones with the SenseMe feature), and we can see prototypes automatically detecting mood in music. We believe we are at the beginning of a new trend in exploiting the current knowledge of musical emotions. In the future the personalization of these techniques will help to provide very accurate recommendations linked to one's emotional sensitivity. Taking into account the taste, the current emotional state and all the detailed emotional concepts and relevant musical features, the future of the current tools will be able to trigger specific emotions when asked for by the users. This can be conceived as an automatic system also but of course it should rely on some control by the users. The control might then not be in terms of artist or genre like today but in terms of a precise and personal emotional concept that would be matched by taking into account details about the user. We should also notice that we need to go beyond the current modeling of emotions as if it was just a labeling problem. We must better understand the overall process and provide smarter and more realistic models of music, users and emotions.

To enhance the human-computer interaction, a personal music assistant should also be considered as part of a global system capable of processing multimodal input. This means that the detection of the emotion in music would be one of the many inputs. The global system would take advantage and manage information from visual content (e.g., analysis of the face via a camera), from sound analysis (e.g., emotion in speech for example) and from text content. Merging all this information would lead to a better analysis of the emotion and enable machines to interact with humans in a realistic and impressive experience.

Conclusion

In this chapter we have shown that predicting a small set of emotions in music from the audio content is technically feasible. However the difficulty comes when introducing more subjectivity together with more complex semantic descriptions. For example, what are the differences between sad and melancholic? What is the overlap between both concepts? Would we all agree on this? We will probably not. Moreover how can our system be aware of the personal history of users, their social or cultural contexts, and their current status? We should also take care about the drawbacks of such systems. The marketing and social control issues can be frightening. However all the promising applications in everyday life, and especially in art and therapy are definitely strong arguments to continue these investigations. Detecting automatically emotion in music is at its early stage but we can expect many improvements and exciting applications in the future.

Acknowledgments

This work has been partially funded by the EU Project Pharos IST-2006-045035 (http://www.pharos-audiovisual-search.eu). The authors wish to thank their colleagues from the Music Technology Group for their help and especially Emilia Gómez and Owen Meyers for providing very useful feedback.

References

Besson, M., Faita, F., Peretz, I., Bonnel, A. M., & Requin, J. (1998). Singing in the brain: Independence of lyrics and tunes. *Psychological Science*, *9*(6), 494–498.

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, *19*(8), 1113–1139.

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11818–11823.

Bonny, H. L. (2002). *Music consciousness: The evolution of guided imagery and music* (L. Summer, Ed.). Barcelona Publishers.

Casacuberta, D. (2004). Dj el niño: expressing synthetic emotions with music. *AI & Society*, *18*(3), 257–263.

Clark, D. M., & Teasdale, J. D. (1985). Constraints on the effect of mood on memory. *Journal of Personality and Social Psychology*, *48*, 1595–1608.

Dalla Bella, Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, *80*(3), 1–10.

Damasio, A. (1994). *Descartes' error : Emotion, reason, and the human brain*. New York: Harper Perennial.

Davies, S. (2001). Philosophical perspectives on music's expressiveness. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research.* Oxford: Oxford University Press.

Downie, J. S. (2006). The music information retrieval evaluation exchange (MIREX). *D-Lib Magazine*, *12*(12).

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3), 169–200.

Farnsworth, P. R. (1954). A study of the hevner adjective list. *The Journal of Aesthetics and Art Criticism*, *13*(1), 97–103.

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, *113*(3), 464–486.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology*, *2*(2-3), 145–161.

Frijda, N. H. (1987). *The emotions (studies in emotion and social interaction)*. Cambridge University Press.

Gómez, E. (2006). *Tonal description of music audio signals*. Doctoral dissertation, Universitat Pompeu Fabra.

Gosselin, N., Peretz, I., Noulhiane, M., Hasboun, D., Beckett, C., Baulac, M., et al.. (2005). Impaired recognition of scary music following unilateral temporal lobe excision. *Brain*, *128*(3), 628–640.

Gouyon, F., Herrera, P., Gómez, E., Cano, P., Bonada, J., Loscos, A., et al.. (2008). Content processing of music audio signals. In P. Polotti & D. Rocchesso (Eds.), *Sound to sense, sense to sound: A state of the art in sound and music computing* (pp. 83–160). Berlin: Logos Verlag Berlin GmbH.

Grewe, O., Nagel, F., Kopiez, R., & Altenmüller, E. (2007). Emotions over time: synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion*, *7*(4), 774–788.

Guaus, E., & Herrera, P. (2006). Music genre categorization in humans and machines. In *Proceedings of the 121st convention of the audio engineering society.*

Herrera, P., Bello, J., Widmer, G., Sandler, M., Celma, O., Vignoli, F., et al.. (2005). SIMAC: Semantic interaction with music audio contents. In *Proceegins of the 2nd european workshop on the integration of knowledge, semantics and digital media technologies* (pp. 399–406). London, UK.

Hevner, K. (1936). Experimental studies of the elements of expression im music. *The American Journal of Psychology*, *48*(2), 246–268.

Honing, H., & Ladinig, O. (2008). The potential of the internet for music perception research: A comment on lab-based versus web-based studies. *Empirical Musicology Review*, *3*(1), 4–7.

Hu, X., Downie, S. J., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the 9th international conference on music information retrieval* (pp. 462–467). Philadelphia, PA, USA.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge: The MIT Press.

Juslin, P., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217–238.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(5).

Kenealy, P. (1988). Validation of a music mood induction procedure: Some preliminary findings. *Cognition & Emotion*, *2*(1), 41–48.

Kim, Y., Schmidt, E., & Emelle, L. (2008). Moodswings: A collaborative game for music mood label collection. In *Proceedings of the 9th international conference on music information retrieval* (pp. 231–236). Philadelphia, PA, USA.

Kivy, P. (1989). *Sound sentiment: An essay on the musical emotions*. Temple University Press.

Koelsch, S., Fritz, T., Cramon, D. Y. V., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: an fmri study. *Human Brain Mapping*, *27*(3), 239–250.

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian journal of experimental psychology*, *51*(4), 336–353.

Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *Proceedings of the international conference on machine learning and applications.* San Diego, CA, USA.

Laurier, C., & Herrera, P. (2007). Audio music mood classification using support vector machine. In *Proceedings of the 8th international conference on music information retrieval.* Vienna, Austria.

Laurier, C., & Herrera, P. (2008). Mood cloud : A real-time music mood visualization tool. In *Proceedings of the 2008 computers in music modeling and retrieval conference* (pp. 163–167). Copenhagen, Danemark.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford: Oxford University Press.

Levitin, D. (2007). Life soundtracks: The uses of music in everyday life. *Report prepared for the exclusive use of Philips Consumer Electronics B.V., Eindhoven, The Netherlands, http://www.yourbrainonmusic.com*.

Li, T., & Ogihara, M. (2003). Detecting emotion in music. In *Proceedings of the 4th international conference on music information retrieval* (pp. 239–240). Baltimore, MD, USA.

Logan, B. (2000). Mel frequency cepstral coefï¬cients for music modeling. In *Proceeding of the 1st international symposium on music information retrieval.* Plymouth, MA, USA.

Lu, D., Liu, L., & Zhang, H. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, *14*(1), 5–18.

Luck, G., Riikkilä, K., Lartillot, O., Erkkilä, J., & Toiviainen, P. (2006). Exploring relationships between level of mental retardation and features of music therapy improvisations: a computational approach. *Nordic Journal of Music Therapy*, *15*(1), 30–48.

Mandel, M., Poliner, G., & Ellis, D. (2006). Support vector machine active learning for music retrieval. *Multimedia Systems*, *12*(1), 3–13.

Mandel, M. I., & Ellis, D. P. (2007). A web-based game for collecting music metadata. In *Proceedings of the 8th international conference on music information retrieval* (pp. 365–366). Vienna, Austria.

McKinney, C. H., Antoni, M. H., Kumar, M., Tims, F. C., & Mccabe, P. M. (1997). Effects of guided imagery and music (gim) therapy on mood and cortisol in healthy adults. *Health Psychology*, *16*(4), 390–400.

Menon, V., & Levitin, D. J. (2005). The rewards of music listening: response and physiological connectivity of the mesolimbic system. *Neuroimage*, *28*(1), 175–184.

Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University Of Chicago Press.

North, A. C., & Hargreaves, D. J. (1997). Music and consumer behaviour. In D. J. Hargreaves & A. C. North (Eds.), *The social psychology of music* (pp. 268–289). Oxford: Oxford University Press.

North, A. C., Hargreaves, D. J., & O'Neill, S. A. (2000). The importance of music to adolescents. *British Journal of Educational Psychology*, 255–272.

Nussbaum, C. O. (2007). *The musical representation: Meaning, ontology, and emotion* (1 ed.). Cambridge: The MIT Press.

Panksepp, J., & Bernatzky, G. (2002). Emotional sounds and the brain: the neuro-affective foundations of musical appreciation. *Behavioural Processes*, 133–155.

Patel, A. D. (2007). *Music, language, and the brain*. Oxford: Oxford University Press.

Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, *68*(2), 111–141.

Pignatiello, M. F., Camp, C. J., & Rasar, L. (1986). Musical mood induction: An alternative to the velten technique. *Journal of Abnormal Pychology*, *95*(3), 295–297.

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

Sacks, O. (2007). *Musicophilia: Tales of music and the brain*. New York: Knopf Publishing Group.

Sacks, O., & Freeman, A. (1994). An anthropologist on mars. *Journal of Consciousness Studies*, *1*(2), 234–240.

Scherer, K. R. (1991). Emotion expression in speech and music. In J. Sundberg, L. Nord, & R. Carlson (Eds.), *Music, language, speech, and brain* (pp. 146–156). London: MacMillian.

Scherer, K. R., & Zentner, M. R. (2001). Emotional effects of music: Production rules. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 361–392). Oxford: Oxford University Press.

Schubert, E. (1999). *Measurement and time series analysis of emotion in music*. Doctoral dissertation, University of New South Wales.

Serrà, J., Gomez, E., Herrera, P., & Serra, X. (2008). Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(6), 1138–1151.

Shi, Y.-Y., Zhu, X., Kim, H.-G., & Eom, K.-W. (2006). A tempo feature via modulation spectrum analysis and its application to music emotion classification. In *Proceedings of the IEEE international conference on multimedia and expo* (pp. 1085–1088). Toronto, Canada.

Skowronek, J., McKinney, M., & Van de Par, S. (2007). A demonstrator for automatic music mood estimation. In *Proceedings of the 8th international conference on music information retrieval* (pp. 345–346). Vienna, Austria.

Sloboda, J. (1999). Everyday uses of music listening: A preliminary study. In S. W. Yi (Ed.), *Music, mind and science* (pp. 354–369). Seoul National University Press.

Sordo, M., Laurier, C., & Celma, O. (2007). Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of the 8th international conference on music information retrieval* (pp. 531–534). Vienna, Austria.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford: Oxford University Press.

Thayer, R. E. (1996). *The origin of everyday moods: Managing energy, tension, and stress*. Oxford: Oxford University Press.

Tomkins, S. S. (1980). Affect as amplification: some modifications in theory. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research and experience.* New York: Academic Press.
Trainor, L. J., Tsang, C. D., & Cheung, V. H. (2002). Preference for sensory consonance in 2- and 4-month-old infants. *Music Perception*, *20*(2), 187–194.
Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, *10*(5), 293–302.
Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, *22*(4), 720–752.
Wieczorkowska, A., Synak, P., Lewis, R., & Raś. (2005). Extracting emotions from music data. In *Foundations of intelligent systems* (pp. 456–465). Springer-Verlag.
Witten, I. H., & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* . Morgan Kaufmann.
Wood, J. V., Saltzberg, J. A., & Goldsamt, L. A. (1990). Does affect induce self-focused attention? *Journal of personality and social psychology*, *58*(5), 899–908.
Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008a). A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 448–457.
Yang, Y. H., Lin, Y. C., Cheng, H. T., & Chen, H. H. (2008b). Mr.emo: Music retrieval in the emotion plane. In *Proceedings of the ACM international conference on multimedia.* Vancouver, BC, Canada.

Key Terms

Music information retrieval (**MIR)** is an interdisciplinary science aimed to studying the processes, systems and knowledge representations required for retrieving information from music. This music can be in symbolic format (e.g., a MIDI file), in audio format (e.g. an mp3 file), or in vector format (e.g., a scanned score). MIR research takes advantage of technologies and knowledge derived from signal processing, machine learning, music cognition, database management, human-computer interaction, music archiving or sociology of music.

Music categorization models consider that perceptual, cognitive or emotional states associated with music listening can be defined by assigning them to one of many predefined categories. Categories are a basic survival tool, in order to reduce the complexity of the environment as they assign different physical states to the same class, and make possible the comparison between different states. It is by means of categories that musical ideas and objects are recognized, differentiated and understood. When applied to music and emotion, they imply that different emotional classes are identified and used to group pieces of music or excerpts according to them. Music categories are usually defined by means of present or absent musical features.

Music dimensional models consider that perceptual, cognitive or emotional states associated with music listening can be defined by a position in a continuous multidimensional space where each dimension stands for a fundamental property common to all the observed states. Pitch, for example, is considered to be defined by a height (how high or low in pitch it is a tone) and a chroma (the note class it belongs to, i.e., C, D, E, etc.) dimension. Two of the most accepted dimensions for describing emotions were proposed by Russel (Russel 1980): valence (positive versus negative affect) and arousal (low versus high level of activation). This variety of dimensions could be seen as the different expressions of a very small set of basic concepts.

Musical features are the concepts, based on musical theory, music perception or signal processing, that are used to analyze, describe or transform a piece of music. Because of that, they constitute the building blocks of any Music Information Retrieval system. They can be global for a given piece of music (e.g., key or tonality), or can be time-varying (e.g., energy). Musical features have numerical or textual values associated. Their similarities and differences make possible to build predictive models of more complex or composite features, in a hierarchical way.

Supervised Learning is a machine learning technique to automatically learn by example. A supervised learning algorithm generates a function predicting ouputs based on input observations. The function is generated from the training data. The training data is made of input observations and wanted outputs. Based on these examples the algorithm aims to generalize properly from the input/ouput observations to unobserved cases. We call it regression when the ouput is a continuous value and classification when the ouput is a label. Supervised learning is opposed to unsupervised learning, where the outputs are unknown. In that case, the algorithm aims to find structures in the data. There are many supervised learning algorithms such as Support Vector Machines, Nearest Neighbors, Decision trees, Naïve Bayes or Artificial Neural Network.

Support Vector Machine (SVM), is a supervised learning classification algorithm widely used in machine learning. It is known to be efficient, robust and to give relatively good performances. In the context of a two-class problem in n dimensions, the idea is to find the "best" hyperplane separating the points of the two classes. This hyperplane can be of n-1 dimensions and found in the feature space, in that case it is a linear classifier. Otherwise, it can be found in a transformed space of higher dimensionality using kernel methods. In that case we talk about a non-linear classifier. The position of new observations compared to the hyperplane tells us in which class is the new input.

Personal music assistants are technical devices, that help its user to find relevant music, provide the right music at the right time and learn his profile and musical taste. Nowadays mp3 players are the music personal assistants, with eventually access to a ecommendation engine. Adding new technologies like the ability to detect emotions, sense the mood and movements of the user will makes these devices "intelligents" and able to find music that trigger particular emotions.