Neural and Music Correlates of Music-evoked Emotions

Konstantinos Patlatzoglou

MASTER THESIS UPF / 2016

Master in Sound and Music Computing

Master thesis supervisor:

Dr. Rafael Ramirez

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

Acknowledgements

I would like to thank Dr. Rafael Ramirez for the opportunity to work on the topic of music and emotions within the cognitive neurosciences, as well as for his help and guidance throughout the implementation of this work. I would also express my gratitude to my friend Theodoros Theodoridis, for his support and encouragement during my studies. Finally, I would like to thank Dr. Xavier Serra for the opportunity to study in the SMC master.

Abstract

One of the basic research interests in cognitive neuroscience of music, comes from the affective phenomena that take place in music. The question of how the human brain represents and organizes conceptual knowledge has been investigated by scientists in different fields and still remains an open problem. Several neuroimaging studies on music-evoked emotions, have shown distinct spatial patterns of activity that emerge from brain structures, already known to be involved in emotions. From the musicological point of view, there has been a strong tendency in the aesthetics of music to emphasize on the importance of the musical structure. Leaving aside factors such as the musical context and listener properties, two questions are addressed in this work: 1) Can we train and test a computational model tha predicts fMRI activity related to music-evoked emotions, based on acoustic features extracted from the music? 2) Which are the features most relevant to the task regarding the basic emotions of joy and fear? Using fMRI data obtained from 17 individuals during a music listening session of 24 tracks (which belong to 3 classes of joy, fear and neutral stimuli), along with the extraction of audio descriptors from music using MIR (music information retreival) tools, a machine learning approach is selected for the creation of the model. By training multiple linear regressions, a predictive relationship is achieved between the extracted musical features and the BOLD activation of fMRI images, that correspond to each stimulus-track. The cross validated accuracies of alternative models seem to depend on the various feature and voxel selection strategies. The results show the possibility of such approach, with high accuracies for specific selection strategies. Nevertheless, what should be predicted and precisely how remains a challenge in the field.

Contents

List of figures

List of tables

1. INTRODUCTION

1.1 Background

The a priori perspective we have for the notion of time, seems to constitute the basis for music and our musical understanding. As an art form, music is described through several components, with rhythm and harmony having the main role. It is also a common quantification example (and maybe the first) of a qualitative phenomenon through mathematics. Its relation to mathematics is known already from ancient times, since the concepts of number and rhythm originate together, as well as the concepts of ratios and harmony [1]. Western history is associated in the last three centuries with an aim to explain qualitative phenomena using quantitative methods, in order to test experimentally and interpret these phenomena through an objective manner. Although the creation of music, its practice and its importance may differ throughout the various civilizations in history, its origins seem to indicate several common characteristics that lead to a consensus of music as a universal language.

While difficult to understand due to the high complexity of its nature, music is a subject of study in a variety of disciplines such as musicology, history, sociology, psychology, biology, physics, informatics, etc. In the last decades and along with the technological advancements, it is possible to study music from a cognitive neuroscience perspective, an empirical scientific approach based on brain mechanisms that take place in cognitive processes related to music (listening, performing, composing, reading), with professionals and tools coming mainly from neuroscience, psychology and computer science. This study is mostly based on the interpretation of data taken from several participants and using brain imaging techniques, such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG) and others.

Some of the fundamental elements that describe sound and the role they play in music perception and cognition, include pitch, loudness, duration, timbre and spatial qualities of sound. From the combination of such attributes, higher level concepts derive, such as melodies, harmony, rhythm, dynamics and structure. It is the totality of these concepts and the relation between them that give rise to our understanding of music [2]. The connection of such musical components with the neural activity and the perceptual representation of the brain, is one of the concerns of this specific field of cognitive neuroscience.

Music is one of the functions that activate most of the regions of a human brain, and thus it would be impossible for a specific center to be related to musical cognition (e.g. a music piece containing lyrics would activate a region responsible for speech). Brain processing beyond the auditory cortex is distributed, while several internal representations are constructed that vary from properties of the acoustic stimuli. By observations and experimental/correlation studies, researchers try to develop hypotheses regarding music cognition and behavior, along with the creation of models that try to predict or reveal associations and causalities between musical and cognitive variables.

1.1.1 Music and Emotion

One of the basic research interests in cognitive neuroscience of music and music psychology in general, comes from the affective phenomena that take place in music. In

contrast to more static terms such as mood and temperament, emotions can be considered as relatively brief episodes of complex interactions among neural and hormonal systems, which give rise to subjective experiences (described by phenomenal consiousness and a positive or negative effect) and can generate cognitive processes, along with the objective-observed physiological adjustments and behaviors of the individual. Although several theories exist regarding the nature and components that constitute emotions, psychophysiological changes and behavior are the components that enable for the direct observation and measurement of such phenomenon.

The relationship between music and emotional states has been studied in depth from the various disciplines, while the implications of such findings concern many areas, from philosophy and music theory to composition and performance. This investigation usually refers to the identification of features derived from the content of a musical piece (or a simpler stimulus) which relate to a specific emotional reaction, the underlying mechanism and nature of the reaction, as well as other relevant factors that may influence this reaction, such as the music context (the artist, the performance, visual associations, etc.) and properties of the human subject (current moods and psychological aspects, social context, preferences and personality traits, musical training, etc.). Independent of culture, there are plenty of indications for perceiving emotions in a similar way among people that may or may not have musical training (or contact to western music), which could relate to the origins of music and biological properties of the human brain (e.g. the sensory dissonance as described by Helmholtz [3], the theory of local consonance by Sethares [4], etc.)

A distinction regarding the nature of emotions in music has already been made from a philosophical/psychological point of view, as described by the cognitivists' and emotivists' approaches. The first one refers to a music-conveyed emotion, in which an emotional state is transfered and recognized by the listener, by means of structural features, performance and listener features, or other contextual features. The ability to perceive emotions in music appears already from infancy [5, 6] and develops throughout childhood, as shown by several studies using facial expressions and labeling (for simple emotions the ability is shown at 4-5 years old). Since many features participate in the overall degree of the experience, and contextual or listener features can vary between situations and individuals, research has turned its focus on the role of the musical structure and its objects. These objects could resemble emotional expression, as the dynamic structure of the music is associated with configurations of human behaviors (e.g. postures, gestures, attitudes, etc.), whether these behaviors derive naturally or culturally [7] (language is a common example, from which musicologists and researchers borrow terms to describe phonology, syntax and semantics in music). Some of the structural features that are associated to emotions include tempo, mode, loudness, melody and rhythm (e.g. fast tempo and major mode have been associated with happiness, while slow tempo and minor mode have been associated with sadness). When these features conflict in time or get mixed, research has suggested that the listener can perceive multiple emotions which may or may not fall on a bipolar scale [8, 9].

The second approach refers to a music-evoked emotion, in which music affects the emotional state of the listener itself. For example, the process theory [10] suggests that emotions could be elicited to the listener through the automatic and immediate response of motor or other autonomic activities which prepares us for action, as a result of musical processes. Although the induced emotion is harder to measure due to the subjectivity of the experience, listeners' reports and the observable responses of the physiological changes often favor this approach [11]. Several studies have suggested

that the same structural features that convey an emotion to the listener, may evoke a corresponding or independent emotional reaction too, by absorbing or associating the perceived expression [12] (a corresponding evoked reaction from structural features such as tempo and mode has been found as an example). Familiarity with a piece of music has also been found to play an important role for enhancing the emotional reaction [13]. As in the case of cognitivists' approach, music context and associative memories in general can also be responsible for emotional sources that affect the listener. Juslin and Västfjäll have proposed a model of eight different psychological mechanisms based on these structural, contextual or listener features, in which music can evoke an emotion to the listener [14, 15].

1. ***Brain Stem Reflex***: acoustical characteristics that influence the brainstem can signal a potentially important and urgent event, which leads to an emotional reaction (e.g. a sudden, loud or dissonance sound can induce arousal or unpleasantness to the listener)

2. ***Rhythmic Entrainment***: a process in which synchronization occurs between an external rhythm in the music and an internal bodily rhythm, such as the heart rate. This proprioceptive feedback can affect emotional components due to bodily changes (e.g. increasing the arousal)

3. ***Evaluative Conditioning***: an emotion is induced from a musical stimulus which has been associated with another positive or negative stimulus (or event), through systematic repetition.

4. ***Emotional Contagion***: the perceived emotional expression in the music induces an emotional reaction to the listener, by activating internal representations (neural substrates, muscle feedback) of the emotion (e.g. prosodic information which resemble emotional speech).

5. ***Visual Imagery***: an emotion is induced in the listener due to visual associations with the music (e.g. a natural landscape, a person, etc.)

6. ***Episodic memory***: a musical stimulus can evoke an episodic memory, which in turn can be associated with a specific emotion

7. ***Musical expectancy***: Implicit or explicit musical learning creates patterns, schemas, organization and rules, which produce expectations in the listener. By violating, delaying or confirming these expectations in time, emotional reactions emerge as a result.

8. ***Aesthetic Judgment***: the aesthetic value of a musical piece can vary among individuals, who hold different preferences in messages or ideas that are conveyed through music.

When it comes to the relation between the conveyed and the evoked emotion, research suggests that although the nature of the reactions is not identical, they are highly correlated [16]. Whether one or the other has a stronger effect has been controversial due to different studies that indicate both asymmetries, while multiple

variables seem to affect the perceived or felt outcome (type of music, type of reports, etc.).

1.2 Motivations

Despite the style or type of the music, its structural elements and the different definitions among cultures and societies, the aesthetic examination has always been considered a main topic within music. Music is an intrinsic aspect of people's way of life, having an important role in various human activities, from religious rituals and social ceremonies to individuals that create, perform and listen to it, with one of the main reasons lying in its emotional power. Questions regarding its beauty and why we enjoy it, in parralel with its capacity to influence human psychology and behavior, started in ancient times with the exploration of the mathematical and cosmological dimensions of rhythm and harmony, until recent times, where focus has shifted in the experience of music listening and how it relates to emotions. Although there are lot of contributions from philosophers, musicologists, musicians and other experts on the matter, empirical studies within psychology and neuroscience in the recent years could provide a scientific theory which can be tested and verified, with computational tools available that allow the quantification of the musical structure and the corresponding cognitive processes. As it has been already argued [17], the so called "semantic gap" in our current computational models (as described within the field of Sound and Music Computing) that relates to the lack of description of the higher abstract levels of music, such as the emotional description, can be attributed to the lack of consideration of cognitive models.

The understanding of neural and music correlates of music-evoked emotions has many implications for many aspects in music. Starting with philosophy, the definition of music itself and the importance of emotional expression as the essence of differentiation among organized sounds, with the perceptual domain having the main role (in contrast to the acoustic or the graphemic domain [17]). This also relates to the biological point of view, with open questions regarding the origins of music, the universal features and behaviors, aspects of cognitive processes and any functions/advantages that derive from its practice. More specifically, implications regard the way music theory and composition developed and continue to change, the way musicians use the different acoustic and musical elements to induce emotions, and the connections among acoustics, psychoacoustics and emotional changes. Music is also a useful tool for neuroscience in general, with brain as a dynamic system that changes, and music as a complex mean of communication that provides insights for the different cognitive functions that it involves.

Apart from basic research, investigating the neural and music correlates has some practical relevance for everyday life applications or practices of musicians, with areas that include the role of music in society, music performance, education and therapy. Music therapy notably, an attempt to use music for a variety of medical conditions (such as psychiatric or physical disorders, communication or interpersonal disorders, and others) or to improve health-related activities, has gained the attention of many researchers. Given a formal understanding of the underlying mechanisms of music-evoked emotions, it can be used for a positive impact in cognitive, social and emotional abilities, thus improving our quality of life.

1.3 Research Question

The question of how the human brain represents and organizes conceptual knowledge has been investigated by scientists in different fields and still remains an open problem. Many studies within neuroscience have shown with brain imaging techniques that distinct spatial or temporal patterns of activity emerge, for different objects of certain semantic categories (objects like words, pictures, or musical stimuli [18, 19]). There are several studies on music-evoked emotions, which try to understand the emotional effect of music along different factors (such as compared to other type of stimuli, its interaction with other conceptual objects, the recognition of emotional classes within lesion studies, and others), with specific regions of interest that are associated to the task. Neural correlates (mostly) from fMRI studies have shown distinct patterns of activity in several brain structures, some of which have traditionally known to be crucially involved in emotions [20]. From the musicological point of view, there has been a strong tendency in the aesthetics of music to emphasize on the importance of the musical structure. Leaving aside relevant factors such as the musical context or listener properties, and by concentrating on the indications for global features within music-evoked emotions, there is an investigation of the neural and music correlates of distinct basic emotions. Although there are mainly descriptive theories regarding this connection, the attempt is to predict specific brain activation based on structural features extracted from the music's audio signal. The question regarding the features used and the training of a testable computational model for predicting the brain is the main research goal.

2. STATE OF THE ART

2.1 Emotion Representations

Definitions of emotions vary within academic disciplines, with the scientific community differentiating terms and mechanisms that may or may not play a role in an accurate description. By exluding intertwined concepts such as mood, temperament, dispotition or motivation, theories of emotion concentrate on the involvement of specific components and the interaction that take place among them, namely the subjective experience, cognitive processes, expressive behavior, psychophysiological changes and instrumental behavior (motivation). Whether cognition is an important aspect of emotions (in the form of judgments, evaluations, or thoughts), and whether some of these components are causes of others or simple epiphenomena, remains debatable until today. The general consensus is that more than one of these components are needed for an accurate description, with psychology concentrating on the conscious experience characterized by the psychophysiological changes, the biological reactions and the mental states. The physiology of emotion is closely related to the arousal of the nervous system, with different strengths relating to particular emotional states. Moreover, a characterization of a positive or negative influence (pleasure or displeasure) is given as part of the mental state. These properties indicate relationships among different emotions and introduce topics of contrasting and categorization.

Some theorists have argued for the discretization and consistency in responses to internal or external events of the various emotional states, as biological functions with evolutionary significance and adaptive value for the organism [21]. Other approaches describe them as existing on a continuum of intensity, allowing for quantitative comparison and representation of more complex states [22]. Concentrating on emotional episodes which are brief in time, and not on the general dispositions of character traits, a classification of emotions is used and researched within the scientific community, on these two viewpoints of discretization (emotions as independend constructs) and dimensional characterization of groups (dimensional continuum).

The first viewpoint categorizes emotions as a set of “basic emotions”, states that are discrete, measurable and physiologically distinct from one another, while each one of them is associated with one or several adjectives (closely related). This approach is supported by findings of universal recognition for a certain set of emotions without any conditioning involved, as well as by the fact that distinct expressions match with specific physiology and experience, as reported in experiments. These basic emotions, which are linked to survival issues, could give rise to more complex emotions from the combination of them, along with any cultural associations for each case.

The second viewpoint places emotions in a multidimensional map, which allows visuallization and a measure of distance between the different states. The scales for each dimension seem to indicate aspects of each emotion, with two dimensions commonly having the main role, valence and arousal (also called as “core affect” [23]). Valence refers to the bipolar measure of the negative and positive feeling of the subjective experience, while arousal refers to the activation or deactivation (in terms of energy) of the experience. This idea led to a theory of emotion as a set of components (one of which is the core affect) that are understood as continuous processes, each of which has a dynamic part of appearance (evolutionary or cultury) and contributes to the instantiation of an emotional state as part of a larger group, rather than a distinct indepedent expression.

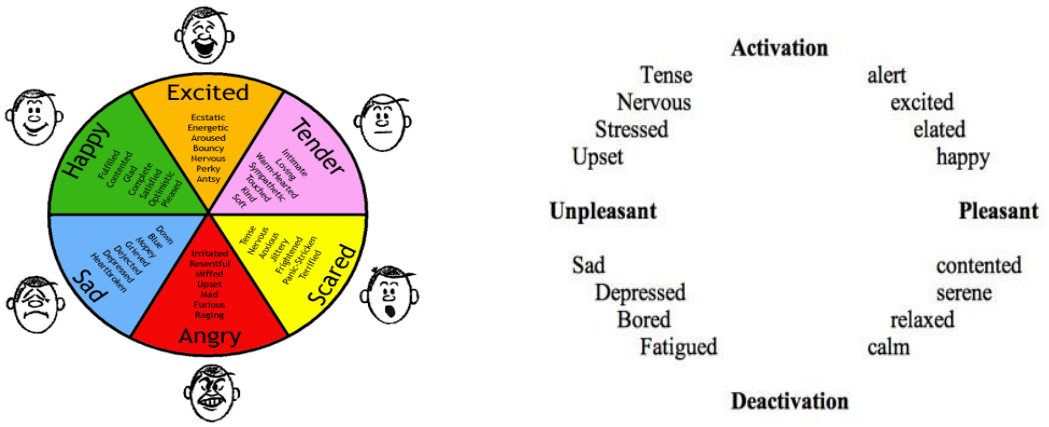


Figure 1. Examples of Basic Emotions (left) and a Multidimensional Map of Emotions (right)

Another categorization that can be made refers to primary and secondary emotions, the former requiring an external input (the sub-cognitive, fast circuit and correlated with limbic structures emotions) and the latter generated by internal thoughts (correlated with the cortical and cognitive slow circuit), although their relation and structural connectivity is not clear. Nevertheless, research in music psychology has indicated both of these categories to be present in music-evoked emotions, based on the various mechanisms from the structural, contextual and listener features (also seen in Juslin and Västfjäll's model). Categorical or dimensional representations of emotions are used throughout music research, with subsets of emotional states that relate to music, either as induced or conveyed.

2.2 Methodological Issues in Measuring Emotions

There are several techniques available in research that allow us to measure emotions, especially in music-listening contexts. Some of these techniques are:

- Word-lists, ratings and self-reports
- Expressive behavior
- Physiological responses
- fMRI and ERP

Wordlists, ratings (e.g. Likert scales) and verbal self-reports are some of the most widely used methods for studying emotions that may be conveyed or evoked to the listener. Expressive behavior is also another way to observe manifestations of emotional states, either through (subliminal) facial expressions or by social contexts (concerts, dances, etc.). Physiological responses refer to bodily changes, such as the heart rate, muscle tension, skin temperature or changes within the nervous system, which emotions create in each case. In general, one or more of the above methods are used for both music conveyed and evoked emotions, in order to obtain evidence.

Regarding the objectivity of the studies, conveyed emotions are considered easier to measure than the evoked emotions, which can be questioned for their

subjective personal experience. Self reports have shown high consistency in the results of identifying or reporting emotions like happiness, sadness, tenderness, threat or anger, among trained and untrained listeners [24, 25]. This method of course is vulnerable to biases during experimentation and thus often accompanied by expressive behavior and physiological responses. These techniques often require the use of instruments (for example facial expressions can be measured with electromyography-EMG), while more recent neuroimaging techniques such as the EEG and fMRI can show neural activity and patterns already associated with emotional states.

2.3 Emotion Classification from Audio Descriptors

2.3.1 Music Information Retrieval

Understanding and modeling sound and music has been a research subfield of Sound and Music Computing for a couple of decades now, with a methodology that focuses on computational approaches and multidisciplinary knowledge coming from signal processing, information retrieval, machine learning, psychology and musicology. Music information retrieval (MIR) specifically, along with the increasing computing power of the recent years, has been able to analyze sound and music in order to automatically extract descriptors (or features) that summarize its content (in recent years focus has turned into modalities of music context and user's properties aswell). This abstract representation of the content can then be used for comparative analysis and several other applications, including the automatic categorization of sound and music (e.g. genre or mood classification).

Data sources can be either in a symbolic representation, such as the score of a musical piece or a midi track, or in a digital audio format such as the wav, mp3 or ogg. Although symbolic representations may provide a more clear or "mathematical" description which can benefit further analysis (however losing information relating to performance, timbre and other aspects), the use of the audio signal has been predominant within research as a mean which is easier to access and data that contain all the available acoustic information. The way features are extracted from the audio signal can vary in techniques, which may include knowledge from psychology (sound perception and music cognition models) or musicology (e.g. musicological concepts of harmony and rhythm), in order to obtain meaningful sound or music descriptors. The amount of data reduction and the selection of the appropriate features that are needed for each case is not a trivial problem, with machine learning and statistics having the main role in such tasks.

The features that describe the music can be categorized into multiple levels of abstraction (usually referred as "low", "middle" and "high" level descriptors) that constitute the whole conceptual framework. Low level features refer to descriptors that are closer to the acoustical properties of the audio signal, such as the frequency, intensity, spectrum, or the onset and duration of a note. Middle level features refer to descriptors that relate to sensorial and perceptual information of sounds, such as pitch, loudness, timbre, intervals, beats, envelope, and others. High level features usually refer to more complex concepts such as melodic information, harmonic or rhythmic descriptions, instrumentation and dynamics, or even higher level concepts such as the emotional expression of a musical piece. Depending on the abstraction level, these

features can be extracted directly from the audio with signal processing techniques for the lower levels, or indirectly, requiring the use of statistics and machine learning for the higher levels. The information derived or the semantics of each descriptor may have lower value (or sense) for the user in the lower abstraction levels, while the techniques used and the insertion of appropriate perceptual/cognitive or musicological models contributes significantly in this effect. Another aspect of the extraction of audio descriptors is related to the temporal scope of the analysis, with the ability to segment the data in time windows and compute the information locally, resulting in instantaneous (time windows of few ms to few seconds), local or global descriptors of a track. Independent of the time scale, any audio descriptor can fall into one of the five musical facets, namely dynamics, rhythm, timbre, tonality and structure [26].

2.3.2 Musical Features and Emotions

There are several studies investigating the relation of musical features to particular emotional states. Even with the recognition that the emotional determination is mostly based on listening tests and emotion tagging, which implies a subjectivity to the results, and the fact that in most studies the research is centered around western musical culture, the problem is targeted in sorting out the universal from culture-specific relations. A main mapping found in literature [27] is presented in Figure 2, which shows a list of musical features mostly associated with five basic emotions.

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

Figure 2. Frequent musical features mapped with five basic emotions

Although a single feature is not capable of asserting the emotion, since many features are associated with multiple emotional states, a set of them could be sufficient enough for the task [28]. This mapping is also supported by studies in linguistics, where a similar correlation appears for attributes found both in speech and in music [29]. Features that are denoted with a single asterisk can be extracted from polyphonic audio content with current technologies, while two asterisks require the use of monophonic signals. For example, tempo can be estimated by locating beats in a track, or key and mode can be estimated by frequency distributions, but features like vibrato or articulation would require the seperation of instruments within the mix. The reliability and robustness of these features can vary depending on the algorithm and the content (e.g. musical style of a track), but the general information extracted is still relevant.

2.3.3 Music Classification with Machine Learning

The automatic categorization or "classification" of musical tracks is a part of MIR, which follows a general schema of four main steps, namely the dataset collection and ground truth, audio feature extraction, classification and evaluation [28].

1. Dataset Collection

A set of audio tracks (training dataset) is chosen for the classification system to learn from. The number of classes (emotions), the number of instances (examples) and the length of the audio tracks are aspects of consideration. The reliability of the ground truth (assigning an emotional class to each class) is also a crucial factor (e.g. tracks denoted by users or experts).

2. Audio Feature Extraction

Audio files are encoded as the digital information representing the waveform. Although lossy formats (mp3, ogg, etc.) may work well for the human ear, missing data and encoding artifacts could affect this analysis (usually pcm 16 bits encoding is preferred, with a sampling rate of 44.1khz). The objective in this step is the extraction of features that represent the most important components of the music. The use of time windows (frames) that segment the audio is an important part of the process, with different parameters (frame rate, window function, hop size) that depend on the algorithm and the descriptor. A large amount of features can be extracted, by summarizing the information in time, using statistics such as the mean, variance or derivatives. Filtering out irrelevant attributes with low influence on the target or high correlation to others, is also an important part prior to classification.

3. Classification

In machine learning and statistics, classification refers to the problem of identifying the class (category) of a given set for a new instance (music track), on the basis of a statistical learning derived from a training set whose classes are already known (also called as "supervised learning"). Each instance is represented by a set of quantifiable properties known as explanatory variables or features (either numeric or nominal). This feature vector, which in case of the audio tracks consists of audio descriptors, is used by the algorithms for predicting the class. The classification system

tries to discover relationships between the features and the classes in order to perform a mapping, in a way that maximizes its predictive accuracy. Some of the most often used algorithms in emotion classification are the k-Nearest Neighbor (k-NN), Decision Trees, Support Vector Machines (SVMs), Logistic Regression, Gaussian Mixture Models, and others.

4. Evalutation

The evaluation of the classifier is done by comparing the predicted classes of the test instances (instances that weren't included during the training) to the ground truth. The performance depends greatly on the features chosen during the training process, while different classifiers tend to perform better than others for different tasks, as indicated by various empirical tests. The usual evaluation measures are precision, recall, f-measure and accuracy, showing information about type I and type II errors. Cross validation is also a commonly used technique that favors the evalutation, in which the initial training dataset is split into K equally distributed sub-samples and the learning procedure repeats K times, each time with K-1 subsamples as training data and the remaining one as test, providing a mean accuracy over all the splits.

2.3.4 State of the Art

By accepting the premise that musical emotions tend to be highly consistent among listeners and thus quite objective, as shown in several studies [27, 30, 31], an attempt for content-based prediction and mathematical modeling through machine learning can be found in literature. Although the approaches may differ in the various steps but concentrating on the audio (instead of lyrics), the classification of emotions is framed into a supervised learning problem with the common schema described above.

Emotional classes are represented either as discrete categories [28, 32, 33, 34, 35] (usually a set of basic emotions) or within a dimensional map (as a 2D or 3D vector) [28, 32], that lead to classification and regression approaches respectively. Even though the categorical approach is the most often chosen, the dimensional models seem to be able to predict and explain the variance of the data, with valence and arousal often criticized for the lack of differentation of emotions that are close neighbours (such as anger and fear) and the ability to account for all the emotional variance found in music [32]. Nevertheless, these two paradigms seem to be highly compatible, with quantitative mappings available.

The ground truth is often created for a large number of various-styled tracks, with human annotators that assign an emotion to each track, consisting of several seconds (usually between 10 to 30 sec). This assignment of an emotional tag can be found in websites and social networks that contain music [28], gathered by questionnaires or games [33, 35] with different degrees of agreement from listeners (especially for emotional categories and labels that are close semantically), or with experts who annotate the music [34].

When it comes to the extraction of the audio descirptors, there are plenty of tools available such as PsySound [36], Marsyas [37], MIRToolbox [38], Essentia [39], and others, which share most of the features and musical concepts, while the implementations of the algorithms are quite similar. Many of the features are considered part of the standard audio descriptors [28] and require a monophonic mixture of the signal (by merging the stereo channels). Regarding the temporal scope, global features

are used to describe each music track, computed either along the whole audio signal or using statistical measures. This simplification does not take into consideration the time development of any emotions, which can provide relevant information and would require different techniques. In the majority of the studies, no more than 10 descriptors were needed to describe the variability of the emotions, with increased number providing statistically non-significant differences.

The algorithms found in different works also vary, with Support Vector Machines as an often used one which provides results with high accuracies. For the case of the categorical classification the results are in general satisfying, with accuracies around 60-90% (depending highly on the number or overlap of categories and the dataset used). Individual accuracies also seem to vary, with specific emotions detected more easily in general, due to high consistency or range of features (such as anger). Satisfying results have been also achieved with dimensional regression, although various implementations can be found [28, 32]. Of course, the comparison between the different implementations may not provide any meaningful information, since the representations of both audio and classification systems differ, as well as the initial data and evaluation methods [28].

Some of the most important audio descriptors for emotion classification found in literature [28, 32, 33, 34, 35] are mentioned below and concern both conveyed and evoked emotions (as highly correlated). These descriptors were derived based on either simple statistical or correlation tests with the emotional classes, or as part of a supervised machine learning algorithm that can reveal linear and non-linear relationships. In some cases, data reduction techniques were used as part of the feature selection process. The totality of the descriptors found can be categorized on the following five musical facets:

- **Dynamics**:
  - Loudness, Loudness variability

- **Timbre**:
  - Attack time
  - Spectral Centroid, Spectral Spread, Spectral Flux, Spectral Complexity, Spectral Entropy
  - Brightness
  - Zero Crossing Rate
  - Dissonance / Roughness
  - Mel Frequency Cepstral Coefficients (MFCCs)
  - Harmonic Richness

- **Rhythm**:
  - Onset Rate
  - Fluctuation
  - Tempo, Tempo variability
  - Rhythm irregularities

- **Tonality**
  - Chromagram
  - Mode
  - Key Clarity / Strength

- Chord strength
- Chords Chage Rate
- Harmonic Change Detection Function (HCDF)

❖ **Structure**:
- Complexity-Repetition - Novelty

Underlines indicate features that appear the most often within the studies. Audio descriptors that require monophonic signals are not present as relevant, since the datasets constitute of polyphonic music. The relation of the above features to the different emotional states (either as basic categories or as dimensional continuum) as found by the machine learning and classification systems, seem to be consistent with previous musicological reports in the literature (e.g. high loudness variability in fear and low for happy/anger, fast tone attacks in happy/anger and slow in sad/tenderness, high dissonance in anger/fear and low in happy/sad/tender, major mode in happy/tenderness and minor in sad/fear, small tempo variability in happy/fear and large in anger). The variability of many properties seem to play an important role on the task, which also relate to the fact that it provides temporal information, which otherwise would be dismissed.

2.4 Functional Magnetic Resonance Imaging

For over a century now, there are techniques developing which allow us to directly or indirectly image the structure and the function of the brain. In the last decades, some of these techniques have become popular in the scientific research within psychology and neuroscience, with functional imaging providing a way to obtain, analyze and visualize neural information in time for different regions of the brain. One of such techniques, which is widely used in cognitive neuroscience of music, is the functional magnetic resonance imaging (fMRI).

2.4.1 Overview

FMRI is a relatively recent functional neuroimaging procedure that is based on the MRI technology and the hemodynamic response of the brain, in order to measure neural activity in time [40]. MRI scans are able to locate certain atomic nuclei in the brain by using strong magnetic fields and radiofrequency pulses, which in turn provides a structural image of the different substances in the brain. Functional changes are captured by the differences in magnetic properties between the levels of oxygenation in blood, as measured by the blood-oxygen-level dependent (BOLD) contrast, which indicate the neural activity [41]. The relation between cerebral blood flow and neuronal activation is known due to the metabolic requirements of the neurons, which use the glucose and oxygen carried in the blood flow in order to function. By comparing the BOLD signal against local or field potentials, researchers have been able to associate the hemodynamic response mostly to the post-neuron-synaptic activity and internal neuron processing (the input and the integrative process of the neuron), rather than the firing of the neurons. As a qualitative signal, it represents a sum of the overall activity

of an area, with higher values indicating higher neural activity. The onset of neural activity leads to a local increase of the blood flow, with a delay of about 2 seconds and reaches a peak within 4-6 seconds, after which it returns back to a normal rate (modeled as the hemodynamic response function). For continuously firing neurons, the BOLD peak can spread and last up until activation stops, where it returns to the baseline. BOLD sensitivity is affected by different brain regions, which may have different inflow and consumption of glucose, although the responses can be compared across participants (subjects) for same region and task [42]. It can also be affected by other factors, including diseases, anxiety, medications, etc. In general, this blood-flow response is fluctuating and can last over 10 seconds, determining ultimately the temporal sensitivity of the measured brain activity.

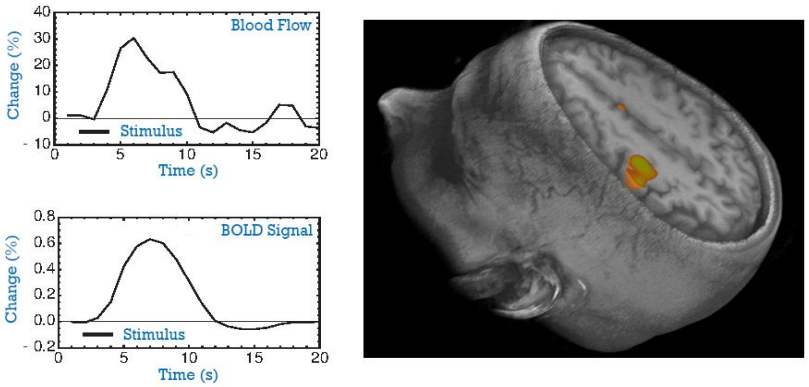


Figure 3. BOLD response to neural activation

The spatial resolution of the BOLD measurement is defined by the number of voxels, which refer to three-dimensional rectangular cuboids determined by the slice thickness and the imposed grid during the scanning. For full-brain studies, voxel size can range from 2-5 mm, an area that could contain few million neurons and billions of synapses. Temporal resolution is constrained in reliably separating the neural activity from the BOLD signal due to the behavior of the hemodynamic response, with sampling time (TR) ranging from 1-4 seconds (lower sampling time would correspond to a curve which can already be achieved by interpolation). This sampling time determines how often a particular brain slice is measured (and thus the head). Although the result can be improved with multiple presentations of a stimulus and combining different sampling times [43], the temporal resolution needed depends on the processing time of the various events that may take place (e.g. in visual stimuli the signal may need tens of milliseconds to reach the visual cortex and up to half a second for the corresponding neuronal activity and awareness of the event). A single voxel's BOLD response produces a signal in time (timecourse) which, apart from noisy contributions from the scanner, irrelevant brain activity and other factors, shows the temporal neural activity.

One assumption that is taken into account in many fMRI studies is that the hemodynamic response (BOLD signal) behaves linearly for multiple activations (e.g. in two simultaneous tasks). This assumption is supported by experiments in which either by increasing the stimulus presence or by having multiple short presentations in contrast to a similar longer one, the BOLD signal also increases or represents the addition of the responses of the multiple stimuli, respectively. For short time intervals below 2 seconds, nonlinear behaviors have been observed, that relate to the refractory period (brain

suppresses further activation of a similar subsequent stimulus) and can vary for different brain regions.

For most fMRI studies the experiment time may last several minutes, with the focus on cognitive processes that take place within few seconds. When it comes to the experimental design, there are different approaches for obtaining data along with some advantages and disadvantages. One common approach is the block design, where two or more conditions are alternated in blocks of time (several fMRI scans) and only one condition present within a single block. As the BOLD activity increases to a stimulus in an additive way, multiple presentations of stimuli contribute to the amplitude of the hemodynamic response. Block design provides intermediate time blocks (resting conditions) between the stimuli or tasks of focus, thus allowing the hemodynamic response to return to a baseline (although brain is never at rest). This introduces variability in the signal which allows differentiating the conditions of interest. On the other hand, noise in the measurements or poor choice of baseline, can affect the statistical power of such approach (e.g. a baseline condition with close to maximum activation may not allow for the representation of certain tasks that tend to increase it [44]). Another approach would be an event-related design, in which there is no sequence of fixed-duration conditions, but rather randomly presented stimuli. In both approaches, by differentiating the conditions in only the cognitive process of interest, any change in the BOLD signal is expected to represent the corresponding process (subtraction paradigm) [45].

A main goal of the fMRI is to localize any neural correlates of sensory, motor and cognitive processes. There are different hypotheses made for identifying brain regions that exhibit increased or decreased responses in conditions which may vary by chance. During the experiment there are usual sources of noise in the data (e.g. head movement), some of which are treated during the preprocessing of the exported signals. Overall, the effective use of such study relies upon knowledge found in different domains, from physics and neuroanatomy, to psychology and behavioral data, to mathematics and statistics.

2.4.2 Sources of Noise and Preprocessing

Analyzing the data obtained by the fMRI can reveal potential correlations between brain activation and cognitive processes or a task, which the subject has undertaken during the experiment. Cognitive states, such as an induced emotion, have been shown to be predicted solely from the fMRI of a subject with high degree of accuracy. In order to analyze the data and perform any statistical search though, various sources of noise must be controlled, as the BOLD signal change of interest is relatively weak. This is done with several preprocessing steps performed on the acquired images, as described below.

Sources of noise refer to signal changes due to elements which are not of concern during the study, namely the thermal noise, the system noise, physiological noise, random neural activity and differences in mental strategies or behaviors across subjects (and within a subject). Thermal noise affects all voxels in a similar way, as higher temperatures distort the current in the fMRI detector due to electrons' activity. System noise, which relates to the hardware itself, can emerge from the drift of the scanner (magnetic field drifting), changes in the receiver coil from brain's current distribution, or the non-uniformity of the magnetic field. One cause for the physiological noise is the head (and brain) movement in the scanner due to breathing,

heart beats, fidgeting, tensing, or other physical responses from the subject. FMRI records slices of activity in time, with head movement resulting in an unwanted change of the mapping between the voxel's absolute location and a specific neuronal area. Another cause of physiological noise, which contributes the most in the total noise [46], comes from changes in the blood flow rate, blood volume and the use of oxygen of the brain over time. Independent and random neural activity is also something present and unavoidable during the experiment, by internal or external stimuli such as thoughts, scanner noise, etc. Moreover, mental strategies and reactions to a stimulus can change over time or task, both within a subject and across subjects, which results in variations in neural activity. Although there's no way to mathematically model the irrelevant activity, training subjects how to respond prior to experiment is an often used method of control [47].

The acquired images from the scanner are in the form of 3D volumes (subject's head), while a single image is obtained every TR. These 3D volumes, which consist of arrays of voxels' BOLD intensity values, when concatenating in time produce a 4D volume of the timecourses of voxels. These data can be preprocessed with some of the following techniques:

- ***Slice Timing Correction***: this is by convention the first step in the preprocessing and refers to the correction of voxels' intensity values to a common timepoint. Since the scanner acquires slices for a single volume in different times, each slice refer to brain activity in a different timepoint within the sampling time. The computation is done by interpolating the discrete values of the timecourse of a voxel (assuming a smooth transition) and representing a single image with a common reference for all slices.

- ***Head Motion Correction***: this is a common correction associated with the head movement. As the BOLD signal of an area may be represented by different voxels along time, each timecourse can include activations from adjacent voxels. By applying a rigid-body transform to each volume, which shifts and rotates the data in various ways, we try to find an optimal transformation that would produce the smoothest timecourse for all voxels. A cost function is used to compare each transformed volume with a reference, although no optimal solution can be found due to the number of the possible candidates.

- ***Distortion Corrections***: there are several techniques to account for the scanner's field non-uniformities, such as the use of shimming coils or the creation of field maps. Mathematical models of estimating the field's noise (e.g.. Markov random fields) can also be used for distortion corrections.

- ***Coregistration Algorithm***: apart from the functional images, a high resolution structural image with MRI is usually acquired, as a mean to segregate or detect brain regions of interest. This is done by aligning the fMRI volumes to the structural one, similarly to motion correction but having different modalities (resolution and intensity values)

- ***Temporal filtering***: this step refers to the removal of specific frequencies from the BOLD signal of the voxels. High-pass, band-pass, or low-pass filters can be used, depending on the spectral range of interest.

- ***Smoothing***: spatial filtering can be applied in order to average nearby voxel intensities, thus creating a smooth spatial map across the brain regions. This is often done by convolution with a Gaussian filter which, depending on the match of the width of the filter and the true spatial extent of the activation, can improve the signal-to-noise ratio.

- ***Spatial Normalization***: in order to analyze and integrate the totality of the results from several participants, a transformation of the subjects' data can be done by aligning each brain to a common brain atlas**,** such as the Talairach or the Montreal Neurological Institute (MNI) one. This normalization has a similar procedure with the head motion correction, by which various transformations aim to reduce the distance of the data to a reference image.

2.4.3 Statistical Analysis – General Linear Model

The preprocessed data of the fMRI can be statistically analyzed under different premises and techniques, in order to assess the effect of the measured variability. One common approach is to consider each voxel independently within the framework of the general linear model (GLM). As mentioned before regarding the linearity of the hemodynamic response, the assumption here is that the instantaneous BOLD measurement corresponds to the scaled and summed activity of several events (or stimuli), that are present at some point in the timecourse. Design matrices can be created with each row representing a time point, each column representing an event, and values (either discrete or continuous) that denote the presence of each event. By marking the events that are active along the timecourse, and by using a model of the hemodynamic response (a function with specific shape but variable amplitude), a prediction of the voxel's BOLD signal can be generated using the procedure of convolution. The mathematical description of the general linear model can be written as

$$\boldsymbol{Y = XB + U,}$$

where in this case **Y** is a matrix of the measurements (from the brain scanner), **X** is a design matrix (containing experimental design variables), B is a matrix of parameters to be estimated, and **U** can be a matrix that contains errors or noise. Noise is usually assumed uncorrelated across measurements, following a multivariate normal distribution. This model incorporates many statistical models (such as ANOVA, t-test, and others), with hypothesis tests being either multivariate or univariate in respect to **Y**. In the case of a single dependent variable and more than one independent ones, **Y**, **B** and **U** are column vectors, while the general linear model can be seen as an ordinary multiple linear regression. The regression model of a single voxel can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ ... + \beta_p X_{ip} + \ \varepsilon_i$$

or in matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 + X_{11} + & X_{12} + & \dots & X_{1p} \\ 1 + X_{21} + & X_{22} + & \dots & X_{2p} \\ \vdots & & \ddots & \vdots \\ 1 + X_{n1} + & X_{n2} + & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

and describes the measured (or predicted) BOLD intensity value Y_i (dependent variable) for the i^{th} measurement in the timecourse, in a linear relation to the p experimental design variables X_{ij} (independent variables, j = 1, 2, ..., p) that correspond to the events during that measurement. Since we can have more observations (n) and hence equations than the unknown parameters β_j, the estimation is done by minimizing the mean square error (this is an optimal solution when the error is distributed as a bell curve, assuming the accuracy of the linearity). These scaling weights can be thought as indications of the importance of each explanatory variable (by their absolute value) in changing the predictor. Linear regression can be used for the goal of prediction (in this case the hemodynamic response), by fitting a predictive model to an observed dataset of Y and X values. The estimation ability of the model is based on various assumptions regarding the independent and dependent variables (e.g. least squares method assumes weak exogeneity, linearity, constant variance and independence of errors, etc.), while other statistical properties can influence its performance.

Another approach in analyzing the data would be to consider the relationship among a group of voxels which contribute to the observed activity, rather than assessing independent voxels. In such techniques (e.g. multi-voxel pattern analysis – MVPA), a statistical analysis (similar to the one already described by machine learning techniques in music classification) is done in order to assess contributions of voxel populations for the different conditions or tasks, by training and testing a classifier.

2.5 Brain Correlates of Music-evoked Emotions

Music is a universal feature found in all human societies, and a prime motivation for our engagement comes from the emotional experience it evokes. The investigation of brain correlates of music-evoked emotions during the past decade has increased our knowledge in the understanding of human emotions in general, although the overlap between the two is still debatable. Nevertheless, music seems to evoke changes in major components such as the subjective feeling, the physiological arousal, the expressive behavior and the action tendencies (e.g. dancing or singing). From the neuroscientific perspective, emotions can be understood as a result of the integrated activity of affect systems (such as the brainstem, diencephalon, hippocampus and orbitofrontal cortex (OFC)) and emotional effector systems (peripheral physiological arousal systems and motor systems), with the corresponding information resulting into an emotional percept represented in areas such as the insular cortex, cingulate and secondary somatosensory cortex. These systems can be regulated and modulated by conscious appraisal, but the functional interconnections involved are not well understood yet.

When it comes to the phylogenetic origins of sound-evoked emotions, the vestibular system plays an important role in acoustic responses, with projections that initiate and support movement, as well as contributing to the arousing effects of music (e.g. motor neurons in response to low-frequency and loud or sudden sounds). Subcortical processing of sounds is responsible not only for the auditory sensations, but also for muscular and autonomic responses (this type of stimulation might contribute to

the impulse to move to a beat) [48, 49]. Apart from these primitive brainstem systems, several other forebrain systems also contribute to the music-evoked emotions. Some of the main pathways underlying autonomic and muscular responses to music include the anterior cingulate cortex (ACC), cochlear nuclei (CN), inferior colliculus (IC), primary motor cortex (M1), middle cingulate cortex (MCC), medial geniculate body (MGB), nucleus accumbens (NAc), premotor cortex (PMC), rostral cingulate zone (RCZ) and vestibular nuclei (VN). The auditory cortex (AC) projects to the orbitofrontal cortex (OFC) and cingulate cortex, while the amygdala (AMYG), OFC and cingulate cortex project to the hypothalamus and influence the endocrine system [20].

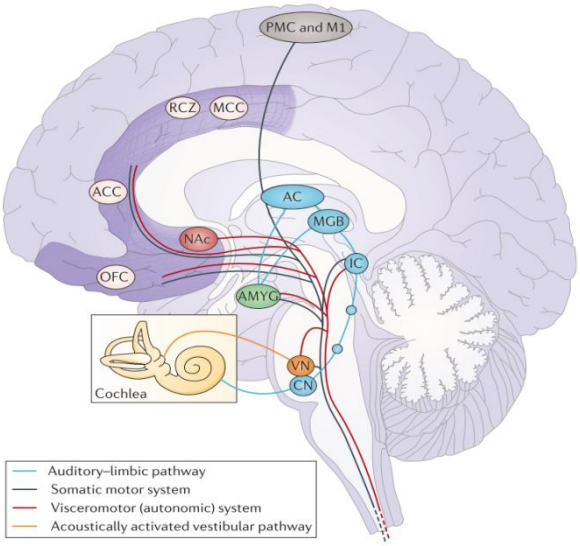


Figure 4. The main pathways underlying autonomic and muscular reponses to music

A meta-analysis of functional neuroimaging studies on music-evoked emotions showed activity changes in core emotion networks (mostly limbic and paralimbic). The various studies used different experimental approaches, while the investigation included music-evoked experience of intense pleasure, emotional responses to consonant or dissonant music, happy/sad/fear-evoking music, musical expectancy violations and music-evoked tension. Among the results, core brain regions that underlie emotions can be found, such as the superficial amygdala, which has a main role in stimuli with universal socio-affective significance (as music), the hippocampal formation, and the dopaminergic mesolimbic reward pathway, which is associated to pleasure. Some of the structures that showed functioncal significance are mentioned below [20]:

- ***Superficial amygdala (SF) - medial nucleus of the amygdala (MeA)***: socio-affective information and modulation of approach-withdrawal behavior

- ***Laterobasal amygdala (LB)*** : positive/negative reward value of music, regulation of neural input into the HF

- ***Central nuclei of the amygdala (CeA)*** : autonomic, endocrine and behavioural responses/expressions of emotion

- ***Hippocampal formation (HF)*** : regulation of hypothalamus-pituitary-adrenal axis activity, vulnerable to emotional stressors, attachment-related emotions

- ***Mediodorsal thalamus (MD)*** : modulates corticocortical communication, movement control, approach-withdrawal behavior

- ***Auditory cortex (AC)*** : central hub of affective-attentional network with limbic, paralimbic and neocortical connections

- ***Bordmann area 7***: conscious appraisal, subjective feeling, attentional functions

- ***Brodmann area 8***: response competition, role in musical tension

- ***Pre-supplementary motor area (SMA)*** : complex cognitive motor programming and preparation of voluntary action plans (e.g. dance)

- ***Rostral cingulate zone (RCZ)*** : interoceptive awareness, internal selection of movements, autonomic regulation

- ***Insula***: autonomic regulation, sensory interoceptive representation of bodily reactions

- ***Head of the caudate nucleus (hCN)*** : initiation/patterning of somatomotor behavior, anticipation of frissons

- ***Nucleus accumbens (NAc)*** : sensitive to rewards and motivates, initiates and invigorates behaviors to obtain rewards

- ***Orbitofrontal cortex (OFC)*** : control of emotional behavior and automatic appraisal, sensitive to expectations violation

The evocation of attachment-related emotions by music seems to be related to several social functions of music, which in numerous social contexts and for the most part of the human history, was an active engagement of a group. These social functions, which refer mostly to communication, cooperation and social cohesion, supported the survival of the individuals and the species, thus providing an evolutionary explanation regarding any adaptative value for music.

A comparison of the neural correlates with neuropsychological findings from patients with brain lesions or degenerative diseases that show impaired recognition of music-evoked emotions, revealed that the activation in several of the mentioned regions has a causal role in music-evoked emotions, rather than being simply correlational. Moreover, the dysfunction of some of these limbic and paralimbic structures in patients with neurological and psychiatric disorders, along with the power of music to evoke changes in their corresponding activity, has many implications for the development of music-based therapies [20].

2.6 Open Problems

The fact that there is no consensus within the scientific community on a definition of emotions, creates already a problem in studying the music-evoked emotions. The different components and mechanisms that may be under attention each time, along with the subjectivity of the human response under different psychological states and enviroments, can give rise to noisy data. By simplifying its complexity, we can create models of representation that enable us to use methods of matching conceptual objects (in this case music tracks) to specific emotional states, where we can test the reliability and consistency. Nevertheless, working with distinct states of basic emotions that are clearly separated (distant states within the dimensional map) and by measuring the physiological changes (which tend to be more objective), we can provide a validity which is supported in many of the available experimental literature. Formalizing emotion representation consistently with human perception (in this case music-specific), as well as standardizing methodologies and techniques in psychology, could be a start for a coherent comparison of the results among studies.

When it comes to modeling sound and music, there are several open problems that are being investigated in the field of Sound and Music Computing. A content-based mathematical approach for modeling the music-evoked emotions, automatically discards aspects outside the acoustic domain, such as lyrics, other contextual information connected to the music (associative memories) and listener properties, which play a significant role in music perception and cognition. Of course, the affective phenomena may not be significantly subject to change, since most people have similar exposition to musical training (implicit or explicit), and thus similar expectations and understanding. Whether a model could be regarded as generic or personalized for a user, depends highly on its construction.

Concentrating on the musical structure, audio and music descriptors need to be related semantically to our own understanding. At this moment, many aspects in music cannot be analyzed, due to the fact that the audio engineering perspective ignores the perceptual domain. Many existing biological inspired models give better results in sound analysis, while music itself is considered as a cognitive construct (cognition plays a role in the design of instruments, rhythmic/harmonic hierararchies, expectations, and others). For example, a (computational) auditory scene analysis would allow us to analyze individual objects (sources) which are perceived as independent streams of information (e.g. singer's voice). By implementing perceptual models, along with the existing musicological models (accepting that they affect our perception), we can analyze and understand the audio signal in a higher abstraction level.

When it comes to measuring the effect of the music evoked-emotions, fMRI has been shown to provide relevant information regarding the temporal and spatial patterns of brain activity (as also compared to other techniques). Some of the problems of such procedure can be assigned to the facts of the indirect imaging of neural activity, the changes in the hemodynamic response for different brain regions, or the noisy contributions in the signal and the arbitrary baselines for signal comparison (which can reduce, eliminate or even reverse the activity patterns). Moreover, the common assumptions of linearity in the statistical models and the seperation of the effects per voxel can dismiss relevant information, while the reverse and forward inference applied within the studies can lead to wrong conclusions (correlation vs causation). In general, although the spatial resolution is considered one of the best among neuroimaging procedures, the low temporal resolution can be restricting, especially when it comes to temporal phenomena such as music. The nature of the cognitive processes and the

gradual profiles of fMRI responses, can determine the efficiency over the different statistical tools of analysis (e.g. mean comparison over correlation methods). Overall, the validity of the study can be influenced by the sample size of participants, the hypotheses made and imposed on the models, and the treatment of the statistical tests.

As mentioned in the Introduction, the research question tackles the problem of moving from a qualitative description of the mechanisms of music-evoked emotions, to a computational model that can be tested. Applying the knowledge obtained from the current trends in modeling sound and music, the available tools for extracting audio descriptors, and the techniques for analyzing neural information acquired in fMRI studies, two specific questions are addressed:

1) **Can we train and test a computational model that predicts fMRI activity related to music-evoked emotions, based on acoustic features extracted from the music?**

2) **Which are the features most relevant to the task regarding the basic emotions of joy and fear?**

3. MATERIALS AND METHODS

In order to address the research questions described in the previous chapter, a methodology is selected based on the literature's state of the art, regarding the data sources, the current technologies, and the algorithms for the implementation of a computational model that makes directly testable predictions. Previous research has provided experimental evidence on the relationship between neural activity associated with semantic categories of objects and their features, by training competing models based on alternative assumptions regarding the relevant features and potentially, the encoding of the brain [19]. As already described in the general schema of classification within machine learning, the acquisition of a dataset with a ground truth and a feature extraction process is necessary for training and testing a prediction model which will lead to the evaluation of such task. The selected emotional classes, the music stimuli and the brain imaging data used in this experiment were part of the fMRI study in [18].

3.1 Participants

The fMRI data were obtained from 17 individuals (aged 20-30 years, M = 23.78, SD= 3.54, 9 females) that participated in the study. All participants were right-handed and had normal hearing, as assessed with standard pure tone audiometry. Seven of the participants had no formal musical training, eight participants had a short formal training (M = 2.81 years) in various instruments but had not played their instrument for several years, and three participants had a long formal training (M = 12.5 years) on an instrument that they were still playing. Exclusion criteria were left-handedness, professional musicianship, past diagnosis of a neurological or psychiatric disorder, a score of >12 on Beck's Depression Inventory, excessive consumption of alcohol or caffeine during the 24 h prior to testing, and poor sleep during the previous night [18].

3.2 Stimuli and Procedure

3.2.1 Music Tracks

The selection of music tracks was based on three discrete classes of stimuli, which intended to evoke [18]:

a) **feelings of *joy***
b) **feelings of *fear***
c) **neither joy nor fear (referred as *neutral*)**

Although the arousal levels of music-evoked joy and fear can be matched in some cases, both emotions are considered as physiologically distinct states (basic emotions) which correspond to a positive and a negative influence respectively, while they also seem to be universally recognized in western music.

Each class comprised of 8 tracks that were chosen to be pronounced representatives of the categories. Joy-evoking stimuli had been used in previous studies [50, 51, 52] and consisted of CD-recorded pieces from various epochs and styles (classical music, Irish jigs, jazz, reggae, South American and Balkan music). Fear-evoking stimuli were excerpts from soundtracks of suspense movies and video games. The high acoustic roughness of the fearful tracks was further increased, in order to increase the emotional effect of the stimulus. Joy and fear evoking tracks were chosen such that each joyful excerpt had a fearful counterpart that matched with regard to tempo, f0 mean, f0 variation, pitch centroid, spectral complexity and spectral flux. Neutral stimuli were created using the MIDI toolbox for Matlab [53], as sequences of isochronous tones with pitch classes randomly selected from a pentatonic scale. The tracks were generated so each one of them was matched to a pair of joy-fear stimuli, with regard to tempo, f0 range and instrumentation. The tones of the midi sequence were produced by high quality natural instrument libraries to resemble real musical compositions. All tracks were rendered as wav-files of same length (30 s), with 1.5s fade-in/fade-out ramps, and equal RMS power [18].

3.2.2 Experimental Design

A block design was used for obtaining the fMRI data, with the three conditions (classes of emotion) alternated in blocks of 48 seconds, which included a single track followed by a rating procedure. The arrangement of the blocks was in a pseudo-random order, so that no more than two musical stimuli of the same emotional class would follow each other. Each track was presented twice to the subjects, resulting in 48 (8 tracks per class) trials overall. The subjects were asked to listen to the music (30s) with closed eyes until the signaling of a beep tone (2s), after which they would commence a rating procedure (12s) followed by an interval of silence (4s). During the rating procedure, subjects evaluated their own subjective experience after listening to the musical excerpt, in terms of valence, arousal, joy and fear (6-point Likert scales). Participants were listening to the music via MRI compatible headphones [18].

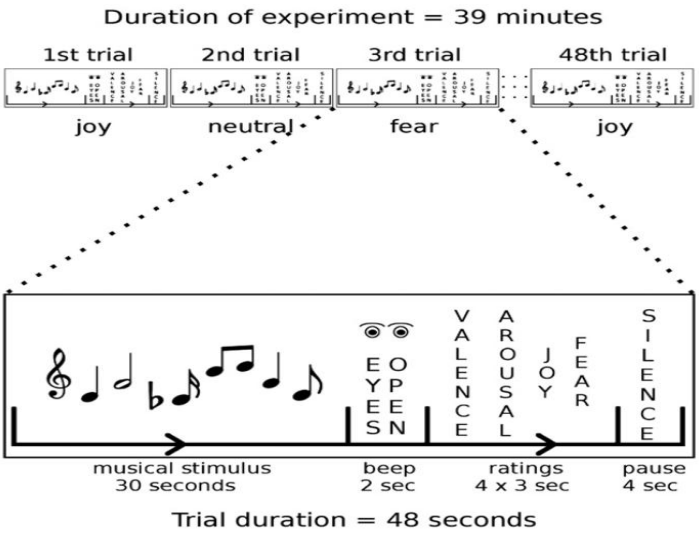


Figure 5. The experimental design of the fMRI study

3.3 fMRI Scanning and Preprocessing

The scanning was performed with a 3 T Siemens Magnetom TrioTim and continuous Echo Planar Imaging (EPI) with a TE of 30 ms and a TR of 2000 ms. Slice-acquisition was interleaved within the TR interval. The matrix acquired was 64 x 64 voxels with a field of view of 192 mm, resulting in an in-plane resolution of 3 mm. Slice thickness was 3 mm with an interslice gap of 0.6 mm (whole brain coverage). The acquisition window was tilted at an angle of 30^o relative to the AC-PC line in order to minimize susceptibility artifacts in the orbitofrontal cortex [18].

The fMRI data were preprocessed with LIPSIA 2.1 [54]. Data were corrected for slicetime acquisition and normalized into MNI-space-registered images with isotropic voxels of 3 cubic millimeters. Highpass filtering was applied for the removal of low frequency drifts (cutoff frequency - 1/90 Hz), as well as spatial smoothing with a 3D Gaussian kernel and a filter size of 6 mm FWHM [18].

3.4 Data Analysis

The results of the behavioral data during the rating procedure showed consistency in the responses among the participants [18]. Valence ratings, which indicate the positive or negative influence (pleasantness/unpleasantness), were higher for the joy-evoking stimuli in contrast to fear and neutral stimuli, which didn't show significant difference. Arousal ratings were lower for the neutral stimuli, with fear and joy having moderate values (no significant difference). Joy ratings were highest for joy stimuli, lowest for fear stimuli, with ratings for neutral being in between. Fear ratings were highest for fear stimuli, lowest for joy stimuli, with ratings for neutral being in between. These ratings of the subjective experiences confirm the ground truth of the dataset in terms of the intended evoked emotions of the stimuli.

The task of creating a computational model that predicts the fMRI activity can be described within the field of machine learning. The model is built by an algorithm that uses example inputs in order to make data-driven predictions in a mathematically optimized way. Given a training dataset to the system, which includes the example inputs and the desired outputs, a mapping is learned via a statistical process (supervised learning). When the output of the system is a continuous variable, the task falls into a regression analysis problem. In this case, the desired estimation refers to the relationships between musical (input) and cognitive variables (output).

As already mentioned, one common approach in analyzing fMRI data is through the general linear model (due to the assumption of the linearity of the hemodynamic response). A multiple linear regression can be trained for each voxel separately, where the independent (explanatory) variables correspond to musical features and the dependent (predicted) variable corresponds to the BOLD intensity of a voxel. For the mapping to occur, a feature vector of a track has to be associated with an fMRI image. An image can be considered as a 3D volume (or array) containing all the voxel values for a specific time point. We expect that a linear combination of relevant acoustic features extracted from the music's signal will be able to predict specific voxels' BOLD intensity, providing us with a spatial pattern of activity. The evaluation of the model is based on the predicted images derived from examples (testing stimuli) that haven't been used during the training.

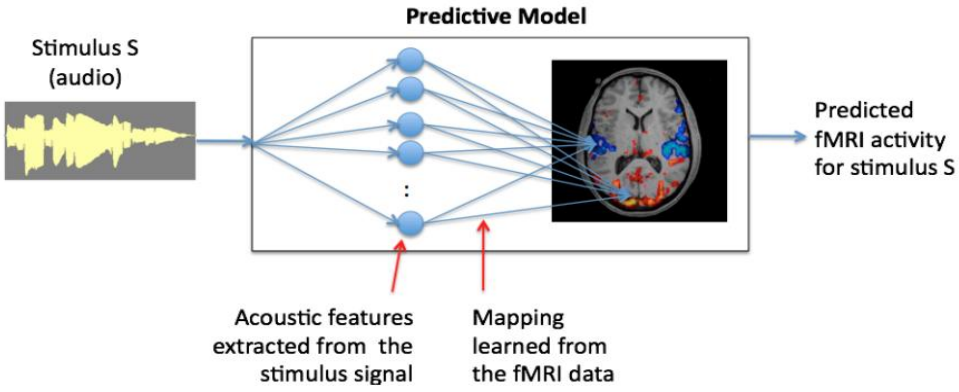


Figure 6. Schematic representation of the model for predicting fMRI activation

3.4.1 fMRI Images

Due to methodological constraints, the majority of fMRI studies in affective neuroscience focus on brief emotional episodes and initial reactions to external or internal stimuli. Emotions of joy and fear, as experienced in everyday life, can span over long time periods, in the range of minutes or even longer. Few available functional neuroimaging studies on the matter suggest that the neural activity underlying the emotion changes over time, as a result of the different connections and temporal responses of the systems involved (e.g. autonomic and endocrine processes) [55]. When it comes to joy and fear-evoking stimuli, research has indicated several correlated structures already mentioned in the previous chapter (e.g. auditory cortex, ventral striatum, hippocampal formation, insula, cingulate cortex, etc). By investigating the neural and music correlates of joy and fear [18], some spatial patterns of fMRI activity emerge, as the BOLD signal seems to increase for joy stimuli in bilateral auditory cortex and bilateral superficial amygdala, while decreases for the case of fear (increased bold activity can also be seen in the right somatosensory cortex during fear). The hemodynamic response which tends to be monotonous (e.g. in the AC the signal increases with time along each trial), doesn't seem to provide any temporal information with regard to a musical analysis. This behavior of the BOLD signal (timecourse), together with the low temporal resolution of the fMRI (TR - 2s), led to the use of a single representative fMRI image for each stimulus.

The creation of a representative fMRI image for each music stimulus is based on the averaging of the BOLD intensity values in time. This supports the notion of an activity pattern that emerges for objects belonging to a certain semantic category (as also seen in [19]), by taking into consideration the statistical variability of the BOLD signal during the different stimuli. This also allows the comparison of the emotional states among the various stimuli, voxel-wise and by mean values. The precise steps followed in computing the fMRI images of the tracks are mentioned below:

- The preprocessed fMRI data of the experiment are divided based on the intended evoked emotion of each stimulus, considering only the 15 scans (timepoints) of each block that correspond to the music listening (a shifting by 2 samples for the acquired scans is used, in order to account for the delay of the hemodynamic response). Each emotional class comprises of 240 scans (15*16 tracks) represented by 60x72x60 shaped 3D volumes.

- The timecourses of the raw fMRI data are normalized using the percent signal change, which transforms each voxel value with respect to the timecourse mean value, as defined by the formula:

$$p = \frac{y_t}{\bar{y}} \cdot 100$$

  This is the most widely used approach in analyzing fMRI data and relates to changes in the level of the measured signal for different voxels, or even across subjects. Voxel values refer to percent changes with respect to a mean derived from all volumes (all classes included) and can be compared to each other regardless the region of the head. In this step we consider which voxels are more (or less) active relatively to the different conditions (emotions).

- The next step is a dimensionality reduction, in which only the voxels that seem to be part of the head of the subject are included for further analysis (60x72x60 voxels). For this purpose, voxels with timecourse mean below 350 or variance 0 are of no interest.

- In the final step, a single fMRI mean image is created for each of the 24 stimuli. The image is created by computing the mean values of the voxels for all scans during the two presentations of each track, while the grand average of all 24 of these images is subtracted from each image. This way each fMRI image represents relative changes with respect to the other stimuli images.

3.4.2 Audio Descriptors

As previously described in detail, music information retrieval (MIR) enables us to summarize the content of a musical track with an abstract representation that is quantitative and comparable. The quest to identify the relevant and universal features of music in relation to the emotions of joy and fear, is constrained in the analysis of the acoustic signal of the musical stimuli. We know from previous studies that a set of audio descriptors can be used to assert an evoked emotion, but whether a descriptor has perceptual value and potentially a role in the encoding of the brain (thus allowing better predictability), is an independent and hard task. The possibility that some descriptors play a role in human auditory perception has been supported by experiments [56, 57], as found for features such as spectral complexity, sensory dissonance, spectral flux, spectral centroid, etc. The music tracks used in this experiment have already been chosen to match with regard to several low level descriptors (acoustical features), which can benefit from statistical biases (we expect that the most relevant features in emotional classification are high level features which incorporate perceptual and musicological concepts).

For the audio analysis and automatic extraction of the descriptors, MIRtoolbox 1.6.1 [38] was used in Matlab. The music tracks included in the experiment were in a lossless audio format (.wav, 16bit - 44.1kHz), while the temporal scope of analysis was chosen as global (descriptors referring to the whole excerpt). Overall, 143 features are extracted (all available functions in the library) which fall into the five facets of Dynamics (e.g. rms, etc), Rhythm (e.g. fluctuation, eventdensity, etc), Timbre (e.g.

attacktime, zerocross rate, brightness, etc), Pitch-Tonality (e.g. inharmonicity, keystrength, mode, etc) and Structure (e.g. novelty). Most of these features can reflect properties of polyphonic audio, such as the chosen music tracks. This total number of features includes also descriptors derived from the statistics of the original features (around 50), which include mean values, standard deviation, linear slope, periodicity frequency, periodicity amplitude and periodicity entropy (computed along frames in time). This way, even though a single feature vector is associated to a single track, temporal information can be still captured within the variables. All features are computed with the default parameters and options of the corresponding functions.

The number of instances used for the training of the computational model can determine the number of the explanatory variables. Since the number of musical stimuli are 24 (as well as the number of obtained fMRI images), a feature vector containing 6 audio descriptors is chosen, as a significantly smaller number is suggested to avoid model's overfitting. This is also consistent with studies that explained the variability of emotions in music using less than 10 descriptors. A feature selection process, which tries to measure and weigh the influence of the different variables concerning a specific task, can be used to determine a ranking of the features. Ideally, variables with low influence or high correlation to others, should be filtered out. For this task, the performance of the system depends highly on the predictability of the chosen features, with respect to the brain activation patterns emerged from the fMRI analysis.

The selection of the audio descriptors that constitute the feature vectors of the musical stimuli is based on 3 alternative strategies:

1. Literature Selection

Several descriptors that emerge during the review of the state of the art in emotion classification (and relate to polyphonic music), are considered as part of the selection process. The frequency of appearance within the studies, along with the musicological consistency, is an important indication of relevance to the task. Dividing the totality of the 143 descriptors into the 5 musical facets, a machine learning approach with classification (joy, fear, neutral) is used to obtain a ranking. The methods used, depending on the size of the subsets, included the CfsSubsetEval / ExhaustiveSearch and InfoGain / Ranker algorithms implemented in Weka [58]. Features with high ranking that also appear in the state of the art, are chosen over others. The 6 features with the highest ranking overall are:

1. *RMS energy std* (Dynamics)
2. *Metrical Centroid mean* (Rhythmic)
3. *Roughness mean* (Timbre)
4. *Key Clarity mean* (Tonality)
5. *Harmonic Change Detection Function std* (Tonality)
6. *Novelty Period Amp* (Structure)

2. fMRI Selection

This approach uses the fMRI images as computed already for the model, by taking into consideration only a subset of the available voxels whose activity relates to the evoked emotion. The methodology for voxel selection can vary, with three different approaches described in the next section. The strategy here is to attribute a ranking to each audio descriptor, by computing the Pearson's correlation between each voxel's value

variability with respect to the descriptor's value variability (24 points). The 6 descriptors with the highest correlation sum (over all voxels) are selected as the feature vector.

3. **Random Selection**

In this case, a random selection of 6 audio descriptors out of the 143 is chosen as the feature vector.

3.5 Prediction Model

The prediction model can be thought as the two-step process, illustrated in Figure 6. Given an arbitrary stimulus-track, the first step is to encode its abstract representation with the extraction of the audio descriptors. The selected descriptors are chosen with one of the three mentioned strategies and each descriptor is standardized by mean removal and variance scaling (unit variance). This can be helpful in regression equations where the variables correspond to arbitrary metrics, in order to facilitate the comparability of their relative importance. These variables constitute the intermediate semantic features, with the variety of their content determining their efficiency.

The second step is to predict the neural activation shown in the representative fMRI image of the corresponding stimulus, as a weighted sum of contributions from these intermediate semantic features. Each predicted value at voxel v in the brain is computed by linear regression (ordinary least squares) as:

$$Y_v = c_{v0} + c_{v1}f_1 + c_{v2}f_2 + \ldots + c_{v6}f_6$$

where f_i is the i^{th} feature of the stimulus-track and c_{vi} are the learned scalar parameters that specify the degree to which each feature contributes to the activation of voxel v. This model can be interpreted as predicting the full fMRI image across all voxels with a weighted sum of images, one per semantic feature f_i. These images are defined by the parameters c_{vi} for each i, producing an "fMRI signature" that indicates the influence of the feature in response to particular brain regions.

One theoretical assumption underlying this model is that the distinction of the emotional classes is reflected in the perceptual or "semantic" properties of the chosen features. This could provide any justification for a neural basis of the music-evoked emotions due to the distributional properties of the sounds that constitute the music. Another assumption regarding the efficiency is related to the linear sum of contributions from these structural features.

A separate computational model is trained for each of the 17 participants and their associated fMRI images, and for all the alternative feature selection strategies mentioned. For the comparison and visualization purposes of the task, only a subset of the available voxels is taken into consideration during the training and evaluation of the model. The experiments are repeated for three different voxel selection methods, which intend to find the brain regions related to the music-induced emotional states. The voxel selection methods, the training and testing of the model, and the evaluation of them are described in detail in the following section. For the implementation of the algorithms, Python 2.7.11 with several external libraries was used.

3.5.1 Voxel Selection

During the music listening part of the fMRI scanning, we expect that only a subset of the voxels in the brain is responsible for activity related to music-evoked emotions. To assess which voxels are appropriate for the analysis, 3 separate selection methods are implemented:

1. **ANOVA Selection**

The uncertainty of the emotional effect can be estimated from the variance of the fluctuations in the data. Analysis of variance (ANOVA) is the extension of the t-test used to assess the difference of mean values, for samples coming from 2 or more independent groups (conditions). In this case, the fMRI images are divided in the 3 emotional classes resulting in three data samples. One-way ANOVA is calculated for each voxel, obtaining a p-value that indicates the error probability of the mean differences being noise fluctuations. For small p-values ($p < 0.05$) we accept the alternative hypothesis that the means differ significantly, which suggests the voxel's correlation to the emotional states. The 2000 voxels with the lowest p-values are selected for the analysis.

2. **t-Test Selection**

A similar procedure with the ANOVA selection is done with multiple two-sampled t-tests, for each pair of the conditions (joy-fear, joy-neutral, fear-neutral). In this case, each voxel is assigned with three p-values that assess only the two conditions involved in the testing. The 2000 voxels with the lowest p-values are selected evenly from the 3 conditions.

3. **Stability Score Selection**

Another way to select a subset of the image voxels is by assigning a "stability score". Using the data from the two presentations, two images are created for each stimulus by computing the mean values of the voxels along the music listening scans. The stability score of a voxel is the Pearson's correlation between the corresponding values of the two images, with all the tracks concatenated. This assigns highest scores to voxels that exhibit consistency in their activity across the same stimuli (e.g. if a voxel exhibits the same 24 responses in both presentations, it would be assigned with correlation 1). The 2000 voxels with the highest stability score are selected for the analysis.

There are several mathematical assumptions underlying each one of these methods, which seem to be consistent in general with the fMRI analysis and empirical results. Although BOLD activation can be noisy, we expect that the supervised methods of the hypothesis testing will reveal similar voxels of interest. Unsupervised selection can differ in such task, since high stability score can be found for voxels with repeatable response pattern for some of the stimuli (which may not differ among conditions), that not necessarily distinguish the emotional states.

3.5.2 Training and Evaluating the Model

Alternative computational models are trained based on the different strategies for feature selection and voxel selection. Each model is trained and evaluated using a cross validation approach, in which the model is repeatedly trained with only 22 of the 24 available stimuli-tracks and their corresponding fMRI images ("leave-two-out" cross validation). The remaining 2 stimuli are used for testing, by first predicting their fMRI image responses and then matching these correctly to their corresponding held-out observed fMRI images. With the constraint that the two test stimuli have to belong to different emotional classes, the leave-two-out train-test procedure is iterated 192 times (all possible pairs between the 3 conditions). The evaluation of the model is based on the accuracy computed as the percentage of correct matches of the test inputs within the 192 trials.

Given a trained computational model, we obtain two new predicted images for the two left-out tracks (p1 and p2), which need to be matched with the observed fMRI images (i1 and i2). A successful matching is considered if the similarity between the images derived from the same stimulus-track (p1=i1 and p2=i2) is higher than the opposite pairing (p1=i2 and p2=i1). The similarity of two fMRI images is computed as the cosine similarity of the voxel arrays. The score of each match is obtained with two similarities, defined as:

$$match1(p_x = i_x and\ p_y = i_y) = cosineSimilarity(p_x, i_x) + cosineSimilarity(p_y, i_y)$$
$$match2(p_x = i_y and\ p_y = i_x) = cosineSimilarity(p_x, i_y) + cosineSimilarity(p_y, i_x)$$

3.5.3 Empirical Distribution

The expected chance accuracy of an uninformed model correctly matching the predicted fMRI images of the test stimuli to the observed ones is 50%. To determine the statistical significance and potential relevance of the chosen features, a distribution of accuracies from 100 models using the random feature selection is generated. These 100 models are trained and tested for each participant, using the same randomly chosen features. Since many of the audio descriptors capture relevant acoustic information regarding the emotional classification, we expect that several alternative feature vectors will result in an accuracy above chance levels.

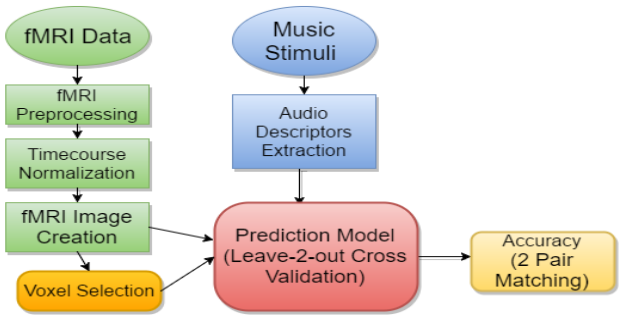


Figure 7. Model Training and Evaluation

4. RESULTS

4.1 Music Classification

Prior to the evaluation of the computational model, the predictability power of several sets of audio descriptors can be estimated, using a classification problem of the 3 emotional states. The dataset and ground truth, comprised of 8 tracks per emotional class (joy, fear, neutral), have been established in the methodology along with the feature extraction strategies. The algorithms used for the classification task are the Logistic Regression and Support Vector Machines. Logistic Regression can be seen as analogous to linear regression, when the dependent variable is categorical. Support Vector Machine (SVM) is an often used algorithm in emotion classification within MIR, with high accuracy results.

The five musical facets can be used for a categorized feature selection, in order to reveal the statistical inference of the different subsets of audio descriptors. Descriptors that belong to different facets can be considered statistically independent from one another. The obtained descriptors for the facets of 'Dynamics', 'Rhythm' and 'Structure' were based on the CfsSubsetEval/ExhaustiveSearch algorithms. For the categories of 'Timbre' and 'Tonality', InfoGain/Ranker was used, due to the larger descriptor subsets (larger search space). The 'InfoGain' feature set comprises of the top 6 features, as evaluated by their information gain ranking. 'Literature' and 'fMRI selection' sets consist of the features derived from the strategies described in the Methodology. The results of the 'fMRI selection subset' in the classification task, regard the 6 features that have been selected the most times, among the 17 participants (mainly from ANOVA and t-Test methods; stability score didn't show consistency among the subjects). The audio descriptors of each set are mentioned below:

- ***Dynamics***: RMS energy std (**1**)
- ***Rhythm***: BeatSpectrum std, Metrical Centroid (**2**)
- ***Timbre***: AttackTime mean, Brightness mean, Roughness mean (**3**)
- ***Tonality***: HCDF std, Key clarity, Tonal centroid (**3**)
- ***Structure***: Novelty PeriodAmp (**1**)
- ***InfoGain***: AttackTime mean, Chromagram-6, Chromagram-11, Key clarity mean, HCDF std, HCDF PeriodAmp (**6**)
- ***Literature***: Rms energy std, Metrical Centroid mean, Roughness mean, Key clarity mean, HCDF std, Novelty PeriodAmp (**6**)
- ***fMRI Selection***: Key Clarity, Chromagram-11, HCDF std, Chromagram-6, Chromagram-9, AttackLeap mean (**6**)

We can observe that most of the descriptors within the 'Literature' and 'fMRI Selection' sets, also appear in the analysis of the individual facets. The descriptors of

'Key Clarity' and 'Harmonic Change Detection Function' are shared between the two strategies of interest. None of the acoustic parameters used to match the musical stimuli are present, as confirmed by the analysis. The accuracy of the classifiers can be asserted to any semantic properties of the features, statistical noise or uncontrolled properties of the tracks. The accuracies of both classifiers and for all feature sets are shown in detail in Table 1. The implementation of the algorithms and the exported results were based on Weka (using the default parameters), using 10-fold cross-validation.

Feature Set	**Logistic Regression**				**Support Vector Machine**			
	Fear	Joy	Neutral	**Total**	Fear	Joy	Neutral	**Total**
Dynamics	75%	12.5%	62.5%	**50%**	75%	0%	50%	**41.6%**
Rhythm	75%	62.5%	87.5%	**75%**	37.5%	62.5%	87.5%	**62.5%**
Timbre	87.5%	37.5%	75%	**66.7%**	75%	37.5%	75%	**62.5%**
Tonality	100%	62.5%	75%	**79.2%**	100%	100%	62.5%	**87.5%**
Structure	75%	50%	75%	**66.7%**	75%	12.5%	75%	**54.2%**
InfoGain	75%	37.5%	75%	**62.5%**	100%	100%	62.5%	**87.5%**
Literature	100%	87.5%	100%	**95.8%**	100%	87.5%	100%	**95.8%**
fMRI Selection	87.5%	75%	62.5%	**75%**	100%	75%	75%	**83.3%**

Table 1. Classification accuracies (10-fold cross-validation) for the three emotional states, derived by different feature sets.

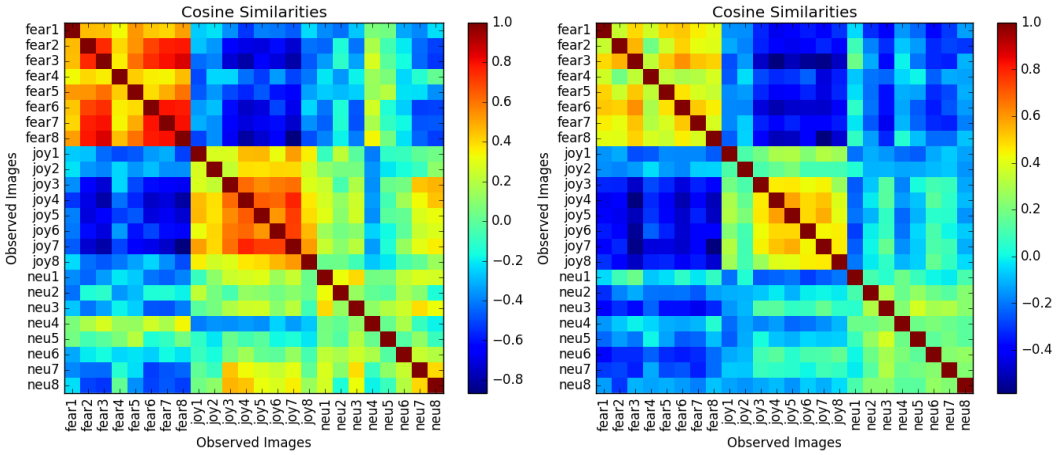
Given that the number of features used for each classification problem is different, the components of Rhythm and Tonality seem to have the highest relevance in distinguishing the emotional states. Fear-evoking tracks are the ones with the highest accuracies in general, with joy and neutral obtaining different prediction rates for different algorithms and feature sets. Overall, the two selection strategies chosen for the prediction model produce the best results, with features from literature giving the highest rate (95.8%). This allows for a consistent model based on the assumption underlying the distributional properties of sound and music, in relation to music-evoked emotions.

4.2 fMRI Classification

In a similar way, we can evaluate the obtained representative fMRI images in their ability to differentiate the 3 emotional states. This is an important step to assess the predictability of the computational model, in terms of the statistical mapping that occurs. A comparison between two fMRI images can be computed as the cosine similarity for the subset of the selected voxels. We expect that images coming from the same emotional class would have higher similarity rates. For the classification task, Logistic Regression and Support Vector Machines are used in a similar manner. The comparison of the observed images and the results of the classifiers are presented in the next sections, for each voxel selection method separately.

4.2.1 ANOVA Selection

Figure 8 shows the cosine similarities between the observed fMRI images, for all pairs of musical stimuli within the 3 emotional classes.



(a) **(b)**

Figure 8. Cosine similarities between the observed fMRI images (ANOVA)
(**a**) For 1 participant, (**b**) Averaged over all the participants

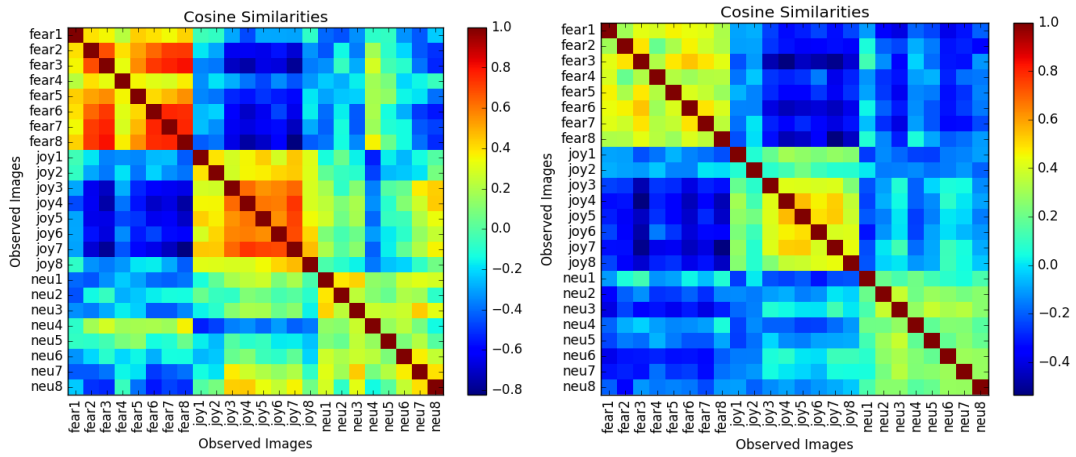
Table 2 shows the accuracies obtained by the two classifiers (10-fold cross-validation), for all participants in the experiment.

Subject No.	**Logistic Regression**				**Support Vector Machine**			
	Fear	Joy	Neutral	**Total**	Fear	Joy	Neutral	**Total**
1	100%	62.5%	100%	**87.5%**	87.5%	100%	100%	**95.8%**
2	87.5%	87.5%	100%	**91.7%**	100%	100%	100%	**100%**
3	87.5%	62.5%	100%	**83.3%**	87.5%	87.5%	100%	**91.7%**
4	100%	87.5%	100%	**95.8%**	100%	87.5%	100%	**95.8%**
5	100%	87.5%	75%	**87.5%**	87.5%	100%	87.5%	**91.7%**
6	100%	87.5%	100%	**95.8%**	100%	87.5%	100%	**95.8%**
7	87.5%	87.5%	75%	**83.3%**	100%	100%	100%	**100%**
8	75%	100%	100%	**91.7%**	100%	100%	100%	**100%**
9	100%	100%	87.5%	**95.8%**	100%	100%	87.5%	**95.8%**
10	87.5%	100%	87.5%	**91.7%**	100%	87.5%	100%	**95.8%**
11	100%	75%	87.5%	**87.5%**	100%	100%	100%	**100%**
12	75%	75%	100%	**83.3%**	100%	100%	87.5%	**95.8%**
13	87.5%	100%	100%	**95.8%**	100%	100%	100%	**100%**
14	100%	87.5%	87.5%	**91.7%**	100%	100%	100%	**100%**
15	87.5%	100%	100%	**95.8%**	100%	100%	100%	**100%**
16	100%	75%	87.5%	**87.5%**	100%	100%	100%	**100%**
17	100%	87.5%	87.5%	**91.7%**	87.5%	75%	100%	**87.5%**

Table 2. Classification accuracies of the observed fMRI images
(ANOVA selection)

4.2.2 t-Test Selection

Similarly, Figure 9 and Table 3 show the cosine similarities and the accuracies obtained by the classifiers, for the t-Test selection method.



(**a**) (**b**)

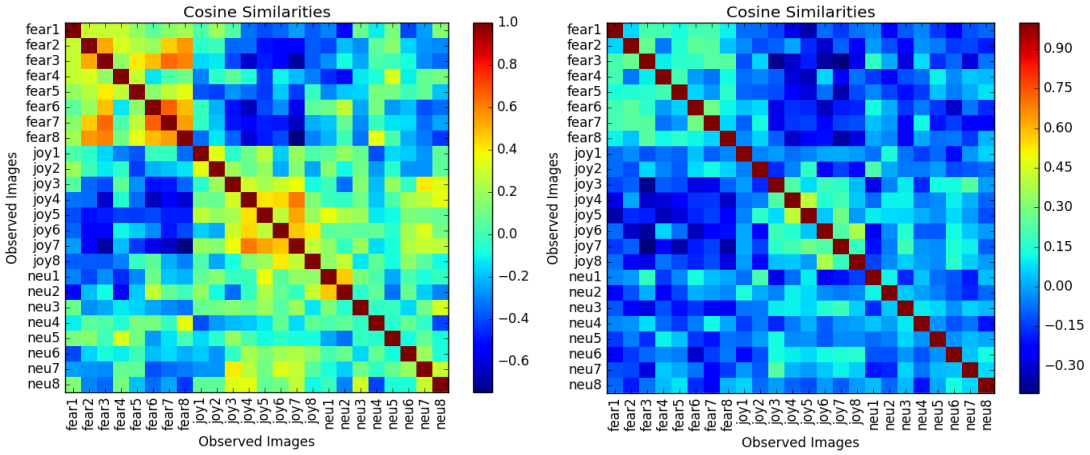
Figure 9. Cosine similarities between the observed fMRI images (t-Test)
(**a**) For 1 participant, (**b**) Averaged over all the participants

Subject No.	Logistic Regression				Support Vector Machine			
	Fear	Joy	Neutral	**Total**	Fear	Joy	Neutral	**Total**
1	100%	50%	100%	**83.3%**	87.5%	100%	87.5%	**91.7%**
2	100%	87.5%	75%	**87.5%**	100%	100%	100%	**100%**
3	87.5%	62.5%	100%	**87.3%**	100%	87.5%	100%	**95.8%**
4	87.5%	87.5%	100%	**91.7%**	100%	87.5%	100%	**95.8%**
5	87.5%	100%	87.5%	**91.7%**	87.5%	100%	100%	**95.8%**
6	100%	100%	100%	**100%**	100%	100%	100%	**100%**
7	100%	100%	87.5%	**95.8%**	100%	100%	100%	**100%**
8	100%	100%	100%	**100%**	100%	100%	100%	**100%**
9	100%	100%	87.5%	**95.8%**	100%	100%	87.5%	**95.8%**
10	62.5%	87.5%	87.5%	**79.2%**	100%	87.5%	100%	**95.8%**
11	100%	87.5%	100%	**95.8%**	100%	100%	100%	**100%**
12	62.5%	62.5%	87.5%	**70.8%**	87.5%	100%	87.5%	**91.7%**
13	100%	100%	75%	**91.7%**	100%	100%	100%	**100%**
14	100%	100%	87.5%	**95.8%**	100%	100%	100%	**100%**
15	87.5%	87.5%	100%	**91.7%**	100%	100%	100%	**100%**
16	100%	87.5%	62.5%	**83.3%**	100%	100%	100%	**100%**
17	100%	87.5%	87.5%	**91.7%**	87.5%	75%	100%	**87.5%**

Table 3. Classification accuracies of the observed fMRI images
(t-Test selection)

4.2.3 Stability Score Selection

Finally, Figure 10 and Table 4 show the cosine similarities and the accuracies obtained by the classifiers, for the stability score method.



(**a**) (**b**)

Figure 10. Cosine similarities between the observed fMRI images (Stability score) (**a**) For 1 participant, (**b**) Averaged over all the participants

Subject No.	Logistic Regression				Support Vector Machine			
	Fear	Joy	Neutral	**Total**	Fear	Joy	Neutral	**Total**
1	25%	50%	25%	**33.3%**	50%	25%	37.5%	**37.5%**
2	87.5%	75%	37.5%	**66.7%**	100%	87.5%	87.5%	**91.7%**
3	50%	50%	25%	**41.7%**	87.5%	75%	50%	**70.8%**
4	50%	50%	50%	**50%**	75%	75%	75%	**75%**
5	25%	37.5%	62.5%	**41.7%**	50%	62.5%	37.5%	**50%**
6	87.5%	87.5%	87.5%	**87.5%**	100%	87.5%	75%	**87.5%**
7	100%	87.5%	37.5%	**75%**	100%	87.5%	75%	**87.5%**
8	87.5%	50%	62.5%	**66.7%**	87.5%	50%	62.5%	**66.7%**
9	37.5%	50%	62.5%	**50%**	62.5%	62.5%	50%	**58.3%**
10	62.5%	37.5%	37.5%	**45.8%**	100%	37.5%	37.5%	**58.3%**
11	62.5%	75%	37.5%	**58.3%**	100%	62.5%	50%	**70.8%**
12	87.5%	62.5%	25%	**58.3%**	87.5%	75%	25%	**62.5%**
13	75%	50%	12.5%	**45.8%**	62.5%	50%	25%	**45.8%**
14	62.5%	25%	37.5%	**41.7%**	25%	12.5%	62.5%	**33.3%**
15	62.5%	75%	62.5%	**66.7%**	100%	50%	87.5%	**79.2%**
16	100%	37.5%	37.5%	**58.3%**	87.5%	37.5%	62.5%	**62.5%**
17	50%	50%	37.5%	**45.8%**	75%	62.5%	50%	**62.5%**

Table 4. Classification accuracies of the observed fMRI images (Stability score selection)

The observed values in figures 8-10 show the similarity for a point at row i, column j, that corresponds to a specific pair of tracks. As already suggested, the high positive values of the blocks along the diagonal indicate that the observed fMRI images within a class are more similar to each other. The mean values among the 17 participants provide a clear distinction for the classes of joy and fear, while neutral stimuli have lower consistency in their responses. This is expected, since the neutral class represents the absence of an emotional reaction, while joy and fear have been reported to evoke specific spatial patterns of brain activation. Although high values can be seen in the case of ANOVA and t-Test methods, stability score is not sufficient in differentiating the tracks, due to the unsupervised nature of its process (image inspection reveals a number of selected voxels with noisy responses and small spatial cohesion, in contrast to ANOVA and t-Test). Fear-evoking tracks, similarly to the music classification task, seem to have the most distinct responses in relation to the other classes (as indicated by the dark blue areas). The overall results are also reflected in the accuracies of the classification problems. ANOVA and t-Test methods produce high accuracies (M: 90-96%), with stability score having significantly lower averages.

4.3 Prediction Model

The evaluation of the computational models was done for each of the 17 participants independently. The process of predicting the fMRI image for a held-out stimulus-track is illustrated in Figure 11. After training the model on the 22 tracks, an image is computed for each of the two test tracks as the weighted sum of the fMRI signatures. Each signature reflects the learned c_{vi} parameters of the 6 used features, as depicted by the voxel colors (red indicates positive values, blue indicates negative values). In the example of Figure 11, features from the literature set are used to predict a joy-evoking fMRI image response. The values of each descriptor are shown left of their respective signatures. At the bottom, the activation value is computed as the linear combination of the 6 signatures, weighted by the descriptor values. The figure shows just one horizontal slice of the brain.

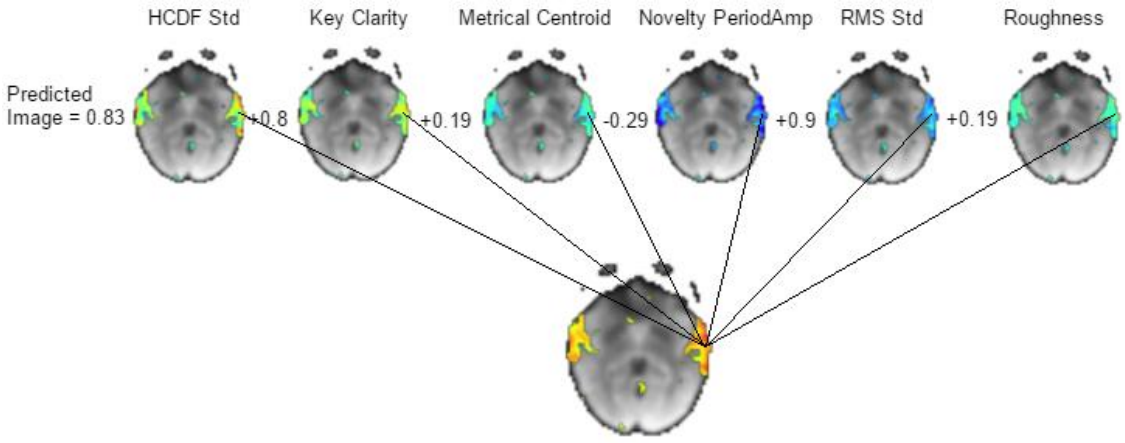


Figure 11. Predicting the fMRI image for a given stimulus-track as a weighted sum of the learned signatures

Figure 12 shows the result of one trial, from the 192 models trained with cross validation (only the horizontal slices with the most selected voxels are depicted). In this example, a successful matching has been achieved by the two held-out test tracks, which belong to the joy and fear classes. The activation patterns reveal the significant BOLD signal differences in the auditory cortex (AC) bilaterally (changes occur in the auditory core, belt and parabelt regions), as expected from previous research [18]. Visual inspection of the produced images among the participants, suggests the auditory cortex as the commonly selected area of interest (mainly derived from the ANOVA and t-Test methods), with the predicted images capturing the substantial activity associated to the emotion. BOLD intensity increases (indicated by red) for joy-evoking tracks, while decreases for the condition of fear (indicated by blue). Neutral tracks show intermediate values around 0 (values correspond to signal percent changes, with respect to grand averages over all stimuli).

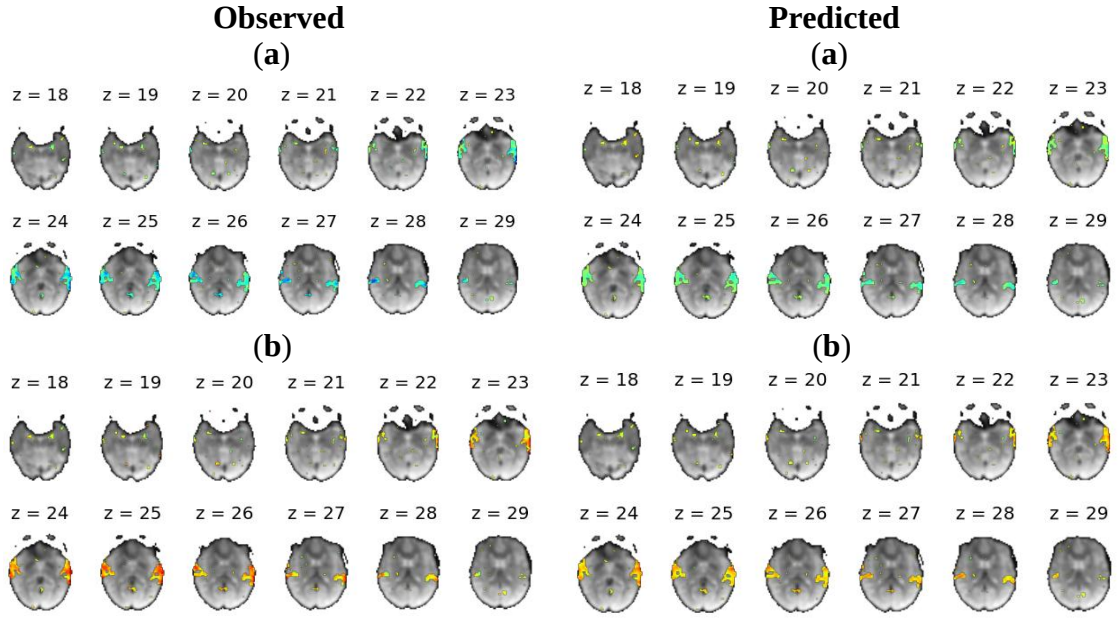


Figure 12. Observed and Predicted fMRI images for two test stimuli
(**a)** Fear-evoking track, (**b**) Joy-evoking track

The cross-validated accuracies in matching the two unseen stimuli-tracks to their unseen fMRI images, were significantly higher than the expected chance accuracy of 50% for all participants. The statistical significance of the literature and fMRI selection feature sets can also be seen in the distribution of accuracies derived from the random feature selection models. As already indicated from the fMRI classification task, stability score resulted in considerably lower prediction rates, due to the inconsistency of the fMRI images. Individual accuracies for each pair of conditions show that joy and neutral tracks are harder to match, with fear being the most distinct class. Literature feature set obtained the highest prediction rate (M=91-93%), with fMRI selection following next (M=85-87%). The fact that the average accuracy of the random selection models is above 50% can be attributed to similar and relevant acoustic information of several audio descriptors. The prediction rates are given in detail in the next sections, for the two feature selection strategies and each voxel selection method separately.

4.3.1 ANOVA Selection

Table 5 shows the prediction model accuracies ("leave-two-out" cross validation) for each feature selection strategy and all 17 participants.

Subject No.	Literature Feature Set				fMRI Selection Feature Set			
	Fear-Joy	Fear-Neutral	Neutral-Joy	**Total**	Fear-Joy	Fear-Neutral	Neutral-Joy	**Total**
1	98.4%	79.6%	81.2%	**86.4%**	92.1%	84.3%	75%	**83.8%**
2	100%	100%	87.5%	**95.8%**	100%	96.8%	43.7%	**80.2%**
3	98.4%	95.3%	76.5%	**90.1%**	100%	92.1%	76.5%	**89.5%**
4	90.6%	90.6%	79.6%	**86.9%**	85.9%	81.2%	59.3%	**75.5%**
5	79.6%	92.1%	90.6%	**87.5%**	78.1%	90.6%	51.5%	**73.4%**
6	96.8%	100%	82.8%	**93.2%**	98.4%	98.4%	56.2%	**84.3%**
7	100%	96.8%	93.7%	**96.8%**	81.2%	96.8%	68.7%	**82.9%**
8	100%	100%	79.6%	**93.2%**	100%	100%	85.9%	**95.3%**
9	95.3%	95.3%	92.1%	**94.2%**	96.6%	95.3%	75%	**89%**
10	98.4%	100%	65.6%	**88%**	96.8%	100%	70.3%	**89%**
11	100%	100%	89%	**96.3%**	100%	96.8%	75%	**90.6%**
12	100%	98.4%	79.6%	**92.7%**	96.8%	90.6%	39%	**75.5%**
13	100%	100%	84.3%	**94.7%**	100%	98.4%	60.9%	**86.4%**
14	89%	96.8%	81.2%	**89%**	93.7%	90.6%	78.1%	**87.5%**
15	98.4%	100%	84.3%	**94.2%**	100%	100%	79.6%	**93.2%**
16	90.6%	93.7%	81.2%	**88.5%**	93.7%	79.6%	78.1%	**83.8%**
17	98.4%	98.4%	84.3%	**93.7%**	100%	100%	70.3%	**90.1%**

Table 5. Prediction model accuracies of literature and fMRI selection feature sets, for the 17 participants
(ANOVA selection)

Figure 13 shows the accuracies of literature and fMRI selection feature sets, in comparison to alternative random feature selection models. The blue histogram shows the distribution of accuracies for 100 models.

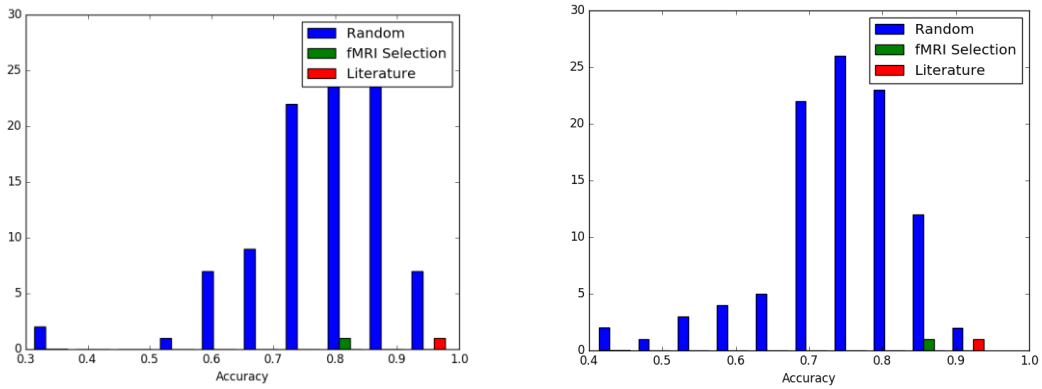


Figure 13. Distribution of accuracies for alternative models (ANOVA)
(**a**) For 1 participant, (**b**) Averaged over all the participants

4.3.2 t-Test Selection

Similarly, Table 6 and Figure 14 show the prediction model accuracies for the t-Test selection method.

Subject No.	Literature Feature Set				fMRI Selection Feature Set			
	Fear-Joy	Fear-Neutral	Neutral-Joy	**Total**	Fear-Joy	Fear-Neutral	Neutral-Joy	**Total**
1	98.4%	92.1%	82.8%	**91.1%**	95.3%	96.8%	79.6%	**90.6%**
2	100%	100%	96.8%	**98.9%**	100%	100%	76.5%	**92.1%**
3	98.4%	98.4%	79.6%	**92.1%**	100%	96.8%	81.2%	**92.7%**
4	93.7%	93.7%	85.9%	**91.1%**	89%	89%	68.7%	**82.2%**
5	75%	90.6%	89%	**84.8%**	68.7%	89%	46.8%	**68.2%**
6	96.8%	100%	82.8%	**93.2%**	98.4%	100%	67.1%	**88.5%**
7	100%	93.7%	93.7%	**95.8%**	96.8%	93.7%	92.1%	**94.2%**
8	100%	100%	89%	**96.3%**	100%	100%	84.3%	**94.7%**
9	69.8%	95.3%	90.6%	**94.2%**	92.1%	95.3%	79.6%	**89%**
10	96.8%	100%	81.2%	**92.7%**	100%	100%	76.5%	**92.1%**
11	100%	100%	90.6%	**96.8%**	100%	93.7%	73.4%	**89%**
12	100%	98.4%	87.5%	**95.3%**	98.4%	90.6%	42.1%	**77%**
13	98.4%	100%	87.5%	**95.3%**	96.8%	92.1%	60.9%	**83.3%**
14	93.7%	96.8%	90.6%	**93.7%**	96.8%	95.3%	79.6%	**90.6%**
15	95.3%	100%	87.5%	**94.2%**	96.8%	100%	76.5%	**91.1%**
16	93.7%	93.7%	85.9%	**91.1%**	87.5%	76.5%	79.6%	**81.2%**
17	92.1%	98.4%	78.1%	**89.5%**	98.4%	96.8%	76.5%	**90.6%**

Table 6. Prediction model accuracies of literature and fMRI selection feature sets, for the 17 participants
(t-Test selection)

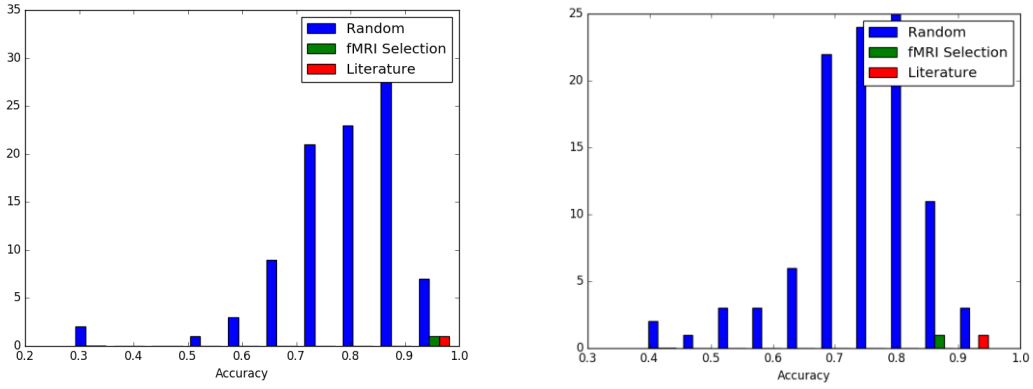


Figure 14. Distribution of accuracies for alternative models (t-Test)
(**a**) For 1 participant, (**b**) Averaged over all the participants

4.3.3 Stability Score Selection

Finally, Table 7 and Figure 15 show the prediction model accuracies for the stability score method.

Subject No.	Literature Feature Set				fMRI Selection Feature Set			
	Fear-Joy	Fear-Neutral	Neutral-Joy	**Total**	Fear-Joy	Fear-Neutral	Neutral-Joy	**Total**
1	70.3%	56.2%	39%	**55.2%**	60.9%	57.8%	68.7%	**62.5%**
2	96.8%	96.8%	54.6%	**82.8%**	92.1%	96.8%	40.6%	**76.5%**
3	85.9%	81.2%	57.8%	**75%**	82.8%	58.9%	65.6%	**78.1%**
4	71.8%	79.6%	53.1%	**68.2%**	78.1%	84.3%	70.3%	**77.6%**
5	28.1%	39%	43.7%	**36.9%**	65.6%	71.8%	62.5%	**66.6%**
6	85.9%	90.6%	48.4%	**75%**	96.8%	82.8%	70.3%	**83.3%**
7	98.4%	82.8%	71.8%	**84.3%**	92.1%	79.6%	37.5%	**69.7%**
8	95.3%	98.4%	43.7%	**79.1%**	93.7%	95.3%	29.6%	**72.9%**
9	70.3%	60.9%	65.6%	**65.6%**	76.5%	57.8%	62.5%	**65.6%**
10	93.7%	96.8%	45.3%	**78.6%**	100%	100%	40.6%	**80.2%**
11	95.3%	89%	54.6%	**79.6%**	82.8%	70.3%	40.6%	**64.5%**
12	79.6%	56.2%	53.1%	**63%**	65.6%	73.4%	51.5%	**63.5%**
13	43.7%	53.1%	42.1%	**46.3%**	65.6%	71.8%	62.5%	**66.6%**
14	46.8%	50%	46.8%	**47.9%**	82.8%	71.8%	75%	**76.5%**
15	64%	85.9%	54.6%	**68.2%**	75%	81.2%	64%	**73.4%**
16	79.6%	65.6%	35.9%	**60.4%**	71.8%	53.1%	65.6%	**63.5%**
17	85.9%	48.4%	50%	**61.4%**	79.6%	76.5%	67.1%	**74.4%**

Table 7. Prediction model accuracies of literature and fMRI selection feature sets, for the 17 participants
(Stability score selection)

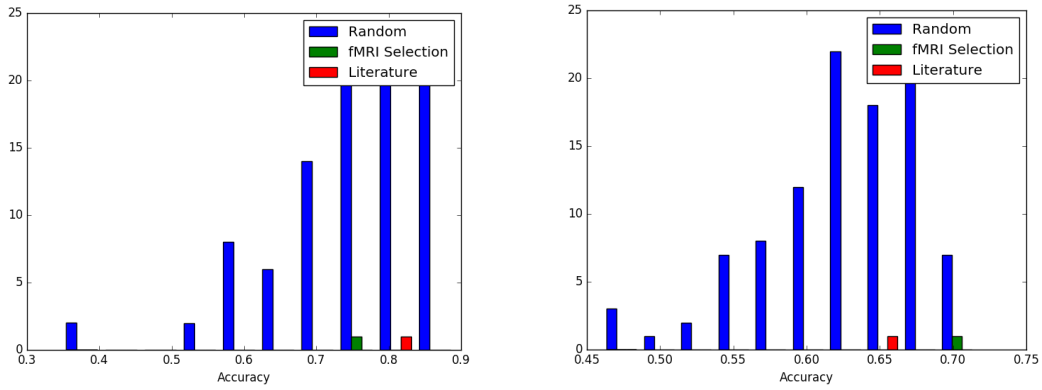


Figure 15. Distribution of accuracies for alternative models (Stability score)
(**a**) For 1 participant, (**b**) Averaged over all the participants

5. DISCUSSION

The attempt to create a computational model that predicts fMRI activity associated with three classes of music-evoked emotions, has been shown to be possible with an approach relying on the content-based analysis of the music's audio signal. A direct predictive relationship can be established between musical features and neural activation, as shown in the results of music and fMRI classification. The prediction model provides insights on how these two representations connect, as the features reveal neural patterns or "signatures" that compose the full fMRI image responses. The trained models succeeded in distinguishing pairs of previously unseen music stimuli, in over 90% of the 192 cross-validated test pairs and across 17 participants. The activation patterns which are shared across the subjects suggest that, given an arbitrary musical stimulus, we can predict the overall brain activity for any individual.

Whether these musical variables can be causally related to the cognitive ones, as measured with the fMRI imaging, is an open problem. Further research and investigation on the neural correlates of the individual musical components, could reveal specific brain regions that are responsible for the encoding of specific elements and concepts. These concepts, which are captured to an extent by audio descriptors, can be improved by incorporating any biological and perceptual properties that transform our understanding and potentially, the encoding in the brain. The idea of analyzing fMRI activity as a result of lower-level features and comparing the similarities in patterns for similar type of stimuli, has previously been studied for other conceptual objects (e.g. predicting picture stimuli based on visual features) with successful results. In this study, the areas indicated by the statistical analysis on the emotional effect regard mainly the auditory cortex (AC) bilaterally, without any other visible structures which are known to play a role.

The success of the model based on the musical features chosen from the literature's state of the art and several machine learning techniques, can be compared to alternative models. Taking the example of the random feature selection model that achieved the highest prediction accuracy, which reaches the accuracy of the 'literature' feature set, 3 out of 6 descriptors have a close relationship under the name and semantic properties that they share (e.g. key clarity can be replaced by the keystrength of a specific pitch class). The fact that several audio descriptors appear during both feature selection strategies, along with their musicological consistency in their relation to the emotions of joy and fear, supports the conjecture of their relevance. Apart from that, any statistical noise inherited from the data must be taken into consideration in any deviations among the results of the competitive models.

The role of the AC has been suggested within the analysis of most of the musical components (as mentioned in studies on pitch, melody, rhythm and tonality perception). For example, it is highly likely that the extraction of the key in tonal music involves the supratemporal cortex bilaterally [59, 60]. The degree in which the observed BOLD contrast during joy and fear evoking stimuli relates to the emotion-specific effects and the musical features, is not clear though. In [18], the observed patterns of activity within the AC are correlated to the subjective feelings (as shown during the emotion ratings) and its role to emotional processing, as indicated through functional interactions with other brain regions (AC as a central hub).

Other approaches to the problem could give further insights regarding this connection. Locating regions of interest and finding voxels with the most accurate responses among the various subjects, could potentially show the involved areas in

encoding the semantic properties of the stimuli. An extension of the current work would be interesting with respect to the number of emotional classes, since the ability of the model to extrapolate multiple semantic categories is weakened, as the number of cognitive states increases.

Moving away from a descriptive theory of music-evoked emotions and their brain correlates, we are able to build models that predict the fMRI activity for arbitrary musical stimuli, and potentially move towards a theory of neural representations. As a restricted form of predictive theory, it could answer how specific semantic features correspond to individual components of neural activation. However, what should be predicted and precisely how remains a challenge in the field.

Bibliography

[1] P. Keisoglou, St & Spirou, “Mathematics-Music: Parallel courses,” *17th Natl. Conf. Greek Math. Soc.*, 2000

[2] Daniel J. Levitin's This Is Your Brain on Music book

[3] Helmholtz, H. L. F. (1885 [1954]). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. 2nd English edition. New York: Dover Publications. [Die Lehre von den Tonempfindungen, 1877. 4th German edition, trans. A. J. Ellis.]

[4] W. A. Sethares, "Local Consonance and the Relationship Between Timbre and Scale"

[5] Dowling, W. J. (2002). "The development of music perception and cognition". *Foundations of Cognitive Psychology: Core Reading*: 481–502.

[6] Gabrielle, A.; Stromboli, E. (2001). "The influence of musical structure on emotional expression". *Music and Emotion: Theory and Research*: 223–243.

[7] Davies, S. (2005). "Artistic Expression and the Hard Case of Pure Music", in: Kieran, M. (Ed.),*Contemporary Debates in Aesthetics and the Philosophy of Art*: 179-91

[8] Hunter, P. G.; Schellenburg, E. G.; Schimmack, U. (2010). "Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions". *Psychology of Aesthetics, Creativity, and the Arts* **4**: 47–56. doi:10.1037/a0016873.

[9] Larsen, J. T.; Stastny, B. J. (2011). "It's a bittersweet symphony: Simultaneously mixed emotional responses to music with conflicting cues". *Emotion* **11**: 1469–1473. doi:10.1037/a0024081.

[10] Jenefer Robinson, Deeper than Reason: Emotion and its Role in Literature, Music, and Art, Oxford: Oxford University Press, 2005; pp. 310-13

[11] Radford, C. (1989). "Emotions and music: A reply to the cognitivists". *The Journal of Aesthestics and Art Criticism* **47**: 69–76. JSTOR 431994.

[12] Sloboda, J. A.; Juslin, P. N. (2001). "Psychological perspectives on music and emotion". *Music and Emotion: Theory and Research*: 79–96.

[13] Ali, S. O.; Peynircioglu, Z. F. (2010). "Intensity of emotions conveyed and elicited by familiar and unfamiliar music". Music Perception: An Interdisciplinary Journal **27**: 177–182. doi:10.1525/MP.2010.27.3.177.

[14] Patrik Juslin & Daniel V¨astfj¨all, ‘Emotional responses to music: The need to consider underlying mechanisms, Behavioural and Brain Sciences, 31, 2008; 559-621.

[15] Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: towards a unified theory of musical emotions. Physics of Life Reviews, 10(3), 235-266.

[16] Hunter, P. G.; Schellenburg, E. G.; Schimmack, U. (2010). "Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions". *Psychology of Aesthetics, Creativity, and the Arts* **4**: 47–56. doi:10.1037/a0016873.

[17] G. A. Wiggins, "Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music," *ISM 2009 - 11th IEEE Int. Symp. Multimed.*, pp. 477–482, 2009.

[18] S. Koelsch, S. Skouras, T. Fritz, P. Herrera, C. Bonhage, M. B. Küssner, and A. M. Jacobs, "The roles of superficial amygdala and auditory cortex in music-evoked fear and joy," *Neuroimage*, vol. 81, pp. 49–60, 2013.

[19] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. K. Chang, V. L. Malave, R. a. Mason, and M. A. Just, "Predicting Human Brain Activity Associated with the Meanings of Nouns (Supplement)," *Science (80-.).*, vol. 320, no. 5880, pp. 1191–5, 2008.

[20] S. Koelsch, "Brain correlates of music-evoked emotions," *Nat. Rev. Neurosci.*, vol. 15, no. 3, pp. 170–180, 2014.

[21] Ekman, Paul (1992). "An argument for basic emotions". *Cognition & Emotion* **6**: 169–200.doi:10.1080/02699939208411068.

[22] Graham, Michael C. (2014). *Facts of Life: ten issues of contentment*. Outskirts Press. p. 63. ISBN 978-1-4787-2259-5.

[23] Russell, J. A.; Barrett, L. F. (1999). "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant". *Journal of Personality and Social Psychology* **76** (5): 805–819.doi:10.1037/0022-3514.76.5.805. PMID 10353204.

[24] Scherer, K. R.; Zentner, M. R. (2001). "Emotional effects of music: production rules". *Music and Emotion: Theory and Research*: 361–387.

[25] Vieillard, S.; Peretz, I.; Gosselin, N.; Khalfa, S. (2008). "Happy, sad, scary, and peaceful musical excerpts for research on emotions". *Cognition and Emotion* **4**: 720–752.doi:10.1080/02699930701503567.

[26] F. Gouyon, P. Herrera, E. Gomez, and P. Cano, "Content processing of music audio signals," *Sound to Sense, Sense to Sound A State Art Sound Music Comput.*, pp. 83–160, 2008.

[27] Juslin, P. N. & 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.

[28] C. Laurier, “Automatic Classification of Musical Mood by Content-Based Analysis,” *Group*, p. 160, 2011.

[29] Scherer, K. R. (1991). Emotion expression in speech and music, pp. 146–156. London: MacMillian.

[30] 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.

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

[32] T. Eerola, O. Lartillot, and P. Toiviainen, “Prediction of multidimensional emotional ratings in music from audio using multivariate regression models,” *Inf. Retr. Boston.*, no. Ismir, pp. 621–626, 2009.

[33] C. Laurier and P. Herrera, “Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines,” *Handb. Res. Synth. Emot. Sociable Robot. New Appl. Affect. Comput. Artif. Intell.*, pp. 9–33, 2009.

[34] R. Panda, B. Rocha, and R. P. Paiva, “Music Emotion Recognition with Standard and Melodic Audio Features,” *Appl. Artif. Intell.*, vol. 29, no. 4, pp. 313–334, 2015.

[35] T. Petri, “Exploring relationships between audio features and emotion in music,” *Front. Hum. Neurosci.*, vol. 3, no. Escom, pp. 260–264, 2009.

[36] http://psysound.wikidot.com/

[37] http://marsyas.info/

[38] Olivier Lartillot, Petri Toiviainen, “A Matlab Toolbox for Musical Feature Extraction From Audio”, *International Conference on Digital Audio Effects*, Bordeaux, 2007.

[39] Wack, N. (2010). Essentia & Gaia: audio analysis and music matching C++ libraries developed by the Music Technology Group. http://mtg.upf.edu/technologies/essentia.

[40] "Magnetic Resonance, a critical peer-reviewed introduction; functional MRI". European Magnetic Resonance Forum. Retrieved 17 November 2014

[41] Huettel, Song & McCarthy (2009, pp. 198–200, 208–211)

[42] Kim et al. (2000, pp. 107–109)

[43] Huettel, Song & McCarthy (2009, pp. 243–45)

[44] Haller S.; Bartsch A. (2009). "Pitfalls in fMRI".*European Radiology* **19**: 2689–2706.doi:10.1007/s00330-009-1456-9.

[45] Grabowski, T., and Damasio, A." (2000). Investigating language with functional neuroimaging. *San Diego, CA, US: Academic Press. 14*, 425-461.

[46] Huettel, Song & McCarthy (2009, pp. 259–62)

[47] Huettel, Song & McCarthy (2009, pp. 262–7)

[48] Todd, N. P. M. & Cody, F. W. Vestibular responses to loud dance music: a physiological basis of the “rock and roll threshold”? J. Acoust. Soc. Amer. 107, 496–500 (2000).

[49] Kandler, K. & Herbert, H. Auditory projections from the cochlear nucleus to pontine and mesencephalic reticular nuclei in the rat. Brain Res. 562, 230–242 (1991).

[50] Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A.D., Koelsch, S., 2009. Universal recognition of three basic emotions in music. Curr. Biol. 19, 573–576.

[51] Mueller, K., Mildner, T., Fritz, T., Lepsien, J., Schwarzbauer, C., Schroeter, M., Möller, H., 2011. Investigating brain response to music: a comparison of different fMRI acqui- sition schemes. NeuroImage 54, 337–343.

[52] Koelsch, S., 2011. Towards a neural basis of music perception—a review and updated model. Front. Psychol. 2, 1–20.

[53] Eerola, T., Toiviainen, P., 2004. Mir in matlab: the midi toolbox. Proceedings of the In- ternational Conference on Music Information Retrieval, pp. 22–27 (Citeseer).

[54] Lohmann, G., Müller, K., Bosch, V., Mentzel, H., Hessler, S., Chen, L., von Cramon, D.Y., 2001. Lipsia — a new software system for the evaluation of functional magnet res- onance images of the human brain. Comput. Med. Imaging Graph. 25, 449–457 (See also at http://www.cns.mpg.de/lipsia).

[55] S. Koelsch, S. Skouras, T. Fritz, P. Herrera, C. Bonhage, M. B. Küssner, and A. M. Jacobs, “The roles of superficial amygdala and auditory cortex in music-evoked fear and joy,” *Neuroimage*, vol. 81, pp. 49–60, 2013.

[56] Agrawal, D., Timm, L., Viola, F.C., Debener, S., Buechner, A., Dengler, R., Wittfoth, M., 2012. ERP evidence for the recognition of emotional prosody through simulated cochlear implant strategies. BMC Neurosci. 13, 113.

[57] Coutinho, E., Dibben, N., 2012. Psychoacoustic cues to emotion in speech prosody and music. Cogn. Emot. 1–27.

[58] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[59] Koelsch, S., Fuermetz, J., Sack, U., Bauer, K., Hohenadel, M.,Wiegel,M., Kaisers, U., Heinke, W., 2011. Effects ofmusic listening on cortisol levels and propofol consumption dur- ing spinal anesthesia. Front. Psychol. 2, 1–9

[60] Liegeois-Chauvel, C., Peretz, I., Babaie, M., Laguitton, V., Chauvel, P., 1998. Contribution of different cortical areas in the temporal lobes to music processing. Brain 121, 1853–1867