



UNIVERSITÀ DEGLI STUDI  
DI TRENTO

---

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE  
ICT International Doctoral School

# INFORMATION RETRIEVAL FROM NEUROPHYSIOLOGICAL SIGNALS

Pouya Ghaemmaghani

Advisor

Prof. Nicu Sebe

Università degli Studi di Trento

---

March 2017



# Abstract

*One of the ultimate goals of neuroscience is decoding someone's intentions directly from his/her brain activities. In this thesis, we aim at pursuing this goal in different scenarios. Firstly, we show the possibility of creating a user-centric music/movie recommender system by employing neurophysiological signals. Regarding this, we employed a brain decoding paradigm in order to classify the features extracted from brain signals of participants watching movie/music video clips, into our target classes (two broad music genres and four broad movie genres). Our results provide a preliminary experimental evidence towards user-centric music/movie content retrieval by exploiting brain signals. Secondly, we addressed one of the main issue of the applications of brain decoding algorithms. Generally, the performance of such algorithms suffers from the constraint of having few and noisy samples, which is the case in most of the neuroimaging datasets. In order to overcome this limitation, we employed an adaptation paradigm in order to transfer knowledge from another domain (e.g. large-scale image domain) to the brain domain. We experimentally show that such adaptation procedure leads to improved results. We performed such adaptation pipeline on different tasks (i.e. object recognition and genre classification) using different neuroimaging modalities (i.e. fMRI, EEG, and MEG). Thirdly, we aimed at one of the fundamental goals in brain decoding which is reconstructing the external stimuli using only the brain features. Under this scenario, we show the possibility of regressing the stimuli spectrogram using*

*time-frequency analysis of the brain signals. Finally, we conclude the thesis by summarizing our contributions and discussing the future directions and applications of our research.*

## **Keywords**

[pattern recognition, brain decoding, neurophysiological signal processing, information retrieval]

# Acknowledgements

*I would like to express my sincere gratitude to my advisor Prof. Nicu Sebe for his constructive guidance and continuous support during my PhD research. I must say that i am very fortunate to have the opportunity to do research under his supervision who incented me to widen my research from various perspectives. Without his precious support, it would not be possible to conduct this research.*

*A special thanks to my parents for their unconditional love and encouragement in all my pursuits. No words can express how utterly grateful i am to you. Thanks for all you've done for me.*



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research Motivation . . . . .	2
1.2	Thesis Structure . . . . .	3
<b>2</b>	<b>Concepts and Fundamentals</b>	<b>5</b>
2.1	Signal Acquisition . . . . .	6
2.1.1	Invasive Methods . . . . .	6
2.1.2	Non-invasive Methods . . . . .	7
2.1.3	Spatio/Temporal Resolution of brain signal acquisition	10
2.2	Signal Processing . . . . .	11
2.2.1	Evoked Response . . . . .	12
2.2.2	Event-related desynchronization/synchronization .	13
2.2.3	Spontaneous Brain Activity . . . . .	14
2.3	Feature Extraction . . . . .	15
2.3.1	Time-domain Features . . . . .	16
2.3.2	Frequency-domain Features . . . . .	16
2.3.3	Time-Frequency Features . . . . .	16
2.4	Decoding Pipeline . . . . .	17
2.4.1	Definition . . . . .	17
2.4.2	Opportunities . . . . .	17
2.4.3	Challenges . . . . .	19
2.5	Summary . . . . .	21

<b>3</b>	<b>Retrieving Genre related Information from Brain</b>	<b>23</b>
3.1	Introduction . . . . .	23
3.2	Related Works . . . . .	25
3.2.1	Content-Centric Music/Movie Genre Classification .	25
3.2.2	Genre and Affective Content . . . . .	26
3.2.3	Affective Contents and Neurophysiological Signals .	27
3.2.4	Spotting the gap . . . . .	28
3.3	Experimental Setup . . . . .	28
3.3.1	Datasets . . . . .	28
3.3.2	Movie/Music Clips Annotation . . . . .	29
3.3.3	Feature Extraction . . . . .	31
3.4	Results and Discussion . . . . .	33
3.4.1	Correlation Results . . . . .	33
3.4.2	Classification Results . . . . .	34
3.5	Conclusion . . . . .	45
<b>4</b>	<b>Domain Adaptation</b>	<b>47</b>
4.1	Introduction . . . . .	47
4.2	Related Works . . . . .	48
4.2.1	Cross-Modal Domain Adaptation . . . . .	48
4.2.2	Domain Adaptation in Brain Studies . . . . .	49
4.3	Study 1: Object Recognition . . . . .	50
4.3.1	Experimental Setup . . . . .	51
4.3.2	Adaptation Method . . . . .	55
4.3.3	Experiments and Results . . . . .	57
4.4	Study 2: Genre Classification . . . . .	65
4.4.1	Materials and Method . . . . .	65
4.4.2	Adaptation Method . . . . .	67
4.4.3	Experiments and Results . . . . .	69



4.5	Conclusion . . . . .	71
<b>5</b>	<b>Towards Mind Reading</b>	<b>73</b>
5.1	Literature Review . . . . .	74
5.1.1	Reconstructing image-based stimuli . . . . .	75
5.1.2	Reconstructing audio-based stimuli . . . . .	76
5.1.3	Spotting the gap . . . . .	78
5.2	Materials and Methods . . . . .	78
5.2.1	Dataset . . . . .	78
5.2.2	Data Analysis . . . . .	79
5.2.3	Regression Analysis . . . . .	80
5.2.4	Correlation Analysis . . . . .	81
5.3	Results . . . . .	81
5.4	Discussion . . . . .	81
5.5	Conclusion . . . . .	83
<b>6</b>	<b>Conclusion</b>	<b>85</b>
6.1	Summary . . . . .	85
6.2	Limitations and Future works . . . . .	87
6.2.1	Music/Movie genre classification . . . . .	87
6.2.2	Domain Adaptation . . . . .	88
6.2.3	Stimuli Reconstruction . . . . .	88
	<b>Bibliography</b>	<b>91</b>



# List of Tables

3.1	Extracted audio-visual features from each movie clip (the number of features is listed in the parenthesis). . . . .	33
3.2	Movie clip titles, ground-truth labels, and predicted labels. The accuracies are obtained by employing the Naive Bayes Classifier. . . . .	40
3.3	Movie clip titles, ground-truth labels, and predicted labels. The accuracies are obtained by employing the Linear SVM Classifier. . . . .	41
3.4	Comparison between the accuracy of MEG, EEG and MCA descriptors with random inputs in the single-subject level scenario. . . . .	42
3.5	Music clip titles, ground-truth labels, and predicted labels of different feature descriptors. The accuracies are obtained by employing the Linear SVM Classifier. . . . .	43
3.6	Music clip titles, ground-truth labels, and predicted labels of different feature descriptors. The accuracies are obtained by employing the Naive Bayes Classifier. . . . .	44
4.1	7-class Classification Accuracy (average accuracy over 12 runs for each subject on each mask area) . . . . .	61
4.2	Average accuracy of all subjects over all folds for each threshold. . . . .	63

4.3	Seven-Class Classification Accuracy (average accuracy over folds for each subject). . . . .	64
4.4	Comparison between the accuracy of Brain features and Adapted-Brain features in the population-level analysis. . .	71
5.1	Correlation analysis. . . . .	82

# List of Figures

2.1	functional magnetic resonance imaging (fMRI) . . . . .	8
2.2	Electroencephalography (EEG) . . . . .	9
2.3	Magnetoencephalography (MEG) . . . . .	11
2.4	The spatio/temporal resolution of the neuroimaging methods.	12
2.5	The P300 wave. A series of ERP components precedes the P300 and they reflect low-level automatic processing of stimuli.	14
2.6	(a) Phase-locked early gamma response to the visual stimulus. (b) Averaging evoked potentials in time domain over trials. The non-phase-locked activity cancels out as a result of averaging. (c) Time-frequency power representation of the evoked gamma response. (d) Time-frequency power computed for each trial. (e) Average of time-frequency powers across all trials. The induced gamma response is visible. [10] . . . . .	15
2.7	Brain Decoding Pipeline: Different stimuli regarding different categories (i.e. house vs. face) are shown to the participant while his/her brain activity is recorded simultaneously. Then a classifier is employed to classify the recorded data into the target stimulus classes. If the classifier performs above chance on the test set, it can be concluded that the stimuli related activities are encoded in the brain signal. . . . .	18
3.1	The framework used in this study regarding the movie/music genre classification by exploiting brain signals. . . . .	25

3.2	Pearson correlation analysis between the MEG responses and audio-visual features. Correlation over each channel is denoted by the gray level, and significant correlations are marked with red $\star$ . . . . .	34
3.3	Comparison between the accuracy of MEG and multimedia features (MCA) with random inputs in the single-subject scenario. (a) using a Naive Bayes classifier. (b) using a Linear SVM classifier. . . . .	37
3.4	Confusion matrix for four-class genre classification using multimedia and MEG features. x and y axes represent predicted and actual labels, respectively. The top row represents the confusion matrices obtained using the Naive Bayes classifier. The bottom row represents the confusion matrices obtained by employing the Linear SVM classifier. . . . .	38
4.1	Domain Adaptation Pipeline. . . . .	53
4.2	(a) The transformation of each subject’s brain anatomy to the standard space (i.e., the MNI space). (b) The shape of Haxby’s VTC area obtained for each subject individually in the MNI space. As it is expected, the shape of VTC in this univariate approach is different for each subject. (c) The shape of the VTC area in the MNI space which is acquired by a brain atlas. . . . .	54
4.3	The average accuracy of the SVM classifier over all folds for each principle components. . . . .	60
4.4	The VTC area in the brain (MNI space) using different value of threshold. . . . .	62
4.5	7-class Classification Accuracy (average accuracy over all runs for each subject using different threshold values). . . . .	62

4.6	Normalized Confusion Matrices. (a) baseline method (before adaptation). (b) proposed method (Adapted-Features).	64
4.7	Overview of our proposed framework: During training, a dictionary learning approach is used to learn a mapping function for brain/multimedia adaptation. Once the mapping function is learned, the genre of a test movie clip is predicted using the adapted brain features. . . . .	65
4.8	Comparison between the accuracy of Brain and Adapted-Brain features in classifying the genre of the music/movie clip in the single-subject level scenario on different datasets. (a) Using a Linear SVM Classifier. (b) Using a Naive Bayes Classifier. . . . .	70
5.1	<b>Stimuli reconstruction pipeline:</b> In the first step of the analysis the spectrogram of the EEG signal and audio wave (after generic preprocessing steps) is computed. Then In the training phase, the EEG features are regressed (Ridge Regression) onto the audio spectrogram to find the proper mapping function. After that, the resulting weight matrix is used on the test EEG data to predict the spectrogram of the audio wave. . . . .	74
5.2	Auditory cortex in human brain . . . . .	80





# Chapter 1

## Introduction

In the past decade, machine learning algorithms have been widely used in neuroscience community. Extracting stimulus-related information from the brain activity by employing machine learning algorithms is known as brain decoding. In a typical brain-decoding paradigm, different types of stimuli are shown to the participant of the neuroimaging experiment, while his/her concurrent brain activity is captured using neuroimaging techniques. Then a machine learning algorithm is employed to categorize the measured brain signal into the target stimuli classes. If the algorithm, can predict the target stimulus category better than the chance level, we can hypothesize that the stimulus-related information exists in the brain data. Among various neuroimaging techniques for recording brain activity, the most widely used methods for noninvasive brain recording in humans are Functional Magnetic Resonance Imaging (fMRI), Magnetoencephalography (MEG) and Electroencephalography (EEG). Once brain signals are recorded, the aforementioned decoding systems can be applied to the measured signal.

This research aims at retrieving semantic-level information from complex stimuli and improving the accuracy of brain decoding systems. In the remainder of this chapter we present the motivation and the outline of the research conducted in this thesis.

## 1.1 Research Motivation

This thesis focuses on addressing some of the current issues in brain decoding and brain signal processing. The main objectives of this research are listed as follows:

1. **Retrieving the music/movie genre related information from neurophysiological signals:** The first objective of this thesis, is to propose a decoding pipeline in order to classify features extracted from brain signals of participants into our target classes (two broad music genres and four broad movie genres). Our results provide an evidence towards the feasibility of the user-centric music/movie content retrieval by exploiting brain signals. This can serve as the first preliminary step towards user-centric music/movie recommender systems.
2. **Improving the performance of brain decoding algorithms using an adaptation paradigm:** The second objective of this thesis is related to the poor performance of employing machine learning algorithms on neuroimaging datasets since such datasets suffer from the constraint of having few and noisy samples. On this, we employed an adaptation paradigm in order to transfer knowledge from another modality to the brain modality. We experimentally show that such adaptation procedure leads to improved results on the neuroimaging datasets.
3. **Reconstructing external stimuli:** The third objective of this thesis aims at one of the ultimate goal of neuroscience which is reconstructing

individuals' experience using their brain signals. Regarding this, we tried to reconstruct the stimulus spectrogram using time-frequency analysis of brain signals. Our approach is a preliminary step towards a complete reconstruction of stimuli using only brain signals.

## 1.2 Thesis Structure

This thesis is organized into six chapters. Chapter 2 contains the background material. It gives an overview of some of the brain signal acquisition techniques, signal processing and feature extraction methods. Furthermore, our brain decoding pipeline is also introduced and some of its applications as well as the current issues are also discussed.

Chapter 3 describes a method to address the specific problem of movie/musical genre classification by exploiting features extracted from the brain activity. The proposed algorithm is employed on two different datasets and yields significant results on both of them. This chapter is based on the papers we published in [28, 31].

Chapter 4 proposes a novel method for improving the performance of “brain decoding” based on “domain adaptation” by transferring knowledge learned in another modality (e.g. large-scale image domain) to the brain modality. We employ this domain adaption pipeline for three different tasks (i.e., object recognition, music genre classification and movie genre classification) using three different neuroimaging modalities (i.e., fMRI, EEG and MEG). Our results show a performance boost in all cases. This chapter is based on the papers we published in [30, 29].

Chapter 5 describes a method for reconstructing the external stimuli using the features extracted from brain signals. The proposed method, is based on regressing the spectrogram of the stimuli using time-frequency analysis of brain signals. From the study in this chapter, we are currently

prepare a submission for EUSIPCO 2017.

Finally, Chapter 6 provides a summary of the contributions and highlights some possible future directions.

# Chapter 2

## Concepts and Fundamentals

Reading someone’s mind has been for many years the domain of science fiction. Recently however, after all new discoveries about the brain, “Mind Reading” has become the province of science [60]. In fact, a challenging goal in neuroscience is decoding mental contents from brain activities. Recent progress in neuroimaging suggests the possibility of brain decoding [38, 79, 126, 105, 52, 81, 137, 82]. Typically, a brain decoding pipeline contains the following steps:

1. **Signal Acquisition:** The first step aims at capturing the brain activity of subjects while they are doing a specific task (e.g. passive observing of different categories of objects). There are numerous types of neuroimaging techniques for recording the brain activity. Each method has its own advantages and disadvantages. In Section 2.1, we briefly review some of the most commonly used signal acquisition techniques.
2. **Signal Processing:** The second step of brain decoding involves the processing of the recorded signal. However, brain signals have several characteristics. These characteristics are sometimes task dependent. We review such characteristics in Section 2.2.

3. **Feature Extraction:** The third step towards brain decoding is feature extraction in order to represent discriminative information regarding a specific task. In Section 2.3, we review some of the most commonly used feature extraction methods.
4. **Classification:** The last step of decoding is feeding the extracted features into a classifier to identify the task. Accurate prediction of the classifier suggests the existence of the task-related information in the brain.

In this chapter we review the above-mentioned steps in more details and we discuss some of the current issues in brain decoding.

## 2.1 Signal Acquisition

In neuroimaging studies, different neuroimaging techniques are used for measuring the brain activity. These techniques are categorized into two broad categories: invasive methods and noninvasive methods, each one having its own advantages and disadvantages. Below, we briefly discuss these techniques.

### 2.1.1 Invasive Methods

The most direct approach for measuring brain activity is by implanting electrodes under the scalp. This is done by neurosurgery. The main advantage of invasive recordings is the high quality signals since they provide very high temporal and spatial resolution and high signal to noise ratio [122]. However, these techniques have many issues that make them inapplicable in many cases. These methods cause significant health risks and discomfort to the users due to their invasiveness. The brain tissues

might get infected or might fail to accept the microelectrode as the foreign substance [59, 88]. Besides, these methods generally cover only a small region of the brain since it is not feasible to put electrodes covering the whole brain. Furthermore, the signal quality deteriorates over time [68]. Thus the usage of invasive methods in real world applications is usually restricted to disabled people and animal studies [23].

### **Intracortical Signal Acquisition**

Intracortical acquisition technique is the most invasive method that measures brain activity inside the gray matter of the brain. In this technique, the microelectrode arrays are implanted inside the cortex to capture spike signals and local field potentials from individual or multiple neurons [88].

### **Cortical Signal Acquisition**

Electrocorticography (ECoG) is a recording method that uses electrodes placed over the exposed surface of the brain through a surgical operation, to measure electrical activity from the cerebral cortex. Since subdural electrodes are not implanted inside the gray matter, the ECoG does not have the high surgical risk and user discomfort as the intracortical signal acquisition methods.

#### **2.1.2 Non-invasive Methods**

Contrary to the invasive methods, non-invasive methods do not demand implanting electrodes in the brain via surgical operations. Thus surgery-related problems are avoided. Different non-invasive modalities have been proposed over the past years. Generally, the type of the task that is going to be decoded determines the modality to be used, since the captured signal

by each modality has different spatial/temporal resolution and signal-to-noise ratio. Here we briefly review some of the most common methods.

### Functional magnetic resonance imaging (fMRI)

Functional magnetic resonance imaging (fMRI) is a neuroimaging technique (Figure 2.1) that measures brain activity indirectly by detecting changes associated with Blood Oxygen Level (BOLD signal). This technique is based on the coupling of changes in the cerebral blood flow with neural activity so that using a specific part of the brain increases the blood flow to that region [75, 45]. The main advantage of fMRI is its high spatial resolution which makes it a proper tool for localizing active regions inside the brain. However, fMRI has a low temporal resolution (in order of 1 or 2 seconds). Besides, the hemodynamic response has a physiological delay (from 3 to 6 seconds). This low temporal resolution makes fMRI inappropriate for rapid/real-time BCI systems [88].



Figure 2.1: functional magnetic resonance imaging (fMRI)



**Electroencephalogram (EEG)**

Electroencephalography (EEG) is a non-invasive neuroimaging technique that can be easily used by placing the electrodes on the top of the scalp (Figure 2.2). EEG measures the brain activity caused by the electrical currents flow during the synaptic transmission [6, 88]. However, the signal quality is weak due to the fact that the signals have to pass different layers (e.g. the scalp and the skull) [88]. Nevertheless EEG has a very high temporal resolution (in order of milliseconds). Besides it is not bulky neither expensive (compared to the other acquisition methods). As a result of this, EEG is the most commonly used brain signal acquisition method for non-invasive BCI systems (specifically real-time BCI).



Figure 2.2: Electroencephalography (EEG)

### Magnetoencephalography (MEG)

Magnetoencephalography is a functional neuroimaging technique that measures the magnetic fields produced by electrical currents in the brain. The neurophysiological processes that make MEG signals are very similar to those that make EEG signals. Both signals are obtained during the synaptic transmission in the dendrites [36, 88]. These electrical currents produce a magnetic field outside the brain which can be captured by using the arrays of SQUIDs (the superconducting quantum interference device). However, the MEG signals can be affected by the other magnetic sources. Thus, in order to diminish the effects of the external magnetic fields, the recordings need to be acquired in a magnetically shielded room. Figure 2.3 shows a MEG device. The advantage of MEG, compared to EEG, is that magnetic fields are less affected by the skull and scalp. Besides MEG has higher spatial resolution than EEG which makes it suitable to localize certain types of activities in the brain. In spite of these advantageous, MEG is rarely used in BCI systems, since the device is too expensive and bulky and this makes it not a suitable technique for many applications [88].

#### 2.1.3 Spatio/Temporal Resolution of brain signal acquisition

Figure 2.4 compares the spatial and the temporal resolution of various neuroimaging methods. As shown in this figure, invasive methods (e.g. ECoG) have higher temporal and spatial resolution compared to the non-invasive methods. Among non-invasive methods, fMRI has higher spatial resolution in comparison with MEG and EEG. However the temporal resolution of fMRI is lower. MEG and EEG both have high temporal resolution. However their spatial resolution is relatively low.

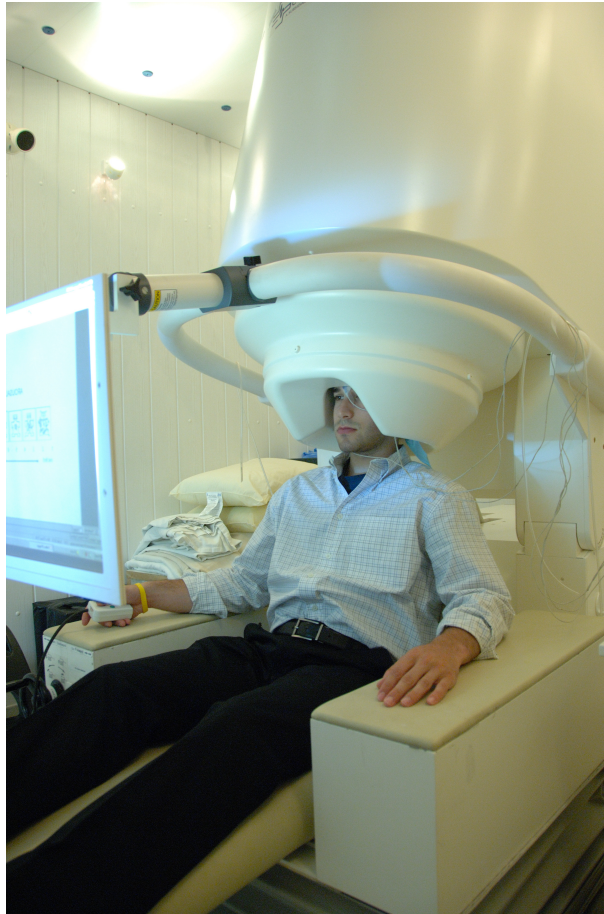


Figure 2.3: Magnetoencephalography (MEG)

## 2.2 Signal Processing

Neurophysiological signals involve numerous simultaneous phenomena related to the cognitive tasks. Most of them are still incomprehensible and their origins are unknown [88]. However the analysis of the neurophysiological signals has revealed several types of characteristics in the brain activity patterns. Below, we briefly discuss the main categories of brain responses.

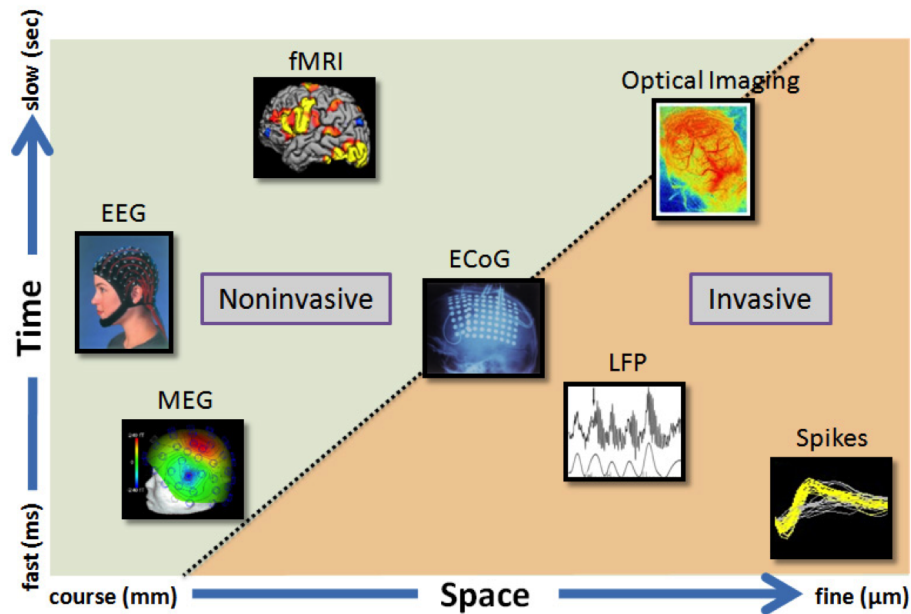


Figure 2.4: The spatio/temporal resolution of the neuroimaging methods.

### 2.2.1 Evoked Response

Evoked responses are automatic brain activities in response to the presentation of certain types of stimuli<sup>1</sup>. These activities are time-locked and phase-locked to the stimulus onset. Thus they are named by their latency and their amplitude. For instance, P300, is a positive event-related potential that is happening 300 milliseconds after the stimulus presentation. Since, the amplitude of such responses is very low (in order of microvolts), low-pass filtering and signal averaging are usually required in order to differentiate the ERP components from the background noise.

#### Slow Cortical Potential

Slow Cortical Potentials (SCP) are event-related potentials that represent the slow voltage changes in the very low frequency bands (below 2 Hz).

<sup>1</sup>Once such responses are measured using EEG, they are called event-related potential (ERP) and once they are captured by MEG, they are called event-related field (ERF).

These activities last from hundreds milliseconds to several seconds [128, 88]. Since the voltage-change in SCPs is very weak, it is necessary to average many trials to obtain the overall trend of the EEG activity. Although analyzing SCPs reveals high inter-subject variability [42], SCPs are shown to be useful in BCI systems regarding moving a cursor on the computer screen [88].

### **Steady-State Evoked Potentials**

Steady State Evoked Potentials (SSEP) are activities evoked in response to periodic stimuli. The stimulus could be either visual, auditory or somatosensory. In case of visual stimuli (flickering light), SSEP evokes a sinusoidal-like waveform with the same fundamental frequency of the stimulus [125, 88].

### **P 300**

The P300 is an event-related potential that has gained increasing attention in the literatures. These potentials are positive peaks that are elicited from EEG approximately 300 ms after the presentation of a rare or unexpected stimuli [25, 88]. This potential is mainly generated in an odd-ball scenario when users see frequent and non-frequent visual stimuli. The appearance of non-frequent items leads to a P300 response in sensors located in the parietal area. Figure 2.5 demonstrates P300 response.

### **2.2.2 Event-related desynchronization/synchronization**

Apart from the evoked responses that are phase-locked changes in the brain activity, there are some other types of neural oscillations that are not phase-locked to the stimulus onset [53, 101]. Such oscillations are not delectable by simply averaging the signal (which is the case in ERP), but might be

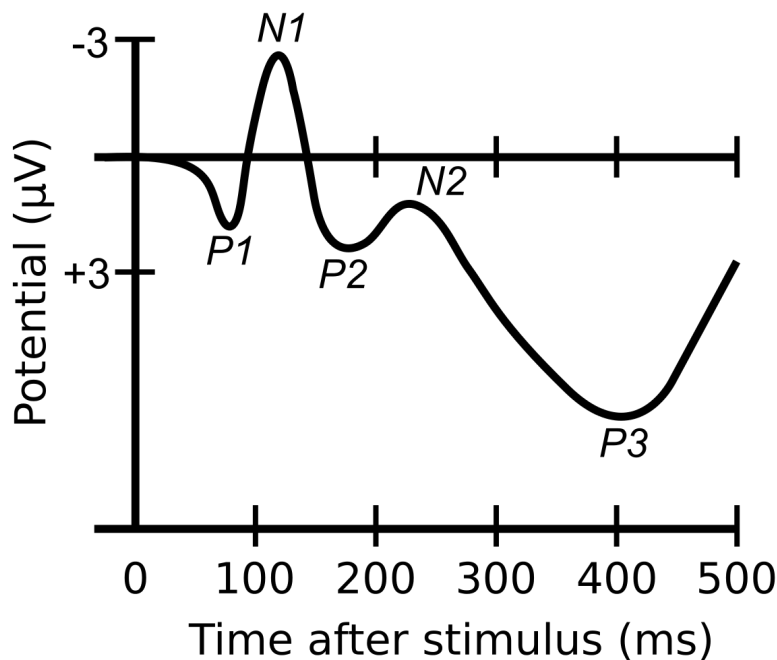


Figure 2.5: The P300 wave. A series of ERP components precedes the P300 and they reflect low-level automatic processing of stimuli.

obtained by frequency analysis. These event-related events represent frequency specific changes of brain activity. The decrease in the signal power is called as "event-related desynchronization" (ERD) and the increase in the signal power is called as "event-related synchronization" (ERS) [101]. These non-phase-locked activities (ERD/ERS), in some literatures are referred to as "induced activity" [19]. Figure 2.6 shows the evoked and the induced responses of a brain activity to a visual stimulus.

### 2.2.3 Spontaneous Brain Activity

The above-mentioned brain responses are brain activities induced by a specific task. The brain activity in the absence of an explicit task is called spontaneous brain activity (also referred to as "resting-state" activity or "ongoing brain" activity). Such activities are generally considered as "noise" where one wants to investigate the brain responses of a given

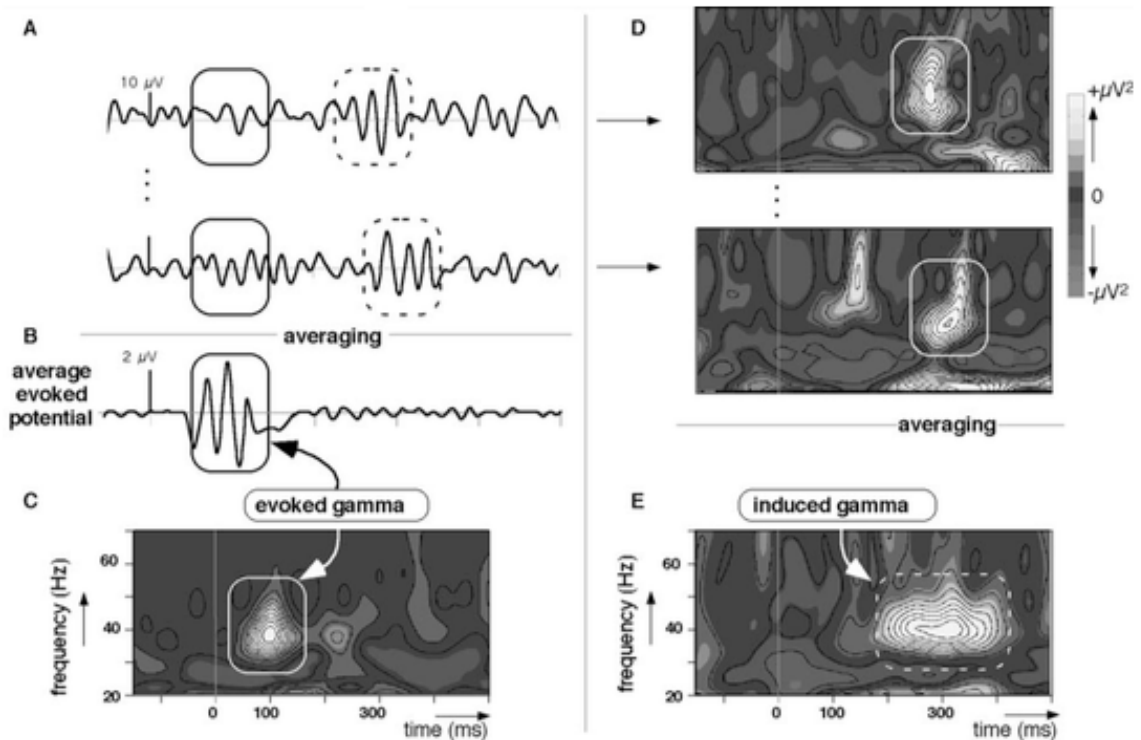


Figure 2.6: (a) Phase-locked early gamma response to the visual stimulus. (b) Averaging evoked potentials in time domain over trials. The non-phase-locked activity cancels out as a result of averaging. (c) Time-frequency power representation of the evoked gamma response. (d) Time-frequency power computed for each trial. (e) Average of time-frequency powers across all trials. The induced gamma response is visible. [10]

task. However, they play a crucial role in our perception and understanding [124, 13].

## 2.3 Feature Extraction

Different tasks result in different activity patterns in the brain signals. Brain decoding is considered as a pattern recognition system that classifies each pattern into a target class according to its features [88]. Once the brain signal is captured, we need to extract features in order to represent the raw neurophysiologic signals into representations that contain the dis-

criminative information needed for that task. However, extracting a set of suitable features is a challenging issue since the information of interest is not easily obtainable as a result of the highly noisy environment. In this section, we review some of the most common feature extraction methods.

### **2.3.1 Time-domain Features**

As mentioned in the previous section, the presentation of certain types of stimuli results in phased-locked alterations in the amplitude of neurophysiological signals at very specific time intervals (e.g. P300). In order to use such information, we need to extract temporal features which are typically the temporal variations of the signal amplitude. However some low-pass filtering is generally needed prior to the feature extraction in order to separate the task-related pattern from the background activity.

### **2.3.2 Frequency-domain Features**

Some tasks induce changes in neurophysiological signals that are not phase-locked to the stimulus onset (e.g. ERS/ERD). Since such oscillations are not phase-locked, temporal features are not useful. Instead, features that are invariant to the stimulus onset shall be used. These changes can be captured from the signal power over specific frequency bands.

### **2.3.3 Time-Frequency Features**

Due to the importance of the information encoded in the both domains (Time and Frequency), time-frequency analysis is performed in order to take into account the information encoded in both domains. These time-frequency features can be estimated using the short-term Fourier transform or wavelets. In both cases, the signal is divided into smaller sequential segments (there might be overlaps between the segments). Then the signal



power is estimated for each segment. The output of such analysis provides a time-frequency representation of the signal.

## 2.4 Decoding Pipeline

One of the ultimate goals of neuroscience and brain studies is decoding someone’s intentions from his/her brain activities. In this section, we first, formulate “brain decoding” and then we review the opportunities and the challenges of such analysis.

### 2.4.1 Definition

Prior works on brain decoding have mostly focused on the classification of the stimuli into a set of pre-defined categories [38, 18, 79, 126, 14, 81, 82]. A typical classification pipeline in neuroscience includes the following steps: First, different stimuli regarding different categories are presented to the participant of the experiment, while his/her concurrent brain activity is recorded (using any of the neuroimaging methods). Once the signal is captured and the features are extracted, a machine learning algorithm is trained on the subset of the samples in order to differentiate different categories of stimuli using the extracted brain features. Accurate prediction of the algorithm in the remaining subset (test-set) is considered as a positive evidence of the hypothesis of the existence of the stimulus-related information in the brain data. Figure 2.7 demonstrates such pipeline.

### 2.4.2 Opportunities

During the past decade, machine learning algorithms have been widely used in the neuroscience community to analyze and interpret neuroimaging datasets. Such investigations revealed new insights regarding brain

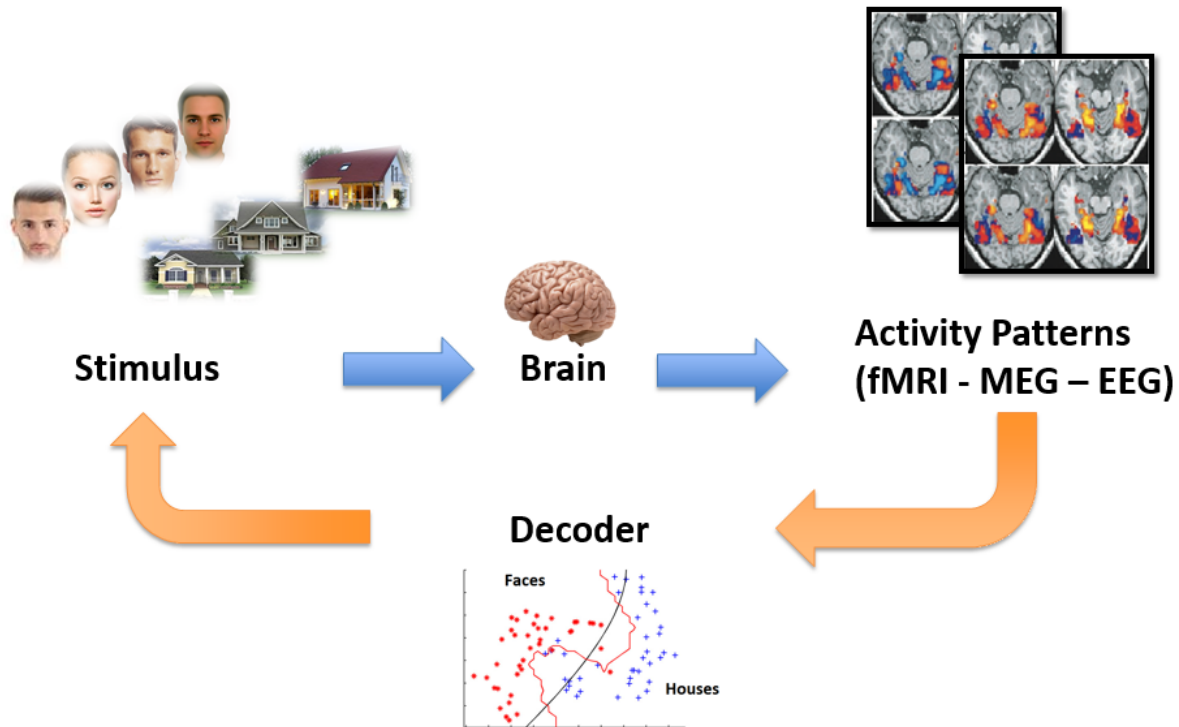


Figure 2.7: Brain Decoding Pipeline: Different stimuli regarding different categories (i.e. house vs. face) are shown to the participant while his/her brain activity is recorded simultaneously. Then a classifier is employed to classify the recorded data into the target stimulus classes. If the classifier performs above chance on the test set, it can be concluded that the stimuli related activities are encoded in the brain signal.

functioning and provide practical applications for brain signals.

### New Insights about Brain

One of the most fundamental question in neuroscience deals with the issue of representation: what information is represented in brain; and how it is represented [90]. Researchers tackle this issue in different ways. Conventional brain analysis methods (Univariate approach) have focused on characterizing the relationship between cognitive states and a specific brain region (i.e. individual brain voxels in an fMRI study). However the information might be distributed over different brain regions. A study by

Haxby et al. [38] illustrated how multi-voxel patterns of activity can distinguish different categories of objects (faces, houses, chairs, shoes, bottles, scissors, cats). They showed that each object category has distinct activity pattern in ventral temporal cortex (VT). This work has been extremely influential and since then, brain decoding algorithms have been widely used in order to discover the hidden information encoded in the brain [18, 79, 94, 55, 54, 40, 39, 41, 91].

### **Applications**

Apart from finding new insights about the encoded information in the brain, decoding brain activity has received substantial attention in Brain Computer interfacing (BCI) and rehabilitation communities particularly specifically due to its potential for helping disabled people [11, 129]. High accuracy of brain decoding systems acquired with versatile techniques such as EEG will ultimately allow subjects to interact with the external world via their mind [123].

### **2.4.3 Challenges**

Employing pattern recognition methods to neuroimaging datasets is also challenging in various aspects including low signal-to-noise ratio, non-stationarity, small sample size and high dimensionality. Below, we review some of the important ones.

#### **Low Signal-to-Noise Ratio and Non-stationarity**

Typically, the signal-to-noise ratio of the brain signal acquisition techniques, particularly the non-invasive methods, is very low. Such noisy signals might yield unwanted results when employing machine learning algorithms. Thus, it is important to pre-process the signal very carefully. Be-

sides, the captured signal is not stationary. Such non-stationarity changes the brain activity patterns (for a specific task) from trial to trial. There are several factors that might contribute to such non-stationarity including: changes in subject's state (e.g. fatigue), artifacts and also changes in the placement of the electrodes.

### **Small Sample Size**

Neuroimaging datasets, in general, are relatively small in-terms of the number of samples. This is mainly due to the cost of recording brain signals and the subject's fatigue since the recording sessions are time consuming and demanding for the subjects. This small number of samples drastically decreases the performance of the machine learning algorithms.

### **High Dimensional Data**

Another issue regarding neuroimaging datasets is due to their dimensionality in space/time resulting in enormous number of features. In order to cope with such difficult issue, various feature reduction methods are proposed. However not all of these methods are applicable since, in this case (neuroimaging datasets), the number of observations is very low (much lower than the number of features) yielding inconsistent results. Thus a significant challenge in brain decoding is dealing with such high-dimensional few-samples datasets.

### **Interpretation of the Results**

Since, one of the aim of brain decoding is obtaining new insights about brain, the interpretation of the results (obtained by employing machine learning algorithms on the neuroimaging datasets) is a crucial step and should not be overlooked [123]. It is important to understand why a specific

feature selection method and a certain classifier yields a specific result rather than dealing with them as “black box”.

## 2.5 Summary

One of the main goals of neuroscience and brain studies is decoding someone’s thoughts from his/her brain activities. In this chapter, we reviewed all the steps of a brain decoding pipeline. These steps include: signal acquisition, feature extraction, and classification. Moreover, we surveyed the opportunities and the challenges of such analysis.

In the next chapter, we will employ the brain decoding pipeline in order to classify the music/movie clips into target genre classes. If the decoder can predict the target genre class better than the chance level, we can hypothesize that the genre-related information exists in the brain data. Such analysis can show the feasibility of a user-centric music/movie recommender system by exploiting the brain signals.



# Chapter 3

## Retrieving Genre related Information from Brain Signals<sup>1</sup>

### 3.1 Introduction

Among all the different sources of entertainment, music and movies are probably the most important ones for entertaining people. Nowadays, thanks to the advances in the technology and with the rapid growth of the Internet, a large amount of music and movies has become available on-line. This has brought forward the need for organizing and managing these large databases. Among all music and movie descriptors, probably the most widely used criteria for indexing and retrieving musics/movies is the genre of the music/movie [104, 139, 4, 74, 16]. As a result of this, genre classification can be considered as an essential part of the music and movie recommender systems.

The most common approach regarding genre classification is the content-based approach. Thus far, various content-based genre classification methods have been proposed based on the variety of audio-visual features.

---

<sup>1</sup>This chapter is based on the two following publications: 1) Pouya Ghaemmaghani, Mojtaba Khomami Abadi, Seyed Mostafa Kia, Paolo Avesani, and Nicu Sebe. Movie Genre Classification by Exploiting MEG Brain Signals. in ICIAP, 2015 [28]. 2) Pouya Ghaemmaghani, and Nicu Sebe. Brain and Music: Music Genre Classification using Brain Signals. in EUSIPCO, 2016 [31].

Regarding movie genre classification, these features include average shot length, color variance, saturation, brightness, grayness, motion, and visual excitement [84, 117, 104, 12, 43, 139]. Regarding music genre classification, these features include MFCC, spectral centroid, spectral flux, zero crossings, energy, pitch, rhythm patterns, harmonic contents and etc [120, 73, 72, 70, 44]. However, regardless of all research conducted during the last years, content-based approaches always depend on the availability of multimedia contents. When such contents are not obtainable, these approaches are not applicable anymore. Besides, the main downside of the content-based approaches is that they are not emotion-centric and are not able to take into account the personal preferences of the people (i.e., they are not able to recommend the most suitable content according to a specific emotional status). Such preferences are important since there are sometimes disagreements between people on the definition of the genre due to the indistinct divisions between the different genres. In view of this, we propose an alternative approach for genre classification that aims at retrieving the people's perception. The rationale behind this is that the recommendation system that captures the people's understanding of the music/movie (e.g. via neurophysiological data), might discern the music genre better. In this study, we present preliminary experimental evidence for the possibility of the music/movie genre classification based on the brain recorded signals of individuals. The brain decoding paradigm is employed to classify recorded brain signals into the target genre classes. Figure 3.1 illustrates the overall framework used in our study. We compare the performance of our proposed paradigm on two neuroimaging datasets that contains the electroencephalographic (EEG) and the magnetoencephalographic (MEG) data of subjects who watched 36 movie clips and 40 music video clips. Our results suggests that the genre of the music/movie clips can be retrieved significantly over the chance-level using the brain signals.



Our study is a primary step towards user-centric music content retrieval by exploiting brain signals.

The rest of this chapter is organized as follows. In Section 3.2 we briefly review the literatures on genre classification in the multimedia content analysis context. Besides, we also review the relevant literatures on the brain decoding. Then, in Section 3.3 we explain the employed datasets, data preprocessing and feature extraction methods. Furthermore, we discuss the method used for annotating the music/movie genres. Section 3.4 elaborates our experimental results with a brief discussion. And finally, Section 3.5 concludes this chapter and highlights some future directions.

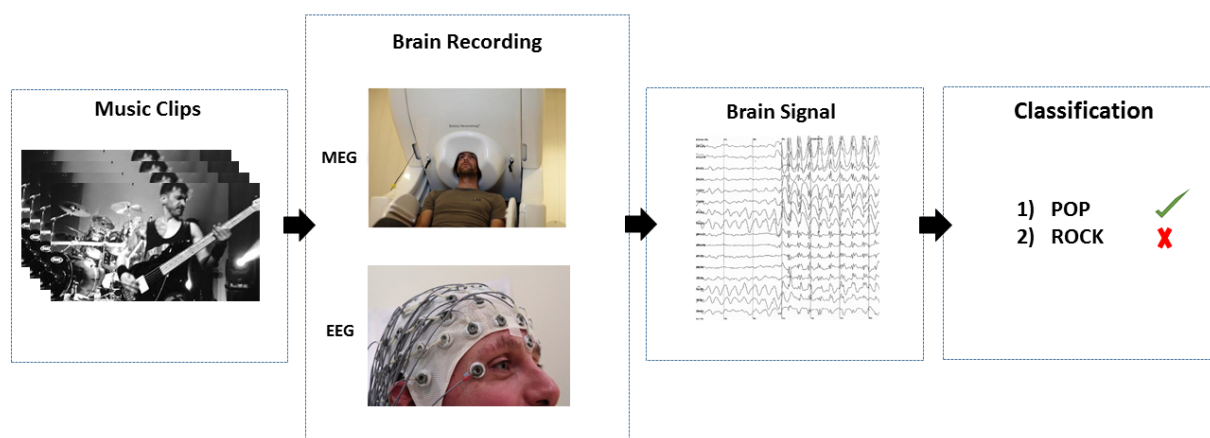


Figure 3.1: The framework used in this study regarding the movie/music genre classification by exploiting brain signals.

## 3.2 Related Works

### 3.2.1 Content-Centric Music/Movie Genre Classification

Regarding movie genre classification, in the literatures, various content-based genre classification approaches have been proposed based on audio-visual features [84, 117, 104, 12, 43, 139]. Rasheed, et al. [104] used

four low-level visual features (average shot length, color variance, motion content and lighting key) to classify over a hundred movie previews into four broad genre categories (Comedy, Action, Drama and Horror). In a similar study [43], the authors used the same low-level visual features and slow and fast moving effects to classify movie previews into three different genres. Elsewhere, Zhou, et al. [139] represented over one thousand movie trailers using a bag-of-visual-words model with shot classes as vocabularies. Then they mapped these bag-of-visual-words models to high-level movie genres.

Regarding music genre classification, there is a large body of works on content-based genre classification approaches. One of the earlier works is introduced by Tzanetakis and Cook [120] where the authors represent a music piece using timbral texture, rhythmic features, and pitch-related features. Their proposed features set has been widely used for music genre classification [73, 72, 70, 44]. Other characteristics such as contextual informational [78], temporal information [17], and semantic information [24] have been investigated in the literatures to improve the accuracy of genre classification. Recently, "sparse feature learning" methods have also been investigated for constructing a codebook for music songs [102, 135, 134, 98]. Elsewhere, Costa, et al. [16] proposed a robust music genre classification approach by converting the audio signal into a spectrogram and extracting features from this visual representation by treating the time-frequency representation as a texture image.

### 3.2.2 Genre and Affective Content

Apart from content-centric audio-visual features, movies and musics can be classified into different genres based on their emotional contents. This emotional content induces an emotional experience in the viewer [131]. In

fact, the emotions that are elicited in response to a clip contain useful information regarding the genre of the clip [110]. For example, in case of Horror and Action movies, it has been shown that movie segments with high emotion intensity cover the major part of the movie highlights [132].

The common approach for predicting multimedia affect is a content-centric approach, in which audio-visual features of the movie are used for affect prediction. Many researchers have investigated the affective contents of the video clips. Xu, et al. [131], analyzed the affective content of comedy and horror movies by detecting emotional segments. Soleymani, et al. [110] showed that a Bayesian classification approach can tag movie scenes into three affective classes (calm, positive excited and negative excited). They used content-based features extracted from each shot of 21 full length movies. In another study [132], a hierarchical model for analyzing movie affective contents was proposed. The proposed model, firstly, detects the emotional intensity level of the movie using fuzzy clustering on arousal features. Secondly, emotion types (Anger, Sad, Fear, Happy and Neutral) are detected using valence related features. Finally, Hidden Markov Models (HMMs) are applied to capture the context information. A similar hierarchical approach using conditional random fields (CRFs) was proposed in [133].

### 3.2.3 Affective Contents and Neurophysiological Signals

Recent works on affective computing, shows the possibility of decoding affects from neurophysiological data. This approach aims at capturing the emotion of the viewer. In [109], authors captured physiological responses of participants while they were watching movie scenes. They showed that the predicted affects from physiological responses of participants are significantly correlated with their self-assessed emotional responses. Koelstra et al. [61], Hadjidimitriou et al. [35], Abadi et al. [1] and Zheng et al.

[138] studied emotional responses of subjects induced by excerpts of music and video clips. These studies indicate that emotional information is encoded in brain signals. In [114], authors aimed at retrieving a music piece somebody listened to based on the EEG data. They obtained significant results (compared to the chance-level) when CNN classifiers are employed [113].

### 3.2.4 Spotting the gap

Our brief literature review reveals that music/movie genre classification has been achieved so far with content-based approaches. On the other hand, brain decoding algorithms were successfully employed on many tasks using various neuroimaging techniques. However, the efficacy of the brain decoding approaches on genre classification has not been explored. Therefore, this study aims at investigating the possibility of classifying movie/musical genres using brain data.

## 3.3 Experimental Setup

In this section, we describe the employed datasets, annotation process and feature extraction method.

### 3.3.1 Datasets

In our experiments, we used two publicly available datasets. These datasets contain the electroencephalographic (EEG) and the magnetoencephalographic (MEG) data of volunteers who watched 40 music video clips and 36 movie clips. The advantage of using these two datasets is that they contain the same music clips (the duration of each clip is 60 seconds) so that the

results can be compared. The details of these datasets are described below:

**MEG dataset:**

The MEG dataset, we employed in this study is the DECAF dataset [1]. This dataset contains the MEG brain signals of 30 volunteers while they were watching 40 music video clips and 36 movie clips. These clips were projected onto a screen placed in front of the subject inside the MEG acquisition room with 20 frames/second and at a screen refresh rate of 60 Hz. The magnetoencephalographic data were recorded in a magnetically shielded room with 1KHz sampling rate and in a controlled illumination using a Electa Neuromag device that outputs 306 channels (102 magnetometers and 204 gradiometers).

**EEG dataset:**

The EEG dataset, we employed in this study is the DEAP dataset [61]. This dataset contains the EEG brain signals of 32 participants while they were watching 40 music video clips. These music clips were projected onto a screen placed about a meter in front of the subject at a screen refresh rate of 60 Hz. The electroencephalographic data were recorded in controlled illumination, at a sampling rate of 512 Hz, using a Biosemi ActiveTwo system that outputs 32 channels.

### 3.3.2 Movie/Music Clips Annotation

The definition of a genre is very subjective so that one song/movie might belong to different genres according to different individuals. As a result

of such arbitrariness in the definition of the genre, many researchers have shown that even major taxonomies are inconsistent [96, 4, 3, 16]. To deal with such a difficulty and given the few number of total samples in our employed datasets (36 excerpts of movie clips and 40 excerpts of music clips), in this study, we asked human annotators to watch the music/movie video clips and assign each clip one specific label. The details of such annotation is as follows:

### **Annotating Movie Clips**

In order to annotate movie genres, three human observers were asked to classify each movie into four genres: Comedy, Romantic, Drama, Horror. The movie genres were picked based on the majority voting between the observers. To evaluate the consistency of the genres across subjects, we measured the agreement between annotators' labeling using the Cohen's Kappa measurement. The average  $\kappa$  across observers is  $77\% \pm 2\%$  ( $p - value < 0.001$ ) that suggests a *substantial agreement* [67] between the annotators. Furthermore, we employed the Cohen's kappa to evaluate the agreement between the movie genres obtained from the majority voting, with the genres obtained from the Internet Movie Database (IMDB). The average  $\kappa$  across the two labels is  $72\%$  ( $p - value < 0.001$ ) that shows a *substantial agreement* between our picked labels (from the majority voting) and the labels obtained from the IMDB. The lack of *full agreement* between these two labels is mainly due to the fact that the employed movie clips in [1] are not necessarily representing the whole movie theme. The genre labels provided by this study augment the dataset proposed in [1]. From here on we refer to the majority voting labels resulting from the annotation process as the ground-truth (see Table 3.2 for the obtained ground-truth labels).

### Annotating Music Clips

Same as movies, three human annotators (two of them are different from the ones who annotate the movie clips) were asked to classify each music clip into one of the two categories; The first category represents the following genres: Pop, Dance, Disco and Tech-no. We refer to this category as the *POP* category. And the second category represents the following genres: Rock and Metal. We refer to this category as the *ROCK* category. The music genre of each clip was picked based on the majority voting between the annotators. To evaluate the consistency of the annotation across subjects, we measured the Cohen’s Kappa agreement between annotators’ labeling. The obtained average  $\kappa$  across observers ( $69.8\% \pm 5\%$ ,  $p - value < 0.001$ ) indicates a *substantial* agreement [67] between the annotators. We refer to the majority voting labels as the ground-truth labels. Table 3.5 presents the name of the music clips together with their ground-truth labels.

### 3.3.3 Feature Extraction

**MEG Features:** The MEG trials are extracted and pre-processed using the MATLAB Fieldtrip toolbox [93] as follows:

1. Down-sampling the MEG signal to 300 Hz.
2. Bandpass frequency filtering (1 - 95 Hz) in order to remove the noise generated by external perturbations such as moving vehicles or muscle activity.
3. Estimating the spectral power of the 102 combined-gradiometer sensors of each trial with a window size of 300 samples. Following [1], (i) we discarded the magnetometer sensors because they are generally prone to noise and (ii) we used a standard Fieldtrip function

to combine the spectral power of planar gradiometers to obtain 102 combined-gradiometer spectral power for each trial.

4. Calculating MEG features by averaging the signal power over four frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz). The output of this procedure for each trial is a 3-dimensional matrix with the following dimensions: 102 (number of the MEG combined-gradiometer sensors)  $\times$  4 (major frequency bands)  $\times L$ , where  $L$  is the length of a video clip in seconds.

**EEG Features:** We used the same pre-processed EEG data as in [61]. These pre-processing steps are as follows:

1. Down-sampling the EEG signal to 128 Hz.
2. EOG artifacts removal.
3. Bandpass frequency filtering (1 - 45 Hz).
4. Estimating the spectral power of each channel of the EEG trials with a window size of 128 samples.
5. Calculating EEG features by averaging the signal power over four frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz). The output of this procedure for each trial is a 3-dimensional matrix with the following dimensions: 32 (number of the EEG sensors)  $\times$  4 (major frequency bands)  $\times$  60 (length of a music clip in seconds).

**MCA features:** For each second of the music video clips, low-level audio-visual features are extracted. These low-level Multimedia Content Analysis (MCA) features are listed in Table 3.1. The extracted multimedia content



analysis (MCA) features include 49 video features and 56 audio features. Hence, for each video, we have 105 (low-level multimedia features)  $\times L$  features.

Table 3.1: Extracted audio-visual features from each movie clip (the number of features is listed in the parenthesis).

<b>Audio features</b>	<b>Description</b>
<b>MFCC features (39)</b>	MFCC coefficients [71], derivative of MFCC, MFCC Auto-correlation (AMFCC)
<b>Energy (1) and Pitch (1)</b>	Average energy of audio signal [71] and first pitch frequency
<b>Formants (4)</b>	Formants up to 4400Hz
<b>Time frequency (8)</b>	mean and std of: MSpectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio [71]
<b>Zero crossing rate (1)</b>	Average zero crossing rate of audio signal [71]
<b>Silence ratio (2)</b>	Mean and std of proportion of silence in a time window [71]
<b>Video features</b>	<b>Description</b>
<b>Brightness (6)</b>	Mean of: Lighting key, shadow proportion, visual details, grayness, median of Lightness for frames, mean of median saturation for frames
<b>Color Features (41)</b>	Color variance, 20-bin histograms for hue and lightness in HSV space
<b>Motion (1)</b>	Mean inter-frame motion [1]
<b>VisualExcitement (1)</b>	Features as defined in [1]

## 3.4 Results and Discussion

### 3.4.1 Correlation Results

We calculate the Pearson correlation between the 102 combined MEG gradiometers in each frequency band ( $\theta$ ,  $\alpha$ ,  $\beta$ , and  $\gamma$ ) and audio-visual features extracted from movie clips. The obtained p-values were first fused over all clips and then over all subjects using the Fisher’s method [27]. We performed the Boferroni correction in order to correct our results for multiple comparisons. Figure 3.2 demonstrates the results of such correlation analysis. This figure shows two visual features (motion and grayness) and two audio features (the forth and the sixth MFCC coefficient). As one can

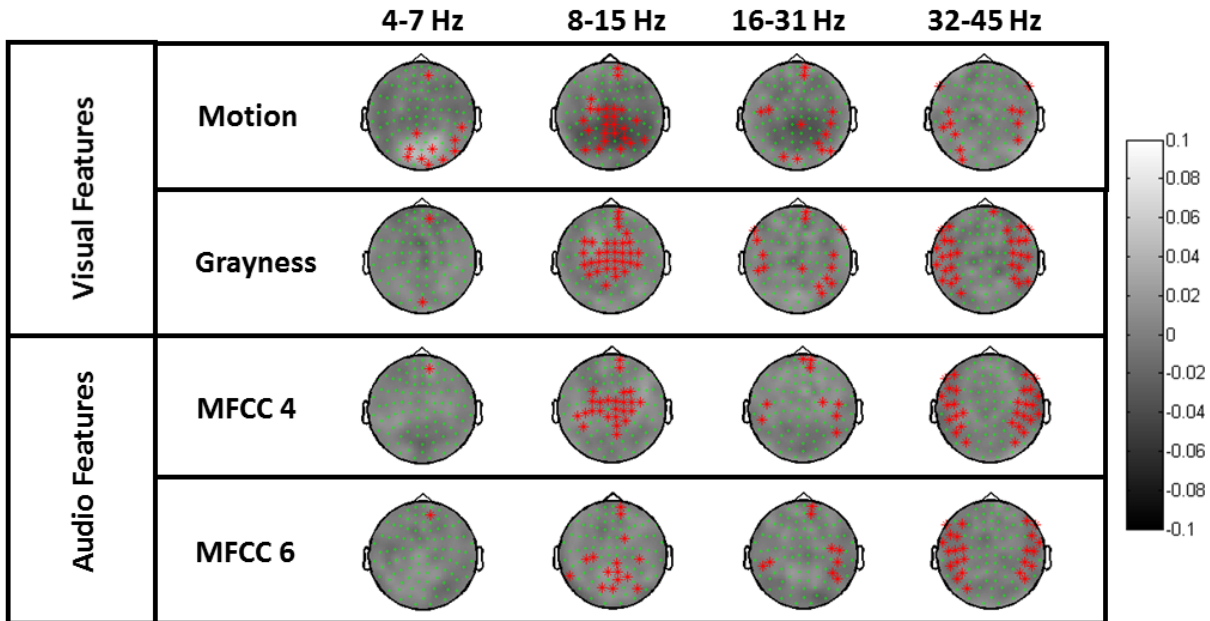


Figure 3.2: Pearson correlation analysis between the MEG responses and audio-visual features. Correlation over each channel is denoted by the gray level, and significant correlations are marked with red  $\star$ .

see, audio-visual features are significantly correlated with MEG sensors in temporal area of the brain in the  $\gamma$  band (32:45 Hz). This part of the brain processes the visual information as well as the audio information. Furthermore in the  $\alpha$  band (8-15 Hz), the extracted motion feature is significantly correlated with the MEG sensors located in the posterior part of the brain, confirming previous studies [54, 41].

### 3.4.2 Classification Results

We employed two classifiers (Naive Bayes classifier, and a Linear SVM classifier), in the classification experiments, under the leave-one-clip-out cross-validation schema to decode the brain/multimedia feature descriptors into our target genre classes.

### Feature Descriptors

The brain/multimedia feature descriptors, employed in the classification experiments, are calculated as follows:

**Movie Descriptors:** We used three types of features descriptors:

1. *MEG-based* descriptors by averaging the MEG features over time.
2. *MCA-based* descriptors by averaging the MCA features over time.
3. *MEG+MCA fusion* by concatenating the MCA descriptors and the MEG descriptors of each subject.

**Music Descriptors:** We fed the following features descriptors into the classifiers:

1. *MEG-based* descriptors by averaging the MEG features over time.
2. *EEG-based* descriptors by averaging the EEG features over time.
3. *MCA-based* descriptors by averaging the MCA features over time.
4. *MEG+MCA fusion* by concatenating the MCA descriptors and the MEG descriptors of each subject.
5. *EEG+MCA fusion* by concatenating the MCA descriptors and the EEG descriptors of each subject.

Note that the fusion of MEG and EEG descriptors is not feasible, since the subjects in these two datasets are not the same (DEAP contains 32 subjects whereas DECAF contains 30 subjects).

### Movie Genres Classification

In the classification experiments we employed Naive Bayes classifier and a Linear SVM classifier under the leave-one-clip-out cross-validation schema

to decode the brain/multimedia feature descriptors into our target movie genre classes (i.e. Comedy, Romantic, Drama, and Horror). The ground-truth labels are used as the target labels in the classification procedure (see Section 3.3.2).

**Subject-level analysis:** At the subject level, the classification procedure was employed on the brain data of each subject separately. Thus, the classification of the MEG-descriptors are repeated 30 times (corresponding to the number of subjects). For each subject, the 36 MEG descriptors (corresponding to the 36 movie clips) are used as samples. Figure 3.3 summarizes the results of the single-subject classification scenario. It compares the accuracy of four-class classification based on the MEG and MCA features with the chance level (27.4% and 28.2% obtained using Naive Bayes classifier and the SVM classifier respectively). The chance level is computed by feeding random numbers with normal distribution into the classification procedure for 100 times. In the MEG case, the average accuracies of 35.6% (using Naive Bayes classifier) and 37.2% (using the Linear SVM classifier) are obtained over 30 subjects which are significantly ( $p - value < 0.001$ ) higher than the chance level. This significant difference suggests the existence of the genre related information in the recorded brain activity. However, employing MCA features provides higher accuracy (45.5% in case of the Naive Bayes classifier and 61.1% in case of the SVM classifier) than employing the MEG features. This could be due to the fact that MEG signals are very noisy.

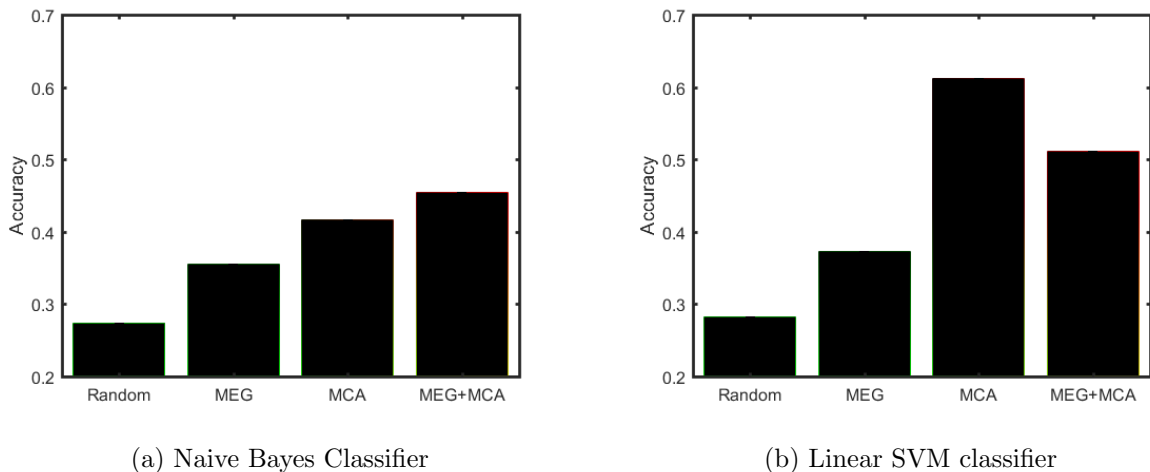


Figure 3.3: Comparison between the accuracy of MEG and multimedia features (MCA) with random inputs in the single-subject scenario. (a) using a Naive Bayes classifier. (b) using a Linear SVM classifier.

To investigate the effectiveness of the feature descriptors, for each movie genre, we computed the confusion matrices regarding the four-class genre classification using MCA and MEG features (using both classifiers). Figure 3.4 shows these confusion matrices. To facilitate the comparison, the confusion matrices are normalized with respect to the total number of samples ( $30 \times 36$  in the MEG case and 36 in the MCA case). Even though the classification accuracy using MCA features is higher than using MEG features, confusion matrices show significantly similar patterns ( $p - value < 2 \times 10^{-5}$ ). In both cases, the comedy and drama genres are predicted with higher confidence while romantic and horror genres are almost indistinguishable from other categories.

**Population-level analysis:** To evaluate the efficacy of MEG descriptors at the population level, for each video clip, we computed the majority vote over the predictions of the single-subject classification across all subjects.

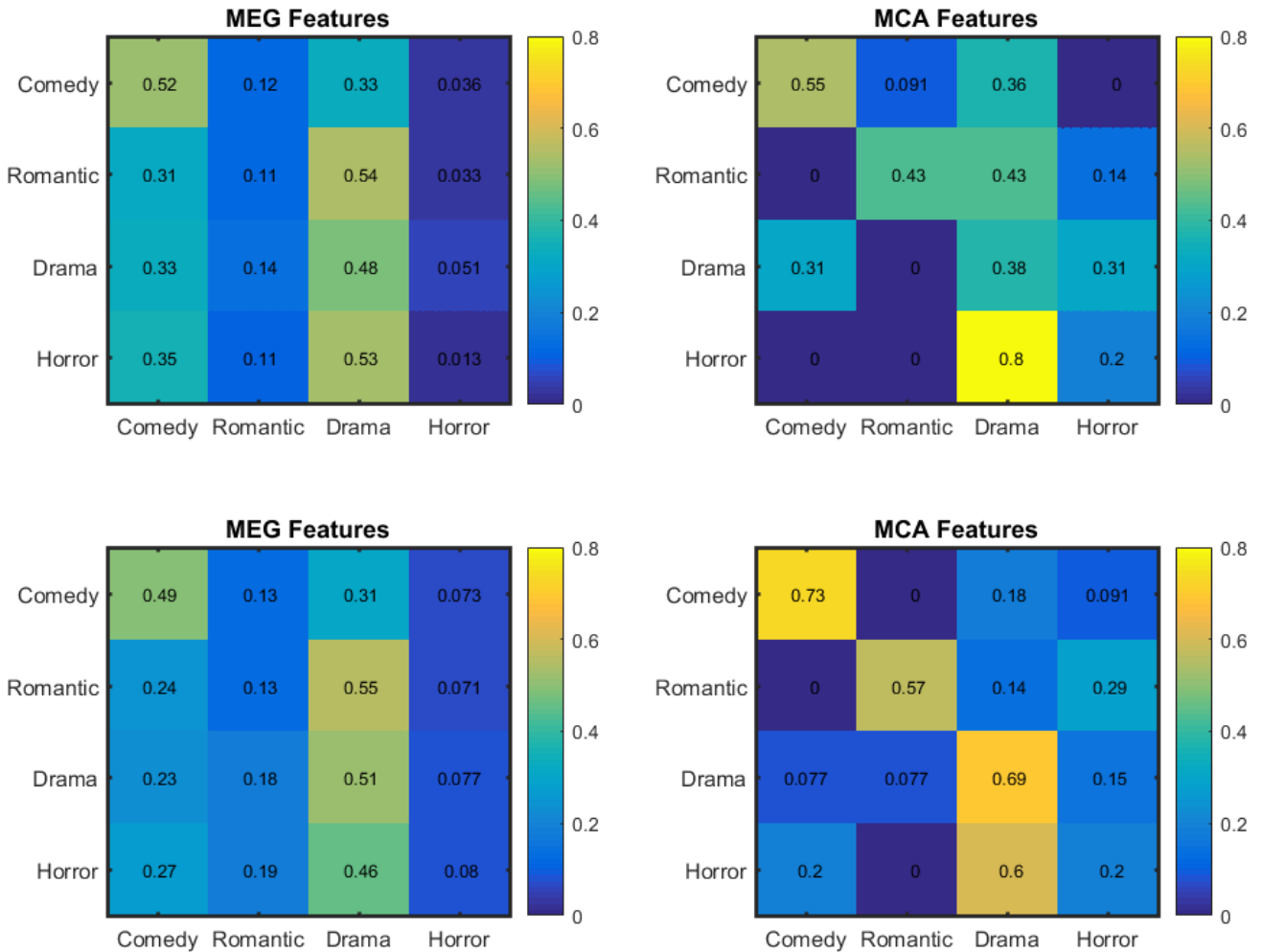


Figure 3.4: Confusion matrix for four-class genre classification using multimedia and MEG features. x and y axes represent predicted and actual labels, respectively. The top row represents the confusion matrices obtained using the Naive Bayes classifier. The bottom row represents the confusion matrices obtained by employing the Linear SVM classifier.

The results are summarized in Table 3.2 and Table 3.3. The population level accuracy using the SVM classifier by employing the MEG+MCA features is 75% which is significantly higher than the classification accuracy of only MCA features (61.1%) or MEG features (58.3%). However, in case of the Naive Bayes classifier, employing MEG+MCA features performs the same as just employing the MEG features. But despite the same perfor-

mance, examining the predicted genres in Table 3.2 and Table 3.3 shows that the combined features are more successful in predicting the romantic genre. This is mainly due to the fact that the MEG features are weaker than the MCA features in classifying the romantic genre. This may suggest the existence of the complementary genre related information in the brain signals and the multimedia contents.

### Music Genre Classification

We adopted a Linear SVM classifier and a Naive Bayes classifier under the leave-one-clip-out cross-validation schema to decode the brain/multimedia feature descriptors into our target music genre classes (i.e. Pop and Rock). The ground-truth labels are used as the target labels in the classification procedure (see Section 3.3.2). The feature descriptors are calculated as follows:

**Subject-level analysis:** At subject level, the classification procedure was employed on the brain data of each subject separately. Thus, the classification of the MEG-descriptors and the EEG descriptors are repeated 30 times and 32 times respectively (corresponding to the number of subjects in each dataset). For each subject, the 40 MEG/EEG descriptors (corresponding to the 40 music clips) are used as samples. Given the unbalanced number of samples for each genres, both accuracy and F-measure are reported as the metrics to compare the classification performance. These metrics are averaged over all subjects.

Table 3.4 compares the results of music genre classification using MEG, EEG and MCA descriptors using both classifiers. The chance level is computed by feeding random numbers with normal distribution into the classification procedure for 100 times. In both MEG and EEG case, the distri-

Table 3.2: Movie clip titles, ground-truth labels, and predicted labels. The accuracies are obtained by employing the Naive Bayes Classifier.

ID	Titles	Ground-Truth	MCA	MEG	MEG+MCA
1	Ace-Ventura: Pet Detective	COMEDY	COMEDY	COMEDY	COMEDY
2	The Gods Must be Crazy II	COMEDY	DRAMA	COMEDY	COMEDY
3	Liar Liar	COMEDY	ROMANTIC	COMEDY	COMEDY
4	Airplane	COMEDY	COMEDY	COMEDY	COMEDY
5	When Harry Met Sally	COMEDY	COMEDY	COMEDY	COMEDY
6	The Gods Must be Crazy	COMEDY	DRAMA	COMEDY	COMEDY
7	The Hangover	COMEDY	DRAMA	COMEDY	DRAMA
8	Up	COMEDY	COMEDY	DRAMA	COMEDY
9	Hot Shots	COMEDY	DRAMA	COMEDY	DRAMA
10	August Rush	ROMANTIC	DRAMA	DRAMA	DRAMA
11	Truman Show	ROMANTIC	DRAMA	DRAMA	DRAMA
12	Wall-E	ROMANTIC	HORROR	COMEDY	DRAMA
13	Love Actually	ROMANTIC	DRAMA	DRAMA	DRAMA
14	Remember the Titans	DRAMA	HORROR	DRAMA	DRAMA
15	Legally Blonde	COMEDY	COMEDY	DRAMA	DRAMA
16	Life is Beautiful	COMEDY	COMEDY	COMEDY	COMEDY
17	Slumdog Millionaire	ROMANTIC	ROMANTIC	DRAMA	ROMANTIC
18	House of Flying Daggers	ROMANTIC	ROMANTIC	DRAMA	ROMANTIC
19	Gandhi	DRAMA	DRAMA	DRAMA	DRAMA
20	My girl	DRAMA	COMEDY	DRAMA	COMEDY
21	Lagaan	DRAMA	COMEDY	DRAMA	COMEDY
22	Bambi	DRAMA	HORROR	DRAMA	DRAMA
23	My Bodyguard	DRAMA	DRAMA	DRAMA	DRAMA
24	Up	ROMANTIC	ROMANTIC	DRAMA	DRAMA
25	Life is Beautiful	DRAMA	DRAMA	DRAMA	DRAMA
26	Remember the Titans	DRAMA	COMEDY	DRAMA	DRAMA
27	Titanic	DRAMA	HORROR	DRAMA	DRAMA
28	Exorcist	HORROR	HORROR	DRAMA	DRAMA
29	Mulholland Drive	DRAMA	COMEDY	DRAMA	COMEDY
30	The Shining	HORROR	DRAMA	DRAMA	DRAMA
31	Prestige	DRAMA	HORROR	COMEDY	DRAMA
32	Alien	HORROR	DRAMA	DRAMA	DRAMA
33	The untouchables	DRAMA	DRAMA	COMEDY	DRAMA
34	Pink Flamingos	HORROR	DRAMA	COMEDY	DRAMA
35	Crash	DRAMA	DRAMA	DRAMA	DRAMA
36	Black Swan	HORROR	DRAMA	DRAMA	DRAMA
	Accuracy		41.7%	55.6%	55.6%



Table 3.3: Movie clip titles, ground-truth labels, and predicted labels. The accuracies are obtained by employing the Linear SVM Classifier.

ID	Titles	Ground-Truth	MCA	MEG	MEG+MCA
1	Ace-Ventura: Pet Detective	COMEDY	COMEDY	COMEDY	COMEDY
2	The Gods Must be Crazy II	COMEDY	HORROR	COMEDY	COMEDY
3	Liar Liar	COMEDY	COMEDY	COMEDY	COMEDY
4	Airplane	COMEDY	COMEDY	COMEDY	COMEDY
5	When Harry Met Sally	COMEDY	COMEDY	COMEDY	COMEDY
6	The Gods Must be Crazy	COMEDY	DRAMA	COMEDY	DRAMA
7	The Hangover	COMEDY	COMEDY	COMEDY	COMEDY
8	Up	COMEDY	COMEDY	DRAMA	COMEDY
9	Hot Shots	COMEDY	DRAMA	COMEDY	COMEDY
10	August Rush	ROMANTIC	HORROR	DRAMA	DRAMA
11	Truman Show	ROMANTIC	DRAMA	DRAMA	DRAMA
12	Wall-E	ROMANTIC	HORROR	DRAMA	ROMANTIC
13	Love Actually	ROMANTIC	ROMANTIC	DRAMA	ROMANTIC
14	Remember the Titans	DRAMA	DRAMA	DRAMA	DRAMA
15	Legally Blonde	COMEDY	COMEDY	COMEDY	COMEDY
16	Life is Beautiful	COMEDY	COMEDY	DRAMA	COMEDY
17	Slumdog Millionaire	ROMANTIC	ROMANTIC	DRAMA	ROMANTIC
18	House of Flying Daggers	ROMANTIC	ROMANTIC	DRAMA	ROMANTIC
19	Gandhi	DRAMA	DRAMA	DRAMA	DRAMA
20	My girl	DRAMA	COMEDY	DRAMA	COMEDY
21	Lagaan	DRAMA	DRAMA	DRAMA	DRAMA
22	Bambi	DRAMA	HORROR	DRAMA	DRAMA
23	My Bodyguard	DRAMA	DRAMA	DRAMA	DRAMA
24	Up	ROMANTIC	ROMANTIC	DRAMA	ROMANTIC
25	Life is Beautiful	DRAMA	DRAMA	ROMANTIC	DRAMA
26	Remember the Titans	DRAMA	DRAMA	DRAMA	DRAMA
27	Titanic	DRAMA	ROMANTIC	DRAMA	DRAMA
28	Exorcist	HORROR	DRAMA	DRAMA	DRAMA
29	Mulholland Drive	DRAMA	HORROR	DRAMA	DRAMA
30	The Shining	HORROR	DRAMA	DRAMA	DRAMA
31	Prestige	DRAMA	DRAMA	DRAMA	DRAMA
32	Alien	HORROR	HORROR	DRAMA	DRAMA
33	The untouchables	DRAMA	DRAMA	DRAMA	DRAMA
34	Pink Flamingos	HORROR	COMEDY	COMEDY	COMEDY
35	Crash	DRAMA	DRAMA	DRAMA	DRAMA
36	Black Swan	HORROR	DRAMA	DRAMA	DRAMA
	Accuracy		61.1%	58.3%	75%

Table 3.4: Comparison between the accuracy of MEG, EEG and MCA descriptors with random inputs in the single-subject level scenario.

Feature-Space	SVM		Naive Bayes	
	Accuracy	F-measure	Accuracy	F-measure
<b>Random</b>	$0.51 \pm 0.10$	$0.60 \pm 0.09$	$0.52 \pm 0.09$	$0.64 \pm 0.08$
<b>MCA</b>	0.70	0.73	0.82	0.87
<b>EEG</b>	$0.60 \pm 0.10$	$0.66 \pm 0.09$	$0.55 \pm 0.10$	$0.58 \pm 0.12$
<b>EEG+MCA</b>	$0.75 \pm 0.05$	$0.78 \pm 0.05$	$0.82 \pm 0.02$	$0.86 \pm 0.05$
<b>MEG</b>	$0.54 \pm 0.10$	$0.62 \pm 0.09$	$0.52 \pm 0.10$	$0.59 \pm 0.10$
<b>MEG+MCA</b>	$0.82 \pm 0.04$	$0.86 \pm 0.03$	$0.80 \pm 0.03$	$0.85 \pm 0.02$

bution of the obtained classification accuracies is better than chance level. This difference implies the existence of genre related information in the recorded brain activity. In the case of EEG descriptors, this difference is significant ( $p - value < 0.001$ ) where the SVM classifier is employed. Furthermore, combining brain features (EEG descriptors and MEG descriptors) of each subject with MCA descriptors provides higher accuracy than employing only EEG/MEG descriptors. Such brain-multimedia features fusion also outperforms the result of MCA descriptors (when the SVM classifier is used) suggesting the existence of complementary music genre related information in the brain signals.

**Population-level analysis:** To evaluate the efficacy of MEG/EEG descriptors at the population level, for each video clip, we computed the majority vote over predictions of the single-subject classification across all subjects. The results are summarized in Table 3.5 and Table 3.6. In case of the EEG descriptors, the population level accuracy (75% using the SVM classifier and 70% using the Navie BAYes Classifier) is higher than the single subject-level accuracy (60% using the SVM classifier and

### 3.4. RESULTS AND DISCUSSION

Table 3.5: Music clip titles, ground-truth labels, and predicted labels of different feature descriptors. The accuracies are obtained by employing the Linear SVM Classifier.

ID	Music Clip Title	Ground-Truth	MCA	MEG	MEG+MCA	EEG	EEG+MCA
1	Emiliana Torrini: Jungle Drum	POP	ROCK	POP	POP	POP	ROCK
2	Lustra: Scotty Doesn't Know	ROCK	POP	POP	POP	POP	POP
3	Jackson 5: Blame It On The Boogie	POP	ROCK	POP	POP	POP	ROCK
4	The B52'S: Love Shack	POP	POP	POP	POP	POP	POP
5	Blur: Song 2	ROCK	ROCK	POP	ROCK	POP	ROCK
6	Blink 182: First Date	ROCK	ROCK	POP	POP	POP	ROCK
7	Benny Benassi: Satisfaction	POP	ROCK	POP	POP	POP	POP
8	Lily Allen: Fuck You	POP	ROCK	POP	POP	POP	ROCK
9	Queen: I Want To Break Free	POP	POP	POP	POP	POP	POP
10	Rage Against The Machine: Bombtrack	ROCK	POP	POP	POP	POP	POP
11	Michael Franti : Say Hey (I Love You)	POP	POP	POP	POP	POP	POP
12	Grand Archives: Miniature Birds	POP	POP	ROCK	POP	POP	POP
13	Bright Eyes: First Day Of My Life	POP	POP	POP	POP	POP	POP
14	Jason Mraz: I'm Yours	POP	POP	POP	POP	POP	POP
15	Bishop Allen: Butterfly Nets	POP	POP	POP	POP	POP	POP
16	The Submarines: Darkest Things	POP	POP	POP	POP	POP	POP
17	Air: Moon Safari	POP	POP	POP	POP	POP	POP
18	Louis Armstrong: What A Wonderful World	POP	POP	POP	POP	POP	POP
19	Manu Chao: Me Gustas Tu	POP	POP	POP	POP	POP	POP
20	Taylor Swift: Love Story	POP	POP	POP	POP	POP	POP
21	Diamanda Galas: Gloomy Sunday	ROCK	POP	POP	POP	POP	POP
22	Porcupine Tree: Normal	ROCK	POP	POP	POP	POP	POP
23	Wilco: How To Fight Loneliness	POP	ROCK	POP	POP	POP	POP
24	James Blunt: Goodbye My Lover	POP	POP	POP	POP	POP	POP
25	A Fine Frenzy: Goodbye My Almost Lover	POP	POP	POP	POP	POP	POP
26	Kings Of Convenience: The Weight Of My Words	POP	ROCK	POP	POP	POP	POP
27	Madonna: Rain	POP	ROCK	POP	POP	POP	POP
28	Sia: Breathe Me	POP	POP	POP	POP	POP	POP
29	Christina Aguilera: Hurt	POP	POP	POP	POP	POP	POP
30	Enya: May It Be (Saving Private Ryan)	POP	ROCK	POP	POP	ROCK	ROCK
31	Mortemia: The One I Once Was	ROCK	ROCK	POP	ROCK	POP	ROCK
32	Marilyn Manson: The Beautiful People	ROCK	ROCK	POP	ROCK	ROCK	ROCK
33	Dead To Fall: Bastard Set Of Dreams	ROCK	ROCK	POP	ROCK	ROCK	ROCK
34	Dj Paul Elstak: A Hardcore State Of Mind	ROCK	ROCK	POP	ROCK	ROCK	ROCK
35	Napalm Death: Procrastination On The Empty Vessel	ROCK	ROCK	POP	ROCK	ROCK	ROCK
36	Sepultura: Refuse Resist	ROCK	ROCK	POP	ROCK	POP	ROCK
37	Cradle Of Filth: Scorched Earth Erotica	ROCK	ROCK	POP	ROCK	ROCK	ROCK
38	Gorgoroth: Carving A Giant	ROCK	ROCK	POP	ROCK	POP	ROCK
39	Dark Funeral: My Funeral	ROCK	ROCK	POP	ROCK	ROCK	ROCK
40	Arch Enemy: My Apocalypse	ROCK	ROCK	POP	ROCK	ROCK	ROCK
-	Accuracy	-	70%	57.5%	87.6%	75%	72.5%

55.5% using the Navie BAYes Classifier) and it is also higher than the classification accuracy of only MCA descriptors (70%). In case of MEG descriptors, the population-level analysis does not perform well. Nevertheless, in single-subject level analysis, as explained in previous section, the average obtained results are better than chance level.

CHAPTER 3. RETRIEVING GENRE RELATED INFORMATION FROM BRAIN

Table 3.6: Music clip titles, ground-truth labels, and predicted labels of different feature descriptors. The accuracies are obtained by employing the Naive Bayes Classifier.

ID	Music Clip Title	Ground-Truth	MCA	MEG	MEG+MCA	EEG	EEG+MCA
1	Emiliana Torrini: Jungle Drum	POP	POP	POP	POP	ROCK	POP
2	Lustra: Scotty Doesn't Know	ROCK	POP	POP	POP	POP	POP
3	Jackson 5: Blame It On The Boogie	POP	POP	POP	POP	POP	POP
4	The B52'S: Love Shack	POP	POP	POP	POP	ROCK	POP
5	Blur: Song 2	ROCK	POP	POP	POP	ROCK	POP
6	Blink 182: First Date	ROCK	POP	POP	POP	POP	POP
7	Benny Benassi: Satisfaction	POP	ROCK	POP	ROCK	POP	ROCK
8	Lily Allen: Fuck You	POP	POP	POP	POP	POP	POP
9	Queen: I Want To Break Free	POP	POP	POP	POP	ROCK	POP
10	Rage Against The Machine: Bombtrack	ROCK	POP	POP	POP	ROCK	POP
11	Michael Franti : Say Hey (I Love You)	POP	POP	POP	POP	POP	POP
12	Grand Archives: Miniature Birds	POP	POP	ROCK	POP	POP	POP
13	Bright Eyes: First Day Of My Life	POP	POP	POP	POP	POP	POP
14	Jason Mraz: I'm Yours	POP	POP	POP	POP	POP	POP
15	Bishop Allen: Butterfly Nets	POP	POP	POP	POP	POP	POP
16	The Submarines: Darkest Things	POP	POP	POP	POP	POP	POP
17	Air: Moon Safari	POP	POP	POP	POP	POP	POP
18	Louis Armstrong: What A Wonderful World	POP	POP	POP	POP	ROCK	POP
19	Manu Chao: Me Gustas Tu	POP	POP	ROCK	POP	POP	POP
20	Taylor Swift: Love Story	POP	POP	POP	POP	ROCK	POP
21	Diamanda Galas: Gloomy Sunday	ROCK	POP	POP	POP	POP	POP
22	Porcupine Tree: Normal	ROCK	POP	POP	POP	POP	POP
23	Wilco: How To Fight Loneliness	POP	POP	POP	POP	POP	POP
24	James Blunt: Goodbye My Lover	POP	POP	POP	POP	POP	POP
25	A Fine Frenzy: Goodbye My Almost Lover	POP	POP	POP	POP	POP	POP
26	Kings Of Convenience: The Weight Of My Words	POP	POP	POP	POP	POP	POP
27	Madonna: Rain	POP	POP	POP	POP	POP	POP
28	Sia: Breathe Me	POP	POP	POP	POP	POP	POP
29	Christina Aguilera: Hurt	POP	POP	POP	POP	ROCK	POP
30	Enya: May It Be (Saving Private Ryan)	POP	POP	POP	POP	ROCK	POP
31	Mortemia: The One I Once Was	ROCK	ROCK	POP	ROCK	ROCK	ROCK
32	Marilyn Manson: The Beautiful People	ROCK	ROCK	ROCK	ROCK	ROCK	ROCK
33	Dead To Fall: Bastard Set Of Dreams	ROCK	ROCK	POP	ROCK	ROCK	ROCK
34	Dj Paul Elstak: A Hardcore State Of Mind	ROCK	ROCK	POP	ROCK	ROCK	ROCK
35	Napalm Death: Procrastination On The Empty Vessel	ROCK	ROCK	POP	ROCK	ROCK	ROCK
36	Sepultura: Refuse Resist	ROCK	ROCK	POP	ROCK	ROCK	ROCK
37	Cradle Of Filth: Scorched Earth Erotica	ROCK	ROCK	POP	ROCK	ROCK	ROCK
38	Gorgoroth: Carving A Giant	ROCK	ROCK	POP	ROCK	ROCK	ROCK
39	Dark Funeral: My Funeral	ROCK	ROCK	POP	ROCK	ROCK	ROCK
40	Arch Enemy: My Apocalypse	ROCK	ROCK	POP	ROCK	ROCK	ROCK
-	Accuracy	-	82.5%	57.5%	82.5%	70%	82.5%

## 3.5 Conclusion

In this Chapter, we presented an approach for classification of music/movie clips into the target genre classes using MEG/EEG brain signals. We experimentally demonstrated the existence of a significant correlation between low-level audio-visual features and the brain signals. This finding shows the possibility of the prediction and the reconstruction of the multimedia features using brain signals.

Furthermore, a classifier has been used to perform the genre class prediction using the features extracted from brain signals. Regardless of the fact that we need to cope with few and noisy samples, our classification results confirm the possibility of user-centric music/movie genre classification using only the brain features. In addition, our analysis suggests the existence of complementary genre related information in the features extracted from brain signals and the multimedia content. To the best of our knowledge, this study is one of the first efforts in the direction of creating user-centric music/movie recommender system using brain signals. As a future plan, this study can be extended in the following directions:

1. Employing more effective (sophisticated) machine learning algorithms in order to improve the classification results.
2. Replicating the same pipeline on other neuroimaging datasets (i.e. using portable brain recording devices such as Emotiv sensors).
3. Reconstructing the external stimuli by exploiting the features extracted from the brain signals.

Our future research will focus on the first and the third issues. Generally, the neuroimaging datasets suffer from having few and noisy samples. This leads to a drop in the performance of machine learning algorithms. In the

next chapter, we will tackle this problem by employing a transfer learning paradigm in order to take into account the rich information existing in other domains. The aim of such approach is to transfer knowledge from a rich modality to the poor-performing modality.

Finally, in Chapter 5, we aimed at the third issue which is reconstruction of the stimuli using the brain activity.

# Chapter 4

## Domain Adaptation<sup>1</sup>

### 4.1 Introduction

The neuroimaging datasets suffer from few samples due to the cost of recording brain signals and subject's fatigue. Additionally, the recordings are very noisy due to the low signal-to-noise ratio and the non-stationarity nature of the signals. These two constraints lead to a sudden drop in the performance of machine learning algorithms. For example, in the previous chapter, we discussed the possibility of music and movie genre classification using brain signals. However the obtained results using low-level multimedia features, instead of brain features, were far better.

In machine learning literature, researchers tackle this problem by employing the transfer learning paradigm. In this paradigm, shared knowledge can be transferred from a large set of samples of source domain to a target domain with fewer samples. In such cases, the performance in the target domain strictly relies on the performance in the source domain and the similarity between the two domains. These methods aim at finding representations such that the domain divergence and consequently the modeling

---

<sup>1</sup>This chapter is based on the two following publications: 1) Pouya Ghaemmaghami, Moin Nabi, Yan Yan, and Nicu Sebe. Sparse-coded Cross-domain Adaptation from the Visual to the Brain Domain. in ICPR, 2016 [30]. 2) Pouya Ghaemmaghami, Moin Nabi, Yan Yan, and Nicu Sebe. A Cross-modal Adaptation Approach for Brain Decoding. in ICASSP, 2017 [29].

error on the target domain would be minimized. Transfer learning can truly be beneficial in cases where collecting data is extremely expensive or even impossible [97, 76, 9, 50, 95, 103]. This situation arises often in brain studies.

Motivated by recent successes in domain adaptation in machine learning literature [111, 100, 121, 34], in this study we investigate the possibility of transferring knowledge from the multimedia domain to the brain domain. We experimentally show that such adaptation procedure leads to improved results for two different tasks (object recognition and genre classification) in the brain domain, outperforming the results of brain features significantly. This is the first study in the direction of transferring knowledge by adapting representations learned on the multimedia domain to the brain modality.

The rest of this chapter is structured as follows: Section 4.2 reviews related literatures on this topic. Section 4.3 and Section 4.4 investigate a domain adaptation pipeline on two different tasks (genre classification and object recognition) using different neuroimaging modalities. And Section 4.5 concludes this chapter by summarizing the key observations and some possible future directions.

## 4.2 Related Works

### 4.2.1 Cross-Modal Domain Adaptation

Convolutional Neural Networks have recently resurfaced as a powerful tool for learning from big data (*e.g.*, ImageNet [106] with  $\sim 1$ M images), providing models with excellent representational capacities. These models have been trained via backpropagation through several layers of convolutional filters [69, 64]. It has been shown that such models are not only able to achieve state-of-the-art performance for the same visual recognition tasks, but the learned representation can be readily applied to other relevant



tasks [22]. These models perform extremely well in domains with large amounts of training data. With limited training data, however, they will generally dramatically over-fit the training data. Attracted by their amazing capability to produce a generic semantic representation, in this paper, we investigate transferring the learned representation from a large set of samples of visual domain to a small set of samples from the brain domain.

There has been a large body of works on representation transfer across domains belonging to the same modality. The representation-transfer aims at encoding the knowledge used to transfer across domains into a learned representation by minimizing the domain discrepancy and the classification error. In [107] an adaptation technique has been proposed that projects the features into a domain-invariant space via a transformation learned from both domains. Others have proposed domain adaptation methods based on a learning asymmetric non-linear transformation [65], subspace alignment [26] and Geodesic flow kernel [32]. This problem is also investigated as "common feature learning" in the field of multi-task learning [2]. More recently, authors in [121] proposed a deep architecture which simultaneously optimizes for domain divergence and uses a soft label distribution matching loss. All these lines of work focused on the problem of domain adaptation within the same modality. In this work we, however, tackle the more difficult problem of domain adaptation across different modalities. This cross-model adaptation problem has received much less attention. While a few methods have been proposed for the text/image [111, 87] and depth/image [34] adaptation, as far as we know, we are the first showing that such cross-model adaptation can be used for brain signals.

### 4.2.2 Domain Adaptation in Brain Studies

Brain signals have inherent variability because of low signal-to-noise ratio, non-stationarity and also the different physical and mental conditions of

the subject. Such variability complicates the analysis of data in a consistent way and degrades the performance of the brain decoding algorithm. This has limited the scope of many brain studies from finding a global decoder that can be applied to all subjects to subject-specific decoders [83, 48, 62].

To accommodate the variability in brain signals, recently transfer learning approaches have been applied on various brain datasets. The aim of such transfer learning algorithms is to find discriminative features that are common across subjects aiming at reducing subject variability and enabling information sharing across subjects that consequently leads to increasing in performance of the decoding system.

We target a completely different perspective compared to these works. Investigating on a common feature space across subjects is out of scope of this study. We differently aim at exploring cross-modal domain adaptation in which each individual’s brain data takes advantage of the semantic representations obtained in another modality.

### 4.3 Study 1: Object Recognition

Recent progress in Deep Neural Nets (DNN) provides the transfer learning community the opportunity to learn generic representations which are capable of capturing the semantics, hence they can be transferred across domains [22, 121] and modalities [34]. Due to the transferability power of such representations specifically in an object recognition task, in this study we investigate the possibility of transferring them for the same object recognition task using brain signals. Prior works in brain studies have shown that there is a region in the human brain called the “Ventral Temporal Cortex” (VTC) containing information about colour, object categories, concepts and semantics [33, 38, 94]. Inspired by this, and because of the importance of VTC in visual perception and object recognition, in this

study, we address the specific problem of transferring knowledge learned in ImageNet [106] to the brain domain.

### 4.3.1 Experimental Setup

#### Dataset

In this work, we used a well-known dataset introduced in a study on face and object representation in human ventral temporal cortex [38]. This dataset is the only publicly available fMRI dataset for object recognition and has been studied massively by many researchers [18, 15, 41, 91]. It consists of the fMRI data of 6 subjects in which each subject had undergone 12 sessions (runs). In each run, the subjects passively viewed greyscale images of eight object categories (faces, houses, cats, bottles, scissors, shoes, chairs, and nonsense patterns)<sup>2</sup>, grouped in 24s time blocks separated by rest periods. Each image was shown for 500ms and was followed by a 1500ms inter-stimulus interval. Full-brain fMRI data were recorded with a volume repetition time of 2.5s, thus, a stimulus block was covered by roughly 9 volumes.

#### Feature Selection

In fMRI studies, voxels (3D dimensional pixel that refers to small part of the brain) are the typically considered features for the decoding algorithms. However, among the whole set of voxels in the brain, not all of them are needed for object recognition task [47, 38, 37, 21]. As such, a subset of brain voxels dealing with visual categorization in human brain needs to be selected. Ventral Temporal Cortex (VTC) is the area in the brain where high-level visual regions reside. VTC is involved in visual perception and

---

<sup>2</sup>We discard “nonsense patterns” category in all our experiments.

recognition. Such recognition is achieved by organizing representations in a nested spatial hierarchy that serves as a neural infrastructure which enables flexible access to category information at several levels of abstraction [33].

In this work, we propose using only voxels within this area (VTC) in order to select the relevant features for object recognition task. However, selecting VTC voxels is not trivial [38]. Traditionally, this has been done using a *univariate approach*, but in this work we propose an *atlas-based approach* for voxel selection. Below, we elaborate and compare these two methods.

**VTC voxels - A univariate Approach:** In [38] VTC voxels are obtained for each subject separately using an univariate analysis by thresholding the brain volumes that are sensitive to one specific task (e.g. face). As a result of this analysis, voxels that are sensitive to specific tasks are selected. However, this procedure has two major drawbacks: 1) these voxels are too subject dependent. Thus, the shape of the VTC is different from subject to subject and consequently the area they cover in the brain is different as well. Figure 4.2b depicts this issue. 2) Since the masks are obtained prior to the analysis, the features are extracted without splitting the data into disjoint training and testing sets. This yields a double dipping effect [63] in brain decoding that leads to over-fitting on each individual’s data. As a result of this, if we use VTC voxels obtained in one subject to decode another individual, the results drastically decrease in most cases.

**VTC voxels - Atlas-Based Approach:** An alternative method for selecting voxels in the VTC area is an atlas-based approach. This approach has been used previously in neuroimaging studies in order to align individual’s brain anatomy to a standard template. However, it has not been used as a paradigm for selecting voxels. In this approach, instead of selecting

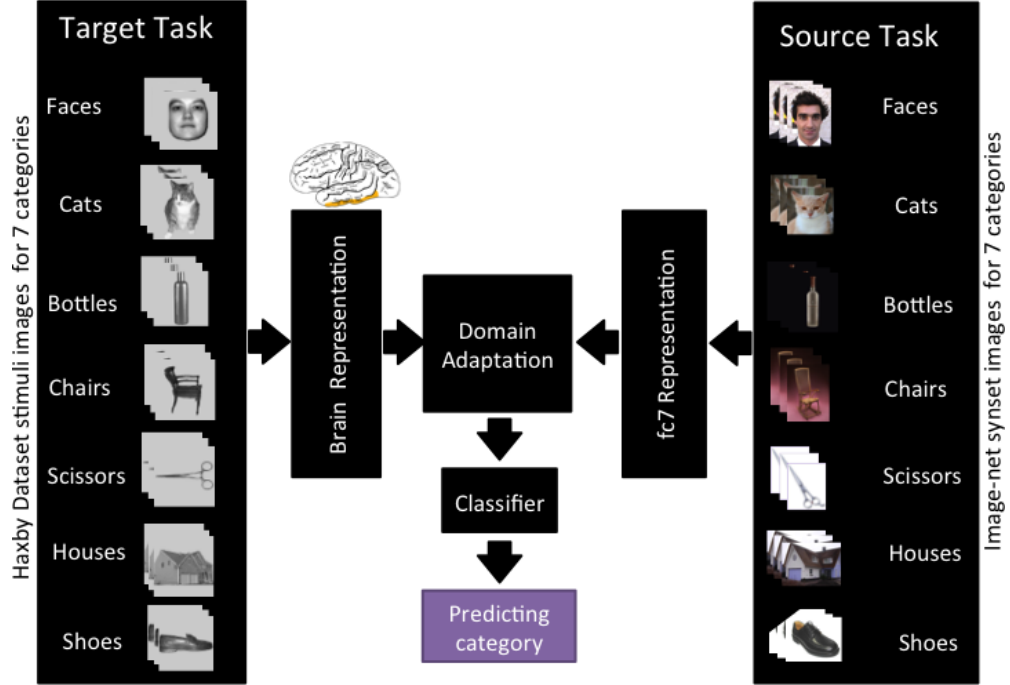


Figure 4.1: Domain Adaptation Pipeline.

VTC voxels in each individual, we define our Region of Interest (ROI) using brain atlases. In this study, we used the “Harvard-Oxford cortical and subcortical structural atlases” that provide probabilistic atlases covering 48 cortical and 21 subcortical structural brain areas [20]. These atlases are available in FSL which is a comprehensive library of analysis tools for fMRI brain imaging data [130, 49]. These brain regions are obtained in the MNI (Montreal Neurological Institute) space. MNI is a 3-dimensional coordinate system of the human brain obtained from 241 MRI brain scans which is used widely as a standard space in order to map the location of brain areas independent from individual differences. Our selected VTC area in MNI space includes the following brain regions: 1) temporo-occipital part of Inferior Temporal Gyrus, 2) posterior part of Parahippocampal Gyrus and 3) temporo-occipital part of Fusiform Gyrus. Since this VTC area is acquired in a standard space (i.e., the MNI space) using the brain corti-

cal atlas, it can be used for every subject. Figure 4.2c shows the shape of this area in the MNI space. Once the VTC area in the MNI space is obtained, we map these regions to each subject’s brain anatomy to match the subject’s anatomical and functional data. This can be done as follows: 1) First, we align the subject’s fMRI data to the MNI space. Figure 4.2a shows such mapping. 2) The inverse transform of the mapping function in the first step, can bring us to the subject’s anatomy from the MNI space. We used this inverse function in order to acquire atlas-based VTC voxels obtained in the MNI space from each individual’s fMRI data. We refer to this voxel selection approach as “Atlas-based VTC”. In Section 4.3.3 we compare this approach with the univariate approach and discuss the effects of voxel selection on the classification results.

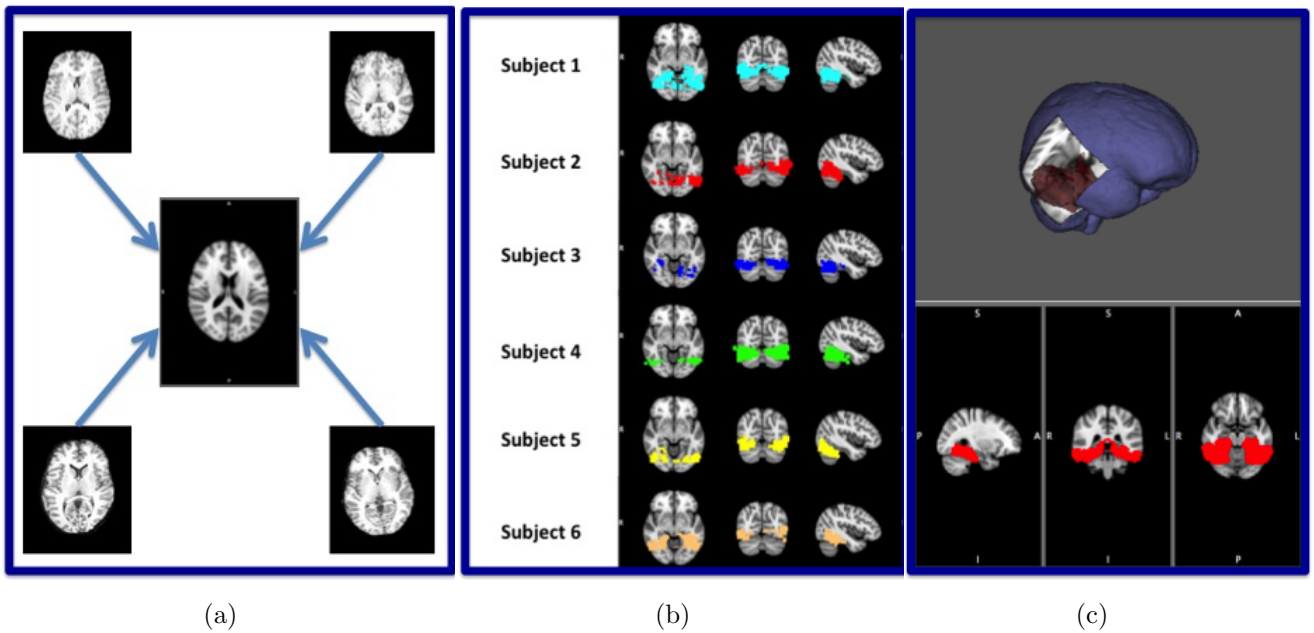


Figure 4.2: (a) The transformation of each subject’s brain anatomy to the standard space (i.e., the MNI space). (b) The shape of Haxby’s VTC area obtained for each subject individually in the MNI space. As it is expected, the shape of VTC in this univariate approach is different for each subject. (c) The shape of the VTC area in the MNI space which is acquired by a brain atlas.

### 4.3.2 Adaptation Method

Sparse coding was shown to be able to find succinct representations of stimuli from the brain [92]. In this section, we describe the details of our domain adaptation sparse coding method. Figure 4.1 demonstrates the overview of our adaptation paradigm.

The source task (i.e., the image domain) consists of data samples denoted by  $\mathbf{X}_s = \{\mathbf{x}_s^1, \mathbf{x}_s^2, \dots, \mathbf{x}_s^{n_s}\} \in \mathbb{R}^{n_s \times d}$ , where  $\mathbf{x}_s^i \in \mathbb{R}^d$  is a  $d$ -dimensional feature vector and  $n_s$  is the number of samples in the source task. The target task is defined as the brain fMRI data. Similarly, the target task consists of data samples denoted by  $\mathbf{X}_t = \{\mathbf{x}_t^1, \mathbf{x}_t^2, \dots, \mathbf{x}_t^{n_t}\} \in \mathbb{R}^{n_t \times d}$ , where  $\mathbf{x}_t^i \in \mathbb{R}^d$  is a  $d$ -dimensional feature vector and  $n_t$  is the number of samples in the target task.

To better adapt useful knowledge from the source domain to the target domain, we are going to learn a shared subspace across the two domains, obtained by an orthonormal projection  $\mathbf{W} \in \mathbb{R}^{d \times b}$ , where  $b$  is the dimensionality of the subspace. In this learned subspace, the data distributions between the source domain and the target domain should be similar to each other. The benefits of this strategy is that we can improve the coding quality of the target task by transferring knowledge from the source task. This can be realized through the following optimization problem:

$$\begin{aligned}
 \min_{\mathbf{C}_s, \mathbf{D}_s, \mathbf{C}_t, \mathbf{D}_t, \mathbf{W}, \mathbf{D}} & \|\mathbf{X}_s - \mathbf{C}_s \mathbf{D}_s\|_F^2 + \lambda_1 \|\mathbf{C}_s\|_1 \\
 & + \|\mathbf{X}_t - \mathbf{C}_t \mathbf{D}_t\|_F^2 + \lambda_2 \|\mathbf{C}_t\|_1 \\
 & + \lambda_3 \|\mathbf{X}_s \mathbf{W} - \mathbf{C}_s \mathbf{D}\|_F^2 + \lambda_4 \|\mathbf{X}_t \mathbf{W} - \mathbf{C}_t \mathbf{D}\|_F^2 \\
 \text{s.t.} & \begin{cases} \mathbf{W}^T \mathbf{W} = \mathbf{I} \\ (\mathbf{D}_s)_j \cdot (\mathbf{D}_s)'_j \leq 1, \quad \forall j = 1, \dots, l \\ (\mathbf{D}_t)_j \cdot (\mathbf{D}_t)'_j \leq 1, \quad \forall j = 1, \dots, l \\ \mathbf{D}_j \cdot \mathbf{D}'_j \leq 1, \quad \forall j = 1, \dots, l \end{cases}
 \end{aligned} \tag{4.1}$$

where  $\mathbf{D}_s, \mathbf{D}_t \in \mathbb{R}^{l \times d}$  are overcomplete dictionaries ( $l > d$ ) with  $l$  prototypes of the source and target task;  $(\mathbf{D}_s)_j$  and  $(\mathbf{D}_t)_j$  in the constraints

denotes the  $j$ -th row of  $\mathbf{D}_s$  and  $\mathbf{D}_t$  respectively;  $\mathbf{C}_s \in \mathbb{R}^{n_s \times l}$  and  $\mathbf{C}_t \in \mathbb{R}^{n_t \times l}$  corresponds to the sparse representation coefficients of  $\mathbf{X}_s$  and  $\mathbf{X}_t$  respectively. In the last two terms of Eqn.(4.1),  $\mathbf{X}_s$  and  $\mathbf{X}_t$  are projected by  $\mathbf{W}$  into the subspace to explore the relationship between the source and the target tasks.  $\mathbf{D} \in \mathbb{R}^{l \times b}$  is the dictionary learned in the shared subspace between the source and the target tasks.  $\mathbf{D}_j$  in the constraints denotes the  $j$ -th row of  $\mathbf{D}$ .  $\mathbf{I}$  is the identity matrix.  $(\cdot)'$  denotes the transpose operator.  $\lambda$ 's are the regularization parameters. The first constraint guarantees the learned  $\mathbf{W}$  to be orthonormal, and the other constraints prevent the learned dictionary to be arbitrarily large. In our objective function, we learn dictionaries  $\mathbf{D}_s$ ,  $\mathbf{D}_t$  for the source and the target task respectively and one shared dictionary  $\mathbf{D}$  between the source and the target tasks.

**Optimization:** To solve the proposed objective problem of Eqn.(4.1), we adopt the alternating minimization algorithm to optimize it with respect to  $\mathbf{D}$ ,  $\mathbf{D}_s$ ,  $\mathbf{C}_s$ ,  $\mathbf{D}_t$ ,  $\mathbf{C}_t$  and  $\mathbf{W}$  respectively in five steps as follows:

**Step1: Fixing  $\mathbf{D}_s$ ,  $\mathbf{C}_s$ ,  $\mathbf{W}$ ,  $\mathbf{D}_t$ ,  $\mathbf{C}_t$ , Optimize  $\mathbf{D}$ .** If we stack  $\mathbf{X} = [\mathbf{X}_s; \mathbf{X}_t]$ ,  $\mathbf{C} = [\mathbf{C}_s; \mathbf{C}_t]$ , Eqn.(4.1) is equivalent to:

$$\begin{aligned} & \min_{\mathbf{D}} \|\mathbf{XW} - \mathbf{CD}\|_F^2 \\ & s.t. \quad \mathbf{D}_j \mathbf{D}_j^T \leq 1, \quad \forall j = 1, \dots, l \end{aligned}$$

This is equivalent to the dictionary update stage in the traditional dictionary learning algorithm. We adopt the dictionary update strategy of Algorithm 2 in [76] to efficiently solve it.

**Step2: Fixing  $\mathbf{D}$ ,  $\mathbf{C}_s/\mathbf{C}_t$ ,  $\mathbf{W}$ , Optimize  $\mathbf{D}_s/\mathbf{D}_t$ .** This is the same as Step 1 which is equivalent to the dictionary update stage in the traditional dictionary learning for  $k$  tasks. We adopt the dictionary update strategy of Algorithm 2 in [76] to efficiently solve it.



**Step3: Fixing  $\mathbf{D}_s/\mathbf{D}_t$ ,  $\mathbf{W}$ ,  $\mathbf{D}$ , Optimize  $\mathbf{C}_s/\mathbf{C}_t$ .** Eqn.(4.1) is equivalent to:

$$\begin{aligned} \min_{\mathbf{C}_s, \mathbf{C}_t} & \|\mathbf{X}_s - \mathbf{C}_s \mathbf{D}_s\|_F^2 + \lambda_1 \|\mathbf{C}_s\|_1 \\ & + \|\mathbf{X}_t - \mathbf{C}_t \mathbf{D}_t\|_F^2 + \lambda_2 \|\mathbf{C}_t\|_1 \\ & + \lambda_3 \|\mathbf{X}_s \mathbf{W} - \mathbf{C}_s \mathbf{D}\|_F^2 + \lambda_4 \|\mathbf{X}_t \mathbf{W} - \mathbf{C}_t \mathbf{D}\|_F^2 \end{aligned}$$

This formulation can be decoupled into  $(n_s + n_t)$  distinct problems. We adopt the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [7] to solve the problem.

**Step4: Fixing  $\mathbf{D}_s$ ,  $\mathbf{C}_s$ ,  $\mathbf{D}$ ,  $\mathbf{D}_t$ ,  $\mathbf{C}_t$ , Optimize  $\mathbf{W}$ .** If we stack  $\mathbf{X} = [\mathbf{X}_s; \mathbf{X}_t]$ ,  $\mathbf{C} = [\mathbf{C}_s; \mathbf{C}_t]$ , Eqn.(4.1) is equivalent to:

$$\begin{aligned} \min_{\mathbf{W}} & \|\mathbf{XW} - \mathbf{CD}\|_F^2 \\ \text{s.t.} & \mathbf{W}^T \mathbf{W} = \mathbf{I} \end{aligned}$$

Substituting  $\mathbf{D} = (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{XW}$  back into the above function, we achieve

$$\begin{aligned} \min_{\mathbf{W}} & \|(\mathbf{I} - \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T) \mathbf{XW}\|_F^2 \\ & = \min_{\mathbf{W}} \text{tr}(\mathbf{W}^T \mathbf{X}^T (\mathbf{I} - \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T) \mathbf{XW}) \\ \text{s.t.} & \mathbf{W}^T \mathbf{W} = \mathbf{I} \end{aligned}$$

The optimal  $\mathbf{W}$  is composed of eigenvectors of the matrix  $\mathbf{X}^T (\mathbf{I} - \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T) \mathbf{X}$  corresponding to the  $s$  smallest eigenvalues.

We summarize our algorithm for solving Eqn.(4.1) as Algorithm 1.

Finally, the classification algorithm can be applied to  $\mathbf{C}_t$  with corresponding labels to train classification models to be used in the target domain. We refer to  $\mathbf{C}_t$  as ‘‘Adapted-Features’’.

### 4.3.3 Experiments and Results

In this section, we first introduce the details of the source and the target domains, then explain the classification scenario and finally elaborate the experiments in detail and discuss the results.

**Require:**

Data sample matrix  $\mathbf{X}$ ; Subspace dimensionality  $b$ , Dictionary size  $l$ , Regularization parameters  $\lambda_s$ .

**Ensure:**

Optimized  $\mathbf{W} \in \mathbb{R}^{d \times b}$ ,  $\mathbf{C} \in \mathbb{R}^{n \times l}$ ,  $\mathbf{D}_s \in \mathbb{R}^{l \times d}$ ,  $\mathbf{D}_t \in \mathbb{R}^{l \times d}$ ,  $\mathbf{D} \in \mathbb{R}^{l \times b}$ .

1: Initialize  $\mathbf{W}$  using any orthonormal matrix;

2: Initialize  $\mathbf{C}$  with  $l_2$  normalized columns;

3: **repeat**

    Compute  $\mathbf{D}$ ,  $\mathbf{D}_s$ ,  $\mathbf{D}_t$  using Algorithm 2 in [76];

    Adopting FISTA [7] to solve  $\mathbf{C}$ ;

    Compute  $\mathbf{W}$  by eigen decomposition of  $\mathbf{X}^T(\mathbf{I} - \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}')\mathbf{X}$ ;

**until** *Convergence*;

**Algorithm 1:** Domain adaptation method.

**Datasets:**

**Source Domain - ImageNet:** We use ImageNet images [106] selected from the synsets corresponding to the seven object categories of: faces, houses, cats, bottles, chairs, shoes and scissors.<sup>3</sup> The number of images for each category of interest is more than 1000 images. For each sample, we extract the output of *fc7* layer of pre-trained AlexNet model [64] using the standard CNN Caffe toolbox [51].

**Target Domain - Brain data:** Our target domain is Haxby's fMRI dataset. For each subject, features are extracted using the Atlas-based approach explained in previous sections. The number of samples corresponds to the number of volumes, and the number of features corresponds to the number of Atlas-based VTC voxels.

---

<sup>3</sup>These categories are the same categories used in the Haxby's fMRI dataset (excluding the non-sense pattern images).

### Classification Scenario

As a common practice in brain decoding, we adopt a classification procedure to classify extracted brain features into object categories. We employed a linear support vector machine to classify brain features. The within-subject analysis is performed as following: The classifier is trained and tested on the data for subject “s”. This is repeated for all six subjects. However, this does not constitute training on the test data since different brain scans regarding the visual stimuli are used for training and testing. For this, leave-one-out cross validation is performed by run: One run is taken as the test set and the remaining runs are taken as the training set. In other words, when testing on run “r”, the classifier is trained on all runs except run “r”. Such cross validation prevents training on the test data and ensures that there is no information leakage from the training set to the test set. This is repeated for all twelve runs, thus performing twelve-fold cross validation.<sup>4</sup>

### Experiments

We first study the cross-subject generalization of two feature selection methods explained in Section 4.3.1. We, then, evaluate the effectiveness of the adaptation method. In all experiments, we employed the same classification scenario mentioned above.

**Experiment 1:** We firstly investigate the performance of the *fc7* features on classifying the synset images (see Section “Datasets”). For this, 10-fold cross validation schema is performed using the above-mentioned Linear SVM classifier. Besides, we employed principle component analysis (PCA) to reduce the dimension of *fc7* features. Figure 4.3 shows the average per-

---

<sup>4</sup>We discard one of the sessions in Subject 5 from our analysis since the data for that session are corrupted for this subject. Thus, eleven-fold cross validation is performed for this subject

formance of the SVM classifier over all folds for each principle components.

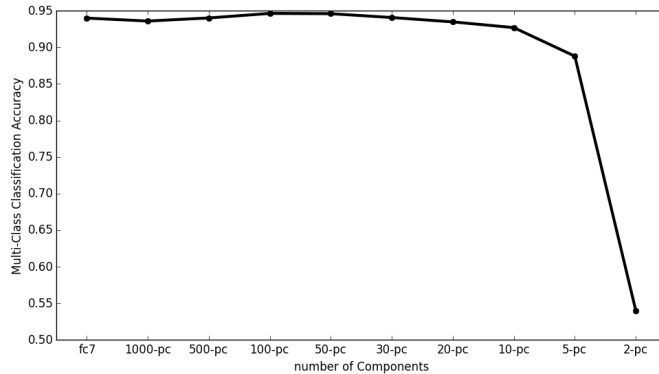


Figure 4.3: The average accuracy of the SVM classifier over all folds for each principle components.

**Experiment 2:** In this experiment, we study the cross-subject generalization power in both univariate and Atlas-based feature selection approaches. For this purpose, the VTC voxels are selected for each subject, using not only its own VTC area, but also other subjects’ VTC areas. This experiment reflects the subject bias that exists in the univariate feature selection approach. In the Atlas-based approach, however, a unique Atlas-based VTC area is used for feature selection on all subjects.

Results of these feature selection approaches are demonstrated in Table 4.1. The maximum performance of each subject is gained when his own VTC area is used for voxel selection. Using voxels obtained by VTC area of another subject drastically decreases the performance. This performance drop, is as a result of subject bias and double dipping effect [63]. As an example, the average classification accuracy for “Subject 3” using his own VTC mask is 0.68. However the accuracy drops to 0.21 if we use Subject 4 VTC mask for this subject which is slightly higher than chance (0.14). This is due to the fact that the features are selected in a subject-specific fashion prior to the analysis (without splitting the data into training and

test sets) and then a classifier is applied on the selected features. In other words, such univariate analysis, is prone to be over-fitted for each subject data, so it cannot be used for brain decoding in general. Besides, the last column of this table shows the results of Atlas-based feature selection strategy. Although this approach does not provide the best result for each subject, it performs fairly good on all subjects. Note that, this approach is not subject-dependent and can be applied to every individual. This is particularly helpful for a better generalization in brain decoding. Thus, in this study we consider atlas-based approach as our baseline in the second experiment.

Table 4.1: 7-class Classification Accuracy (average accuracy over 12 runs for each subject on each mask area)

Subjects	Sub1 VTC	Sub2 VTC	Sub3 VTC	Sub4 VTC	Sub5 VTC	Sub6 VTC	Atlas VTC
Subject 1	<b>0.79</b> $\pm$ 0.16	0.64 $\pm$ 0.16	0.66 $\pm$ 0.1	0.61 $\pm$ 0.1	0.57 $\pm$ 0.07	0.6 $\pm$ 0.11	0.64 $\pm$ 0.10
Subject 2	0.57 $\pm$ 0.07	<b>0.70</b> $\pm$ 0.08	0.54 $\pm$ 0.07	0.44 $\pm$ 0.07	0.50 $\pm$ 0.06	0.65 $\pm$ 0.06	0.59 $\pm$ 0.07
Subject 3	0.41 $\pm$ 0.07	0.30 $\pm$ 0.08	<b>0.68</b> $\pm$ 0.07	0.21 $\pm$ 0.05	0.27 $\pm$ 0.09	0.46 $\pm$ 0.05	0.43 $\pm$ 0.08
Subject 4	0.44 $\pm$ 0.08	0.32 $\pm$ 0.09	0.33 $\pm$ 0.07	<b>0.57</b> $\pm$ 0.10	0.33 $\pm$ 0.07	0.49 $\pm$ 0.08	0.49 $\pm$ 0.10
Subject 5	0.64 $\pm$ 0.08	0.61 $\pm$ 0.08	0.66 $\pm$ 0.06	0.60 $\pm$ 0.10	<b>0.77</b> $\pm$ 0.03	0.66 $\pm$ 0.05	0.66 $\pm$ 0.08
Subject 6	0.60 $\pm$ 0.06	0.49 $\pm$ 0.06	0.48 $\pm$ 0.04	0.43 $\pm$ 0.05	0.40 $\pm$ 0.07	<b>0.78</b> $\pm$ 0.05	0.65 $\pm$ 0.10

**Experiment 3:** As explained in Section 4.3.3, we used probabilistic atlases to extract voxels within the VTC area of the brain. These probabilistic atlases are defined by a threshold value that explains the probability of a specific voxel that belongs to specific cortical brain area (i.e. VTC area). This threshold can be chosen between 0 and 1. In order to study the sensitivity of the threshold on our results, we set the threshold parameter in the range of 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8. We observe that with higher number of the threshold, the size of the VTC area shrinks. Figure 4.4 shows this effect.

Besides the shape of the VTC area, we also investigate the effect of the values of the threshold on the classification results. Figure 4.5 demon-

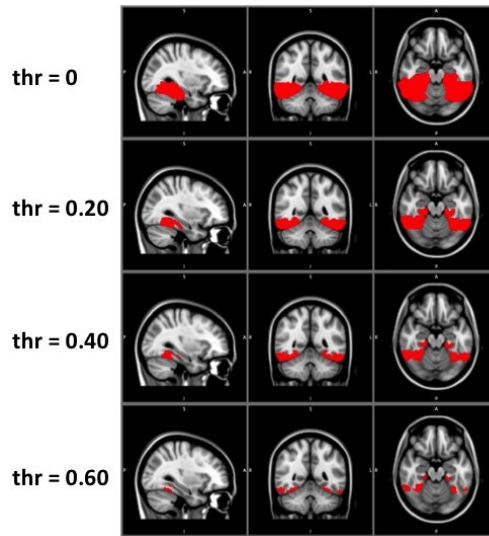


Figure 4.4: The VTC area in the brain (MNI space) using different value of threshold.

strate the classification results of each subject using various number of the thresholds.

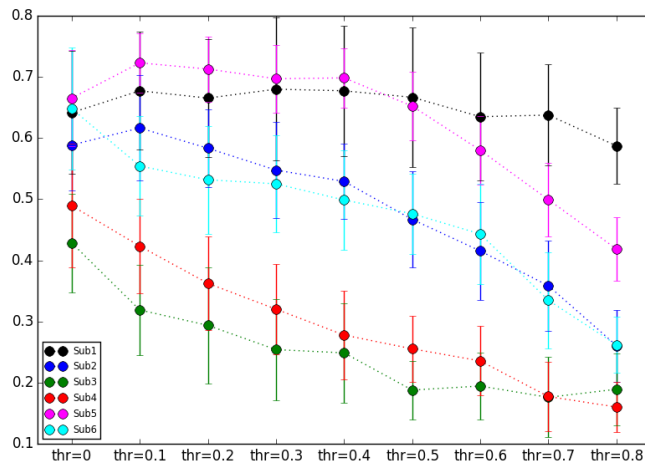


Figure 4.5: 7-class Classification Accuracy (average accuracy over all runs for each subject using different threshold values).

To allow the threshold-wise analysis, we calculate the average results of all subjects over all runs for each threshold. Table 4.2 compares the results of such analysis. The average accuracy using VTC area obtained from  $thr = 0$  is superior compared to the average accuracy using other threshold

values. As a result of this, we fix this value (the best parameter) in all our experiments.

Table 4.2: Average accuracy of all subjects over all folds for each threshold.

Threshold	Accuracy
thr = 0.0	<b>0.58 ± 0.13</b>
thr = 0.1	0.55 ± 0.16
thr = 0.2	0.52 ± 0.17
thr = 0.3	0.50 ± 0.19
thr = 0.4	0.48 ± 0.19
thr = 0.5	0.45 ± 0.19
thr = 0.6	0.41 ± 0.18
thr = 0.7	0.36 ± 0.18
thr = 0.8	0.31 ± 0.16

**Experiment 4:** To evaluate the effectiveness of our adaptation method, we use the same experimental setup as the second experiment, only replacing the VTC brain features with the Adapted-Features. In this experiment, the features are computed using the adaption method explained in Section 4.3.2.

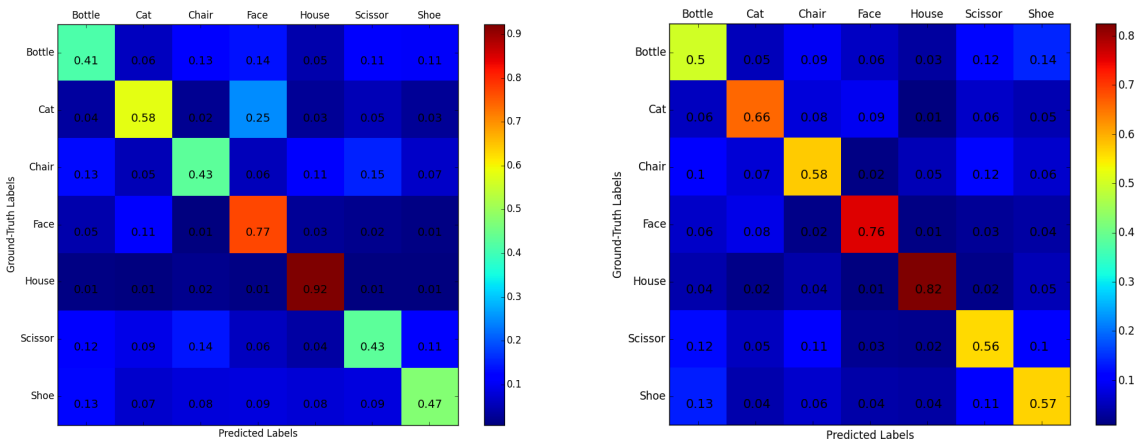
Table 4.3 summarizes the results of experiment 4. The average accuracy using the Adapted-Features is significantly superior compared to the average accuracy obtained by Brain-Features (Atlas-based approach). This difference suggests the impact of transferring knowledge from visual modality to brain modality. Regardless of the big differences in these modalities, the semantic representations learned in ImageNet are transferred successfully to brain features. Besides, our result shows significant improvement in 5 out of 6 subjects and is on a par with the baseline method for the other subject.

To allow the category-wise analysis, the confusion matrices for object classification are illustrated in Figure 4.6a and 4.6b. To facilitate the comparison, the confusion matrices are normalized with respect to the total number of samples (639). In both cases, “face” and “house” categories are

Table 4.3: Seven-Class Classification Accuracy (average accuracy over folds for each subject).

Subjects	Atlas-based Features	Adapted-Features
Subject 1	0.64 ± 0.10	<b>0.78 ± 0.09</b>
Subject 2	0.59 ± 0.07	<b>0.65 ± 0.11</b>
Subject 3	0.43 ± 0.08	<b>0.47 ± 0.07</b>
Subject 4	0.49 ± 0.10	<b>0.57 ± 0.08</b>
Subject 5	0.66 ± 0.08	<b>0.73 ± 0.05</b>
Subject 6	<b>0.65 ± 0.10</b>	0.63 ± 0.06
Average	0.58 ± 0.13	<b>0.64 ± 0.13</b>

predicted with higher confidence compared to the other categories. In 5 out of 7 categories, the classification using Adapted-Features outperforms the baseline method with a large margin. The “House” and the “Face” category, however, is predicated better using Atlas-based Features. This is probably due to the importance of Fusiform Face Area (FFA) for face recognition [56, 57] and the effect of this area might be lost after adaptation.



(a) Atlas-based Brain-Features

(b) Adapted-Features

Figure 4.6: Normalized Confusion Matrices. (a) baseline method (before adaptation). (b) proposed method (Adapted-Features).



## 4.4 Study 2: Genre Classification

Prior works in neuroimaging studies have shown that low-level audio-visual features such as orientation, direction of motion and color of visual stimulus are encoded in the human brain [33, 54, 118]. Some of these low-level audio-visual features were used also in multimedia retrieval literature for specific tasks (e.g., genre classification [104, 139]). Inspired by these facts, in this study, we address the specific problem of cross-modal adaptation by learning jointly a sparse dictionary on the low-level audio-visual features and brain features. Figure 4.7 illustrates the overview of the framework used in this study.

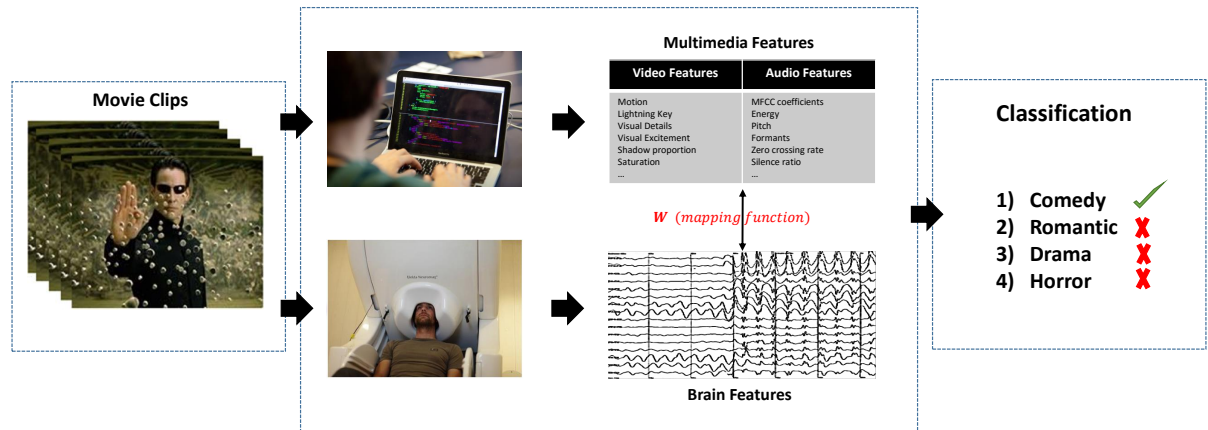


Figure 4.7: Overview of our proposed framework: During training, a dictionary learning approach is used to learn a mapping function for brain/multimedia adaptation. Once the mapping function is learned, the genre of a test movie clip is predicted using the adapted brain features.

### 4.4.1 Materials and Method

In this section, we describe the employed datasets, feature extraction scheme, and the adaptation procedure.

## Datasets

In this study, we employed two publicly available neuroimaging cross-modal datasets: DEAP dataset [61] and the DECAF dataset [1]. (see Section 4.3.3 for more details on the employed datasets). We specifically selected these two datasets in this study because, for each music/movie clip, the corresponding brain features (i.e. MEG features and EEG features) and multimedia features can be extracted.

## Annotation

Each music clip (in both datasets) is labeled with one of the following two broad genres: Pop or Rock. And Each movie clip is assigned with a label out of the following four genres: Comedy, Romantic, Drama and Horror (see Section 3.3.2 for more details on genre annotation). Note that the music video clips used in the DECAF dataset are the same clips used in the DEAP dataset.

## Source Domain: Multimedia Features

As it is explained in Section 3.3.3, for each music/movie clip, the low-level audio-visual features are extracted. These low-level Multimedia Content Analysis (MCA) features are described in Table 3.1. These MCA features are extracted for each second of the movie clips and then they were averaged by the length of the clip.

**Target Domain: Brain Features**

Brain features (MEG features and EEG features) were extracted using the same principles described in Section 3.3.3.

**MEG features:** Using the MATLAB Fieldtrip toolbox [93] and following [1], the MEG trials are extracted and pre-processed as follows: 1) Upon down-sampling the MEG signal to 300 Hz, High-pass and Low-pass filtering with cut-off frequencies of 1 Hz and 95 Hz are performed respectively. 2) Then, the spectral power of the 102 combined-gradiometer sensors of the MEG trials is estimated with a window size of 300 samples. 3) MEG features are calculated by averaging the signal power over time and four major frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz).

**EEG features:** We used the publicly available pre-processed EEG data [61]. These pre-processing steps include: EEG signal down-sampling to 128 Hz, EOG artifacts removal and bandpass frequency filtering (4 - 45 Hz). Then, for every trial, the spectral power of each channel is estimated with a window size of 128 samples. EEG features are calculated by averaging the signal power over time and over four major frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz).

**4.4.2 Adaptation Method**

To benefit sparsity-inducing properties, we first sparsify the features in both modalities. Once sparse representations are obtained, we adopted the Semi-Coupled Dictionary Learning (SCDL) approach [127] in order to adapt the sparse MCA features to the sparse brain features. This was done

for each subject separately. We refer to these features as **Adapted-Brain** features.

The intuition behind such cross-modal adaptation is that a mapping function can be found to associate the given sample in the brain domain to the corresponding sample in the multimedia domain. Since each pair of samples in two modalities refer to the same video clip, it is reasonable to assume that there exists a hidden space where the knowledge can be transferred across the two modalities. Therefore, we employ a coupled dictionary learning method with the assumption that there exists a dictionary pair over which the representations of two modalities have a stable mapping. Once the dictionary pair and mapping are learned, cross-modal domain adaptation can be performed.

We denote  $\mathbf{X}$  and  $\mathbf{Y}$  as source and target domain feature matrix, respectively.  $\mathbf{D}_x$  and  $\mathbf{D}_y$  are the dictionaries learned in the source and the target domain.  $\mathbf{\Lambda}_x$  and  $\mathbf{\Lambda}_y$  are the codes learned in the source and the target domain. We propose to optimize the following objective function below:

$$\begin{aligned} & \min_{(\mathbf{D}_x, \mathbf{D}_y, \mathbf{W})} \|\mathbf{X} - \mathbf{D}_x \mathbf{\Lambda}_x\|_F^2 + \|\mathbf{Y} - \mathbf{D}_y \mathbf{\Lambda}_y\|_F^2 \\ & + \gamma \|\mathbf{\Lambda}_y - \mathbf{W} \mathbf{\Lambda}_x\|_F^2 + \lambda_x \|\mathbf{\Lambda}_x\|_1 + \lambda_y \|\mathbf{\Lambda}_y\|_1 + \lambda_w \|\mathbf{W}\|_F^2 \end{aligned} \quad (4.2)$$

$$s.t. \quad \|d_{x,i}\|_{l_2} \leq 1, \|d_{y,i}\|_{l_2} \leq 1, \quad \forall i$$

where  $\gamma$ ,  $\lambda_x$ ,  $\lambda_y$ ,  $\lambda_w$  are regularization parameters to balance the terms in the objective function. The objective function in (4.2) is not jointly convex to  $\mathbf{D}_x$ ,  $\mathbf{D}_y$ ,  $\mathbf{W}$ . However, it is convex w.r.t. each of them if others are fixed. An iterative algorithm is designed to alternatively optimize the variables.

### 4.4.3 Experiments and Results

For the sake of compatibility with the analysis performed in the previous chapter, we employed the same classifiers (Linear SVM and Naive Bayes) under the leave-one-clip-out cross-validation schema to classify brain features (*Brain* features and *Adapted-Brain* features) into the target genre classes. Such evaluation, provides us with comparing brain features before and after adaptation. The above-mentioned pipeline was performed in the following scenarios:

#### **Subject-level analysis:**

At subject level, both classifiers were employed on the brain data of each subject separately. Figure 4.8a and 4.8b compares the results of the movie/music genre classification using brain features (MEG and EEG) before and after adaptation on all datasets (DECAF-MOVIE, DECAF-MUSIC and DEAP-MUSIC). In both DECAF and DEAP datasets, the distribution of the obtained classification accuracies using the Adapted-Brain features is far superior to the Brain features (regardless of the employed classifier). This difference implies the effectiveness of adapting brain domain to the multimedia domain. In the case of the DECAF-MOVIE and the DECAF-MUSIC datasets, this difference is significant ( $p - value < 0.01$ ).

#### **Population-level analysis:**

To evaluate the efficacy of the Adapted-Brain features at the population level, the genre of each music/movie clip is computed by majority voting over the predicted labels of single-subject predictions across all subjects.

The results are summarized in Table 4.4. In case of movie genre clas-

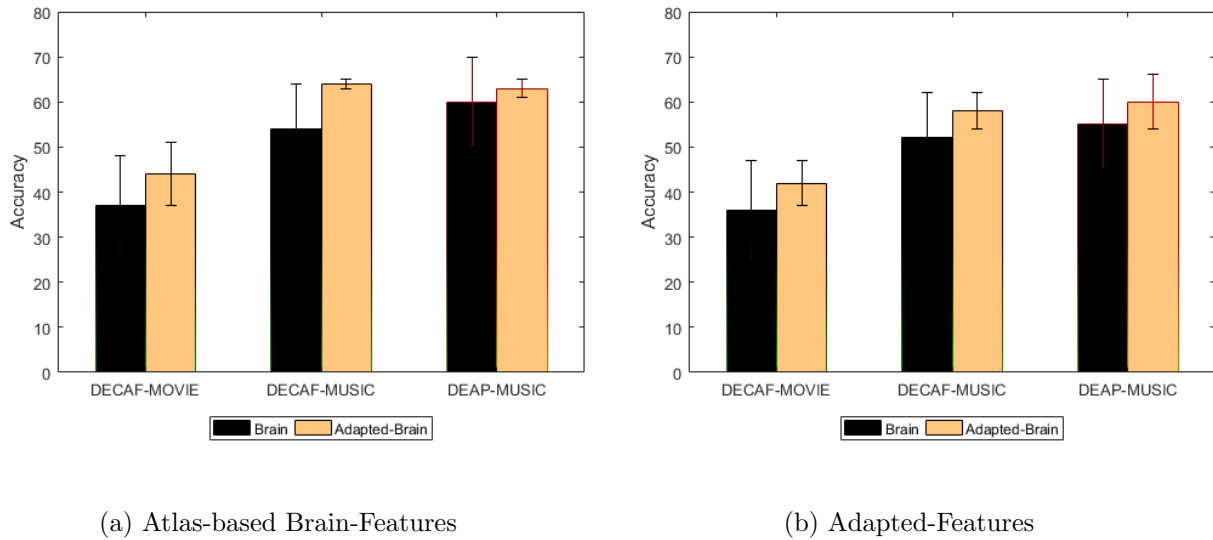


Figure 4.8: Comparison between the accuracy of Brain and Adapted-Brain features in classifying the genre of the music/movie clip in the single-subject level scenario on different datasets. (a) Using a Linear SVM Classifier. (b) Using a Naive Bayes Classifier.

sification (DECAF-MOVIE), the population level accuracy for Adapted-MEG features, using Naive Bayes classifier, is 63.9% which is higher than the accuracy of MEG features (55.6%). However, under SVM classifier, the obtained accuracy using the Adapted-MEG features are on a par with the accuracy of the MEG features. In case of music genre classification using MEG signals (DECAF-MUSIC), the population level accuracy for Adapted-MEG (65% using SVM classifier and 62.5% using Naive Bayes classifier) is higher than the accuracy of MEG features (58.3% using SVM classifier and 55.6% using Naive Bayes classifier). However, in the case of music genre classification using EEG signals (DECAF-MUSIC), the population level accuracy for Adapted-EEG features is below the accuracy of EEG features. Considering the higher accuracy of the Adapted-EEG features in the Subject-Level analysis, this phenomena is probably due to the low agreement between the predictions of all subjects.

Table 4.4: Comparison between the accuracy of Brain features and Adapted-Brain features in the population-level analysis.

Dataset	Feature-Space	Accuracy Using SVM	Accuracy Using Naive Bayes
DECAF-MOVIE	MEG	<b>58.3%</b>	55.6%
	Adapted-MEG	55.6%	<b>63.9%</b>
DECAF-MUSIC	MEG	57.5%	57.5%
	Adapted-MEG	<b>65%</b>	<b>62.5%</b>
DEAP-MUSIC	EEG	<b>75%</b>	<b>70%</b>
	Adapted-EEG	62.5%	62.5%

## 4.5 Conclusion

In this chapter, we proposed an adaptation framework in order to transfer the semantic representations learned on a rich and/or large-scale domain to the brain domain. We showed that despite the big difference between these two modalities, the adaptation procedure led to improved results for the both object classification and genre classification tasks, outperforming the the previous state of the art in all settings. We evaluated our approach on three different neuroimaging modalities (MEG, EEG and fMRI) and our cross-modal domain adaptation approach led to improved results in all of them. This is the first study in the direction of transferring knowledge from the multimedia domain to the brain modality. We believe that such approaches can overcome the limitations of the neuroimaging studies (namely, few and noisy samples) and consequently boost the performance of the decoding algorithms. Our work shows that, brain features are more informative than what we thought before. However this information is not easily obtainable because of the nature of brain studies (i.e. few and noisy samples). Using a rich representation from another modality, can help us discover these hidden information in brain signals, thus improving the performance of brain decoding algorithms. As a future plan, this study can be extended by exploring other transfer learning algorithms in order to investigate the possibility of improving results.





## Chapter 5

# Towards Mind Reading: Reconstructing External Stimuli from Brain Signals

Despite the rapid advances in Brain-computer Interfacing (BCI) and continuous effort to improve the accuracy of brain decoding systems, the urge for the systems to reconstruct the experiences of the users has been widely acknowledged.

This urge has been investigated by some researchers during the past years in terms of reconstruction of the naturalistic images [58, 86], abstract images [66], video [89] and audio [99]. However, almost all of these studies are conducted using either invasive methods or bulky expensive non-invasive methods such as MEG or fMRI which are not appropriate for out-of-lab scenarios. In this study, instead, we try to tackle this issue using EEG that is the most commonly used signal acquisition technique in BCI. In particular, we aim at regressing the stimuli spectrogram using the spectrogram analysis of the EEG signal. Figure 5.1 illustrates the overall framework used in our study. To our knowledge, we are one of the first showing the possibility of reconstructing stimuli spectrogram using EEG signals.

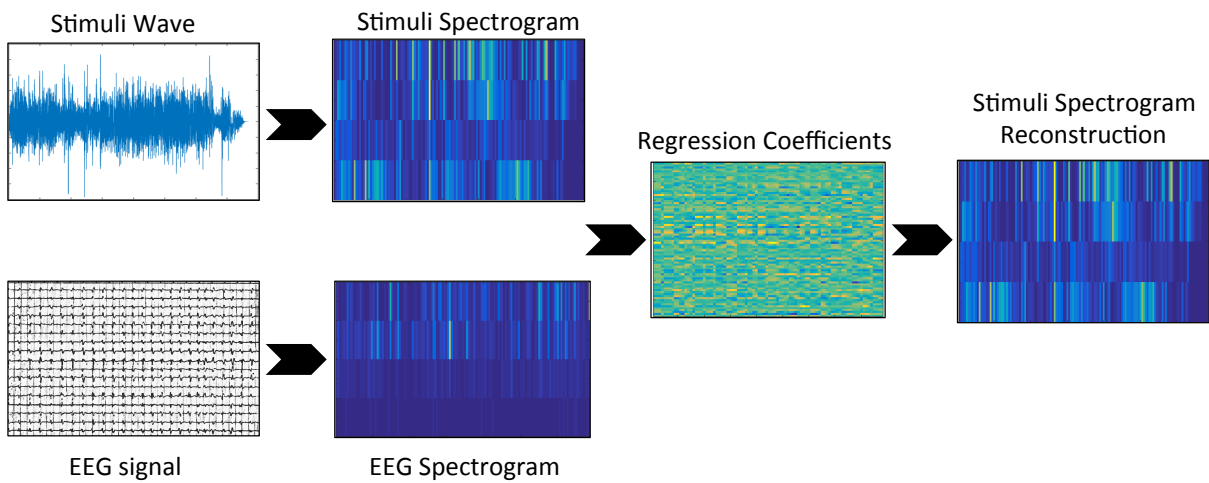


Figure 5.1: **Stimuli reconstruction pipeline:** In the first step of the analysis the spectrogram of the EEG signal and audio wave (after generic preprocessing steps) is computed. Then In the training phase, the EEG features are regressed (Ridge Regression) onto the audio spectrogram to find the proper mapping function. After that, the resulting weight matrix is used on the test EEG data to predict the spectrogram of the audio wave.

The remainder of this chapter is organized as follows: Section 5.1 reviews related literature on this topic. Section 5.2 explains the employed dataset and the data analysis methods used for data compilation. The results of such analysis are presented and discussed in Sections 5.3 and 5.4. Finally Section 5.5 concludes the chapter with the key observations and some possible future directions.

## 5.1 Literature Review

Reconstruction of someone’s experiences from his/her brain activity patterns can be considered as the ultimate goal of ”Mind Reading”. During the past recent years, many researchers tried to tackle this task in many different ways using different neuroimaging modalities and different methodologies. Here, we briefly review some of the major works on this

topic.

### 5.1.1 Reconstructing image-based stimuli

Reconstructing image-based stimuli has received more attention compared to the other types of stimuli. In the experiment conducted by Key et al. [58], the authors used fMRI to record brain activity of the subjects while they were watching 1,750 natural images. The authors used brain responses of the voxels from early visual areas. Based on these data, they built an encoding model for each voxel. This model can describe the dependency between the voxel and a particular set of image features. In the identification test, the image with highest correlation between the predicted activity pattern (based on the receptive-field models) and the measured activity pattern (based on fMRI) was selected. 82% of the images (99 out of 120 ) were identified correctly. Such encoding models were used in other studies in order to decode the mental images of the remembered scenes [112, 85]. In two very related studies, Naselaris et al. [86] and Nishimoto et al. [89] succeeded to reconstruct natural images and movies using a Bayesian framework and a database of natural image priors. In [89], natural movie stimuli were first passed through a set of Gabor filters (differing in position, orientation, direction, spatial, and temporal frequency). Then, during training, the hemodynamic response of each voxel to the stimulus is learned using L1-regularized linear regression model. This model was used to predict the bold response of a movie (in the test-set). Reconstruction was accomplished by averaging over 5 movies (from a big separate dataset that contained around 8,000,000 seconds of natural movies) with the highest probability of producing the same bold response. In another related study by Schoenmakers et al. [108], reconstruction of the handwritten characters from brain response was investigated. However, as authors in [89, 108] conclude, the exact reconstruction is not possible and the quality

of such reconstructions are highly dependent to the number and the quality of the priors. In other words, changing the prior, can yield to a completely different reconstruction.

Unlike these studies, Thirion et al. [119], Miyawaki et al. [80] and Kuo et al. [66] tried to reconstruct visual stimuli from brain activity, without any kind of image priors. This means that there exists a function that maps the visual stimuli with the corresponding brain responses elicited from such stimuli. By inverting this function, the stimuli can be reconstructed given the brain activity patterns. However, in the studies, authors just used very simple-abstract images with limited shapes. Even in such case, the correlation coefficient between the reconstruction and the original stimulus was not significant for all stimulus and all subjects.

### 5.1.2 Reconstructing audio-based stimuli

Reconstructing audio-based stimuli has received less attention compared to the visual types of stimuli (i.e. images and videos). However, recently, researchers began to explore this type of stimuli as well. One of the inspiring works is the work of Pasley et al. [99] in which the authors tried to reconstruct the spectro-temporal auditory features of spoken words and continuous sentences from neural responses using an invasive neuroimaging technique (ECoG). However the spectrogram analysis of the audio input contained the information in the very low frequencies (0.2-7 Hz). Despite this, they obtained significant correlation (average accuracy  $r = 0.28$ ), between the reconstructed spectrogram and the original wave spectrogram.

In a similar study [77], Martin et al., tried to reconstruct the spectrotemporal features of overt (reading out loud) and covert (silent reading) speech from the human cortex using ECoG. On this, they first built a neural decoding model (using high gamma band (70-150 Hz) to reconstruct spectrotemporal auditory features of the overt speech. Then they tried to

see if such model can reconstruct auditory speech features in the covert speech condition based on the hypothesis that these phenomena share a common underlying neural representation. For the overt condition, reconstruction accuracy (the correlation between original and predicted speech features) was significant in each subject. For the covert speech, however, the accuracy was lower but it was still higher than the baseline (resting-state prediction).

Strum et al., [115], tried to reconstruct the musical stimuli power-slope using a non-invasive approach (EEG). Using a Ridge Linear Regression approach, they were able to regress the temporally embedded EEG features into power-slope of the musical stimuli. They measure the canonical correlation between the signal-slope and the reconstructed slope. However in most of the cases the correlation value is very low and insignificant.

In an interesting and recent work by Huth et al. [46], authors tried to build a semantic model for each voxel by regressing the brain activity patterns of each voxel with the semantic features obtained from more than two hours of narration of the stories (from The Moth Radio Hour). Semantic features were constructed based on the word co-occurrence statistics in a large corpus of text. Once semantic models are learned, they can be employed to predict the brain activity patterns (BOLD response) of the new stimuli. Such prediction can be validated by correlating the predicted brain response with the original brain response. Such analysis revealed high correlation in some voxels, but the average performance (in between-subject analysis) is rather weak (less than 0.08). Although authors did not investigate the possibility of reconstruction of the new stimuli (narration of the story) from the brain responses, however, using such semantic voxel models might enable them to do so in further experiments.

### 5.1.3 Spotting the gap

Our examination of the related literature reveals that reconstructing the experiences of the users are recently capturing attention across many communities and will be a hot topic with the continuous advances in signal acquisition techniques and machine learning algorithms. However, these studies are conducted using the neuroimaging techniques that are not appropriate for out-of-lab scenarios. Since EEG is probably the most common signal acquisition technique in BCI, hence, it would be an interesting question to investigate whether EEG responses to the musical stimuli can be utilized to obtain information about the stimulus spectrogram or not. Insights into such question could be acquired by examining the relationship between EEG signal spectrogram on one hand and the stimulus spectrogram on the other hand. In light of this, in the present context, we propose an approach to regress the spectrogram of the stimulus using spectrogram analysis of the EEG signals in order to complement a possible link between stimulus structure and brain signals.

## 5.2 Materials and Methods

In this section, we describe the employed dataset and data analysis procedure.

### 5.2.1 Dataset

In our experiments, we used DEAP dataset [61] that contains the EEG data of volunteers who watched 40 music video clips. The complete description of this dataset is explained in Chapter 3.

### 5.2.2 Data Analysis

#### EEG Signal Processing:

In this study, the pre-processed EEG data [61] is employed. These generic pre-processing steps include: down-sampling of the EEG signal to 128 Hz, EOG artifacts removal and bandpass frequency filtering (4 - 45 Hz). Then, for every trial, the spectral power of each channel is estimated with a window size of 64 samples (500 milliseconds) with 8 samples overlap between windows (We employed Short-time Fourier transform (Equation 5.1) in order to estimate the signal power for a specific frequency and time-intervals). EEG features are calculated by averaging the signal power over four major frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz). The output of this procedure for each trial is a matrix with the following dimensions: 32 (number of the EEG sensors)  $\times$  4 (major frequency bands)  $\times$  137 (number of segmented temporal windows).

$$X_{(\tau,\omega)} = \int_{-\infty}^{+\infty} x(t)w(t - \tau)e^{-j\omega t} dt \quad (5.1)$$

#### Sensor Selection:

Not all brain regions involve in processing the auditory information in humans. The auditory cortex (Figure 5.2) that is locate bilaterally at the top of the temporal lobes, involves in perception and understanding of voices [8, 136]. In light of this, In this study, we used the sensors located in the temporal area of the brain (T7 and T8). In order to increase the signal to noise ratio, for each subject and on each clip, we averaged the time-frequency outputs of these two sensors.

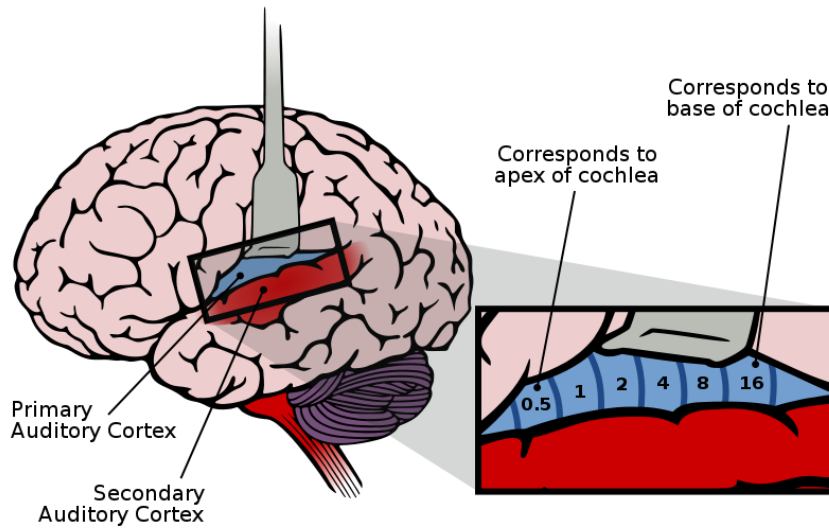


Figure 5.2: Auditory cortex in human brain

### Stimuli Processing:

For each stimulus (music video clip), the signal power is determined as follows: 1) Downsampling the audio signal to the sampling frequency of the EEG signal (128 Hz). 2) Segmenting the audio signal into the 12.5 % overlapping time frames of 500 milliseconds width. 3) Calculating the average signal power for each window for every frequency. 4) Averaging the signal power over four major frequency bands: theta (3:7 Hz), alpha (8:15 Hz), beta (16:31 Hz) and gamma (32:45 Hz).

### 5.2.3 Regression Analysis

We adopted Linear Ridge Regression models under leave-one-sample-out cross-validation schema in order to minimize the distance between the spectrogram of the EEG signals and the spectrogram of the audio signal. Thus, For a given stimulus on a particular subject a mapping function is learned based on the EEG/Audio spectrogram of 39 stimulus (training set). Then, the mapping function is applied to the EEG spectrogram of the remaining



stimulus (test set) in order to predict the spectrogram of the audio stimulus. Such procedure is repeated 40 times so that each stimulus for each subject is once used in the test set.

#### **5.2.4 Correlation Analysis**

The prediction of the spectrogram of the audio stimulus from EEG Signals is served as the basis for examining the relation between the stimulus spectrogram and the reconstructed stimulus spectrogram. In order to do so, we calculate the Pearson correlation between these two spectrograms (the original one and the reconstructed one) for each stimulus on each subject.

### **5.3 Results**

As explained in Section 5.2.4, we calculate the Pearson correlation between the spectrogram of the stimulus (in the test set) and the reconstructed spectrogram of the same stimulus using the EEG signals. This procedure, finds 40 correlation coefficient (r-values and their corresponding p-values) for each stimulus of each subject. In order to compare the results on a subject level basis, the correlation coefficients are averaged for each subject. Accordingly, for each subject, the obtained p-values are fused over all clips using the Fisher's method [27, 5]. The results of such analysis are demonstrated in Table 5.1.

### **5.4 Discussion**

The results of our regression-based method demonstrate the feasibility of the reconstruction of the spectrogram of the audio stimulus directly from

Table 5.1: Correlation analysis.

Subjects	Correlation Coefficient	p-value
Sub 1	0.0090	0.0918
Sub 2	0.0314	< 0.001
Sub 3	0.0701	< 0.001
Sub 4	0.0371	< 0.01
Sub 5	0.0068	0.6506
Sub 6	0.0290	< 0.05
Sub 7	-0.0011	0.4689
Sub 8	0.0340	< 0.001
Sub 9	0.0224	< 0.001
Sub 10	0.0432	< 0.001
Sub 11	0.0267	< 0.01
Sub 12	0.0057	0.8974
Sub 13	0.0168	< 0.001
Sub 14	0.0395	< 0.001
Sub 15	0.0293	< 0.05
Sub 16	0.0573	< 0.001
Sub 17	0.0097	0.4307
Sub 18	0.0768	< 0.001
Sub 19	0.0061	0.7059
Sub 20	0.0438	< 0.001
Sub 21	0.0764	< 0.001
Sub 22	0.0399	< 0.001
Sub 23	0.0022	0.9008
Sub 24	0.0255	< 0.001
Sub 25	0.0600	< 0.001
Sub 26	0.0486	< 0.001
Sub 27	0.0785	< 0.001
Sub 28	0.0244	< 0.05
Sub 29	-0.0044	0.3365
Sub 30	0.0363	< 0.05
Sub 31	0.0403	< 0.001
Sub 32	0.0432	< 0.001
<b>Average</b>	0.0333	$1.13 \times e^{-136}$

the EEG signals at the single-subject level. The obtained correlation coefficients are significant (p-value < 0.05) in 24 (out of 32) subjects. Nevertheless, several issues call for further exploration. First and foremost, the obtained correlation coefficient is very weak (although it is significant and it is consistent with the correlation value reported in the other literatures [77, 115, 46]). On this, other regression methods can be explored. One

of the promising technique is applying canonical correlation in order to find a subspace to maximize the correlation between two representations. Secondly, the variance of the correlation coefficient between stimuli within (and between) subjects has not been explained yet and should be explored in future experiments. Thirdly, other characteristics of the music stimulus (rather the signal spectrogram) need to be examined in the subsequent studies. In addition, it would be interesting to probe whether or not such stimuli can cause significant correlation with physiological responses, such as heart rate.

## 5.5 Conclusion

In this Chapter, we presented an approach regarding reconstruction of the audio stimulus spectrogram directly from the EEG brain signals. The presented results are just a proof-of-concept that multivariate analysis of the brain signals may extract complex stimuli related information from brain. To our knowledge, this study is one of the first efforts that employs EEG signals for such tasks. However the proposed study presents some limitations that need to be addressed in further experiments. The main limitation is regarding the correlation coefficient that is weak. For this, other regression methods, as well as the canonical correlation, should be investigated. As a future plan, this study can be extended in several ways:

1. Applying other regression methods for reconstructing audio spectrogram from EEG signals.
2. Exploring other musical features rather the spectrogram.

3. Examining the correlation between musical features and behavioral and physiological responses such as heart rate and respiration.

# Chapter 6

## Conclusion

In this chapter we summarize our research conducted for this PhD. In this thesis we investigated three aspects of brain decoding: 1) Retrieving the genre-related information from brain signals. 2) Improving the accuracy of machine learning algorithms on specific tasks by transferring the knowledge learned in another modality using a domain adaptation approach. 3) Reconstructing some aspects of the stimuli structure by analyzing the brain signals. Here we summarize our research and discuss the limitations of this work as well as the possible future directions.

### 6.1 Summary

The individuals' cognitive state and perception is determined by their brain activity. So far, brain signals have already been analyzed by the field of signal processing and computer science in order to retrieve information related to specific tasks. Such analysis revealed the feasibility of decoding brain signals. The brain decoding systems vary in tasks and the employed stimuli, ranging from resting-state brain activity to observing complex natural stimuli (e.g. movies). In this thesis, we followed this trend but we additionally worked on the idea of the recommender systems. We aspired to

make music/movie recommender systems by employing neurophysiological signals. But since genre classification can be considered as an essential part of the music and movie recommender systems, in Chapter 3 we conducted two studies on retrieving genre-related information from neurophysiological signals: Music genre classification and movie genre classification. Our analysis revealed the feasibility of such recommender systems and our results can be considered as an initial step towards an implicit music/movie recommender system via neurophysiological signals.

Later on we sought to increase the performance of brain decoding systems. This urge has arisen as a result of poor-performing brain decoding systems as a result of few and noisy samples of neuroimaging datasets. On this, we proposed a domain adaptation approach in order to take into account the rich information exists in other domains. The aim of such approach is to transfer knowledge from a rich modality to the poor-performing modality. These approaches, has been used previously in other tasks. Following up those studies, in Chapter 4, we investigated the possibility of applying transfer learning algorithms on the brain data. Firstly, we proposed an adaptation framework in order to transfer the semantic representations learned on a visual domain to the brain domain. We showed that despite the big difference between the modalities, the adaptation procedure led to improved results for the object classification task, outperforming the baseline method on the fMRI dataset. Secondly, we tried to investigate such adaptation approaches on other tasks (i.e. genre classifications) using other neuroimaging modalities (MEG and EEG). Our experimental results showed a significant boost on the performance of the brain decoding systems once adaptation has been employed.

Finally, in Chapter 5, we aimed at one of the ultimate goals of neuroscience that is reconstruction of someone's experiences from his/her brain activity. This goal has been tackled by many researchers recently in various

ways using different neuroimaging modalities and different methodologies. We tried to complement this research direction by reconstructing the audio stimulus spectrogram from the brain signals. In our analysis, we found significant correlation between the stimuli spectrogram and the reconstructed version of the stimuli spectrogram using the brain signals. Our results are a proof-of-concept that multivariate analysis of the brain signals may extract complex stimuli related information from brain.

## **6.2 Limitations and Future works**

In this section, we discuss the limitations of the research carried out throughout this PhD thesis.

### **6.2.1 Music/Movie genre classification**

We are one of the first showing that brain signals contain genre-related information, thus this can be used in movie/music recommender systems. However, in order to make useful applications of such analysis, more research needs to be conducted to guarantee the accuracy of the music/movie retrieval from neurophysiological signals. Firstly, more effective machine learning algorithms need to be employed in order to see whether the classification results can be improved. Without having a significant accuracy, it would be difficult to find any practical usage of such systems in real life. Secondly, one of the main practical limitations of our approach is related to the employed datasets that contain MEG and EEG brain data. MEG is an expensive and bulky device which is not suitable for out-of-lab scenarios. EEG, compared to MEG, is cheaper and easier to use. However it also needs significant precision once the electrodes are placed. On the other

hand, portable brain recording devices (e.g. Emotiv sensors, and Neurosky device) are easier to use for normal users [116]. Thus, one direction of future studies is replicating the same analysis by exploiting portable brain recording devices or even by using wearable neurophysiological devices.

### **6.2.2 Domain Adaptation**

While our domain adaptation showed a significant boost in the performance of the brain decoding algorithms, the interpretation of such adaptation should be further investigated in the future studies. In our analysis, we did not explore to see what kind of knowledge is shared between these two modalities and what is actually transferred between brain and the other modality. This needs to be explored in the follow-up studies.

Apart from the interpretation, our approach is just applicable in cases where data for both modalities (i.e. brain data and the multimedia data) are available. In other words, the application of this approach is limited and it is not probably suitable in real-time BCI scenarios.

### **6.2.3 Stimuli Reconstruction**

Regarding audio stimulus reconstruction, we showed a proof-of-concept that multivariate analysis of the brain signals may extract complex stimuli related information from brain signals. However the results of our correlation analysis is not very strong and should be further explored in the next studies by employing other (more efficient) regression methods. Besides, other musical features rather the signal spectrogram can be also investigated as the future plan. On top of that, it would be interesting to examine the relation between musical features and physiological responses such as



heart rate and respiration.



# Bibliography

- [1] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015.
- [2] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
- [3] Jayme Garcia Arnal Barbedo and Amauri Lopes. Automatic genre classification of musical signals. *EURASIP Journal on Advances in Signal Processing*, 2007(1):1–12, 2006.
- [4] Jean-Julien Aucouturier and Francois Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.
- [5] Timothy L. Bailey and Michael Gribskov. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48–54, 1998.
- [6] Sylvain Baillet, John C Mosher, and Richard M Leahy. Electromagnetic brain mapping. *IEEE Signal processing magazine*, 18(6):14–30, 2001.

## BIBLIOGRAPHY

---

- [7] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
- [8] Pascal Belin, Robert J Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. Voice-selective areas in human auditory cortex. *Nature*, 403(6767):309–312, 2000.
- [9] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *NIPS*, 2007.
- [10] Olivier Bertrand and Catherine Tallon-Baudry. Oscillatory gamma activity in humans: a possible role for object representation. *International Journal of Psychophysiology*, 38(3):211–223, 2000.
- [11] Niels Birbaumer. Breaking the silence: brain–computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43(6):517–532, 2006.
- [12] Darin Brezeale and Diane J Cook. Using closed captions and visual features to classify movies by genre. In *International Workshop on Multimedia Data Mining*, 2006.
- [13] Niko A Busch, Julien Dubois, and Rufin VanRullen. The phase of ongoing eeg oscillations predicts visual perception. *The Journal of neuroscience*, 29(24):7869–7876, 2009.
- [14] Thomas A Carlson, Hinze Hogendoorn, Ryota Kanai, Juraj Mesik, and Jeremy Turret. High temporal resolution decoding of object position and category. *Journal of vision*, 11(10):9–9, 2011.

- [15] Thomas A Carlson, Paul Schrater, and Sheng He. Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, 15(5):704–717, 2003.
- [16] Yandre MG Costa, LS Oliveira, Alessandro L Koerich, Fabien Gouyon, and JG Martins. Music genre classification using lbp textural features. *Signal Processing*, 92(11):2723–2737, 2012.
- [17] Emanuele Coviello, Antoni B Chan, and Gert Lanckriet. Time series models for semantic music annotation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1343–1359, 2011.
- [18] David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri)“brain reading”: detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003.
- [19] Olivier David, James M Kilner, and Karl J Friston. Mechanisms of evoked and induced responses in meg/eeg. *Neuroimage*, 31(4):1580–1591, 2006.
- [20] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, Marilyn S Albert, and Ronald J Killiany. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.
- [21] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.
- [22] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional

- activation feature for generic visual recognition. *arXiv:1310.1531*, 2013.
- [23] Guido Dornhege. *Toward brain-computer interfacing*. MIT press, 2007.
- [24] Katherine Ellis, Emanuele Coviello, and Gert RG Lanckriet. Semantic annotation and retrieval of music using a bag of systems representation. In *ISMIR*, 2011.
- [25] Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.
- [26] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.
- [27] Ronald A. Fisher. Statistical methods for research workers. *Quarterly Journal of the Royal Meteorological Society*, 82(351), 1956.
- [28] Pouya Ghaemmaghami, Mojtaba Khomami Abadi, Seyed Mostafa Kia, Paolo Avesani, and Nicu Sebe. Movie genre classification by exploiting meg brain signals. In *ICIAP*. 2015.
- [29] Pouya Ghaemmaghami, Moin Nabi, Yan Yan, Giuseppe Riccardi, and Nicu Sebe. A cross-modal adaptation approach for brain decoding. In *ICASSP*, 2017.
- [30] Pouya Ghaemmaghami, Moin Nabi, Yan Yan, and Nicu Sebe. Sparse-coded cross-domain adaptation from the visual to the brain domain. In *ICPR*, 2016.

- [31] Pouya Ghaemmaghami and Nicu Sebe. Brain and music: Music genre classification using brain signals. In *EUSIPCO*, 2016.
- [32] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.
- [33] Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8), 2014.
- [34] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. *arXiv:1507.00448*, 2015.
- [35] Stelios K Hadjidimitriou and Leontios J Hadjileontiadis. Eeg-based classification of music appraisal responses using time-frequency analysis and familiarity ratings. *IEEE Transactions on Affective Computing*, 4(2):161–172, 2013.
- [36] Peter Hansen, Morten Kringelbach, and Riitta Salmelin. *MEG: an introduction to methods*. Oxford university press, 2010.
- [37] Stephen José Hanson, Toshihiko Matsuka, and James V Haxby. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a face area? *Neuroimage*, 23(1):156–166, 2004.
- [38] James V Haxby, M Ida Gobbini, Maura L Furey, Almit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 2001.
- [39] John-Dylan Haynes and Geraint Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5):686–691, 2005.

- [40] John-Dylan Haynes and Geraint Rees. Predicting the stream of consciousness from activity in human visual cortex. *Current Biology*, 15(14):1301–1307, 2005.
- [41] John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006.
- [42] Thilo Hinterberger, Stefan Schmidt, Nicola Neumann, Jürgen Mellinger, Benjamin Blankertz, Gabriel Curio, and Niels Birbaumer. Brain-computer communication and slow cortical potentials. *IEEE Transactions on Biomedical Engineering*, 51(6):1011–1018, 2004.
- [43] Hui-Yu Huang, Weir-Sheng Shih, and Wen-Hsing Hsu. A film classifier based on low-level visual features. In *IEEE Workshop on Multimedia Signal Processing*, 2007.
- [44] Yin-Fu Huang, Sheng-Min Lin, Huan-Yu Wu, and Yu-Siou Li. Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. *Data & Knowledge Engineering*, 92:60–76, 2014.
- [45] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.
- [46] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.
- [47] Alumit Ishai, Leslie G Ungerleider, Alex Martin, Jennifer L Schouten, and James V Haxby. Distributed representation of objects in the



- human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 96(16):9379–9384, 1999.
- [48] Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Schölkopf, and Moritz Grosse-Wentrup. Transfer learning in brain-computer interfaces. *arXiv:1512.00296*, 2015.
- [49] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.
- [50] Chengcheng Jia, Yu Kong, Zhengming Ding, and Yun Raymond Fu. Latent tensor transfer learning for rgb-d action recognition. In *ACM Multimedia*, 2014.
- [51] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, 2014.
- [52] Xi Jiang, Tuo Zhang, Xintao Hu, Lie Lu, Junwei Han, Lei Guo, and Tianming Liu. Music/speech classification using high-level features derived from fmri brain imaging. In *ACM Multimedia*, 2012.
- [53] J Kalcher and G Pfurtscheller. Discrimination between phase-locked and non-phase-locked event-related eeg activity. *Electroencephalography and clinical neurophysiology*, 94(5):381–384, 1995.
- [54] Yukiyasu Kamitani and Frank Tong. Decoding motion direction from activity in human visual cortex. *Journal of Vision*, 5(8):152–152, 2005.

- [55] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.
- [56] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, 17(11):4302–4311, 1997.
- [57] Nancy Kanwisher and Galit Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476), 2006.
- [58] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.
- [59] Philip R Kennedy, Roy AE Bakay, Melody M Moore, Kim Adams, and John Goldwaithe. Direct control of a computer from the human central nervous system. *IEEE Transactions on rehabilitation engineering*, 8(2):198–202, 2000.
- [60] Erica Klarreich. Reading brains. *Communications of the ACM*, 2014.
- [61] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deep: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 2012.
- [62] Sotetsu Koyamada, Yumi Shikauchi, Ken Nakae, Masanori Koyama, and Shin Ishii. Deep learning of fmri big data: a novel approach to subject-transfer decoding. *arXiv:1502.00093*, 2015.

- [63] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540, 2009.
- [64] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [65] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.
- [66] Po-Chih Kuo, Yong-Sheng Chen, Li-Fen Chen, and Jen-Chuen Hsieh. Decoding and encoding of visual patterns using magnetoencephalographic data represented in manifolds. *NeuroImage*, 102:435–450, 2014.
- [67] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.
- [68] Mikhail A Lebedev and Miguel AL Nicolelis. Brain–machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546, 2006.
- [69] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [70] Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Hwai-San Lin. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *IEEE Transactions on Multimedia*, 11(4):670–682, 2009.

- [71] Dongge Li, Ishwar K Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern recognition letters*, 22(5):533–544, 2001.
- [72] Tao Li and Mitsunori Ogihara. Music genre classification with taxonomy. In *ICASSP*, 2005.
- [73] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *ACM SIGIR*, 2003.
- [74] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *ISMIR*, 2005.
- [75] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157, 2001.
- [76] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.
- [77] Stéphanie Martin, Peter Brunner, Chris Holdgraf, Hans-Jochen Heinze, Nathan E Crone, Jochem Rieger, Gerwin Schalk, Robert T Knight, and Brian N Pasley. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in neuro-engineering*, 7:14, 2014.
- [78] Riccardo Miotto and Gert Lanckriet. A generative context model for semantic music annotation and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1096–1108, 2012.
- [79] Tom M Mitchell, Rebecca Hutchinson, Radu S Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning

- to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004.
- [80] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.
- [81] Eva Moledano, Graham Healy, Kevin McGuinness, Xavier Giró-i Nieto, Noel E O’Connor, and Alan F Smeaton. Object segmentation in images using EEG signals. In *ACM Multimedia*, 2014.
- [82] Seong-Eun Moon and Jong-Seok Lee. EEG connectivity analysis in perception of tone-mapped high dynamic range videos. In *ACM Multimedia*, 2015.
- [83] Hiroshi Morioka, Atsunori Kanemura, Jun-ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii. Learning a common dictionary for subject-transfer decoding with resting calibration. *NeuroImage*, 111:167–178, 2015.
- [84] Jeho Nam, Masoud Alghoniemy, and Ahmed H Tewfik. Audio-visual content-based violent scene characterization. In *ICIP*, 1998.
- [85] Thomas Naselaris, Cheryl A Olman, Dustin E Stansbury, Kamil Ugurbil, and Jack L Gallant. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage*, 105:215–228, 2015.
- [86] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

## BIBLIOGRAPHY

---

- [87] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.
- [88] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012.
- [89] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.
- [90] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.
- [91] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.
- [92] Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
- [93] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 2010.
- [94] Alice J O’toole, Fang Jiang, Hervé Abdi, and James V Haxby. Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience*, 17(4):580–590, 2005.

- [95] Xinyu Ou, Lingyu Yan, Hefei Ling, Cong Liu, and Maolin Liu. Inductive transfer deep hashing for image retrieval. In *ACM Multimedia*, 2014.
- [96] François Pachet and Daniel Cazaly. A taxonomy of musical genres. In *RIAO*, 2000.
- [97] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
- [98] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce. Music genre classification via joint sparse low-rank representation of audio features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1905–1917, 2014.
- [99] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang. Reconstructing speech from human auditory cortex. *PLoS Biol*, 10(1):e1001251, 2012.
- [100] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *Signal Processing Magazine, IEEE*, 32(3):53–69, 2015.
- [101] Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.
- [102] Mark D Plumbley, Thomas Blumensath, Laurent Daudet, Rémi Gri-bonval, and Mike E Davies. Sparse representations in audio and music: from coding to source separation. *Proceedings of the IEEE*, 98(6):995–1005, 2010.

- [103] Shengsheng Qian, Tianzhu Zhang, Richang Hong, and Changsheng Xu. Cross-domain collaborative learning in social multimedia. In *ACM Multimedia*, 2015.
- [104] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):52–64, 2005.
- [105] Jonas Richiardi, Hamdi Eryilmaz, Sophie Schwartz, Patrik Vuilleumier, and Dimitri Van De Ville. Decoding brain states from fmri connectivity graphs. *Neuroimage*, 56(2):616–626, 2011.
- [106] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.
- [107] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*. 2010.
- [108] Sanne Schoenmakers, Markus Barth, Tom Heskes, and Marcel van Gerven. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961, 2013.
- [109] Mohammad Soleymani, Guillaume Chanel, Joep JM Kierkels, and Thierry Pun. Affective characterization of movie scenes based on multimedia content analysis and user’s physiological emotional responses. In *IEEE International Symposium on Multimedia*, 2008.
- [110] Mohammad Soleymani, Joep JM Kierkels, Guillaume Chanel, and Thierry Pun. A bayesian framework for video affective representation.



- In *International Conference on Affective Computing and Intelligent Interaction*, 2009.
- [111] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012.
- [112] Dustin E Stansbury, Thomas Naselaris, and Jack L Gallant. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5):1025–1034, 2013.
- [113] Sebastian Stober, Avital Sternin, Adrian M Owen, and Jessica A Grahn. Deep feature learning for eeg recordings. *arXiv preprint arXiv:1511.04306*, 2015.
- [114] Sebastian Stober, Avital Sternin, Adrian M Owen, and Jessica A Grahn. Towards music imagery information retrieval: Introducing the openmiir dataset of eeg recordings from music perception and imagination. In *ISMIR*, 2015.
- [115] Irene Sturm, Sven Dähne, Benjamin Blankertz, and Gabriel Curio. Multi-variate eeg analysis as a novel tool to examine brain responses to naturalistic music stimuli. *PloS one*, 10(10), 2015.
- [116] Ramanathan Subramanian, Julia Wache, Mojtaba Abadi, Radu Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, in press, 2016.
- [117] Masaru Sugano, Roger Isaksson, Yasuyuki Nakajima, and Hiromasa Yanagihara. Shot genre classification using compressed audio-visual features. In *ICIP*, 2003.
- [118] Keiji Tanaka. Mechanisms of visual object recognition: monkey and human studies. *Current opinion in neurobiology*, 7(4):523–529, 1997.

- [119] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis LeBihan, and Stanislas Dehaene. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104 – 1116, 2006.
- [120] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, 10(5):293–302, 2002.
- [121] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015.
- [122] Anirudh Vallabhaneni, Tao Wang, and Bin He. Brain-computer interface. In *Neural engineering*, pages 85–121. Springer, 2005.
- [123] Dimitri Van De Ville and Seong-Whan Lee. Brain decoding: Opportunities and challenges for pattern recognition. *Pattern Recognition*, 45(6):2033–2034, 2012.
- [124] Hanneke Van Dijk, Jan-Mathijs Schoffelen, Robert Oostenveld, and Ole Jensen. Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *The Journal of Neuroscience*, 28(8):1816–1823, 2008.
- [125] Jonathan D Victor and Joelle Mast. A new statistic for steady-state evoked potentials. *Electroencephalography and clinical neurophysiology*, 78(5):378–388, 1991.
- [126] Jun Wang, Eric Pohlmeier, Barbara Hanna, Yu-Gang Jiang, Paul Sajda, and Shih-Fu Chang. Brain state decoding for rapid image retrieval. In *ACM Multimedia*, 2009.

- [127] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR*, 2012.
- [128] Jonathan Wolpaw and Elizabeth Winter Wolpaw. *Brain-computer interfaces: principles and practice*. OUP USA, 2012.
- [129] Jonathan R Wolpaw, Dennis J McFarland, and Theresa M Vaughan. Brain-computer interface research at the wadsworth center. *IEEE Transactions on Rehabilitation Engineering*, 8(2):222–226, 2000.
- [130] Mark W Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M Smith. Bayesian analysis of neuroimaging data in fsl. *Neuroimage*, 45(1):S173–S186, 2009.
- [131] Min Xu, Liang-Tien Chia, and Jesse Jin. Affective content analysis in comedy and horror videos by audio emotional event detection. In *ICME*, 2005.
- [132] Min Xu, Jesse S Jin, Suhuai Luo, and Lingyu Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *ACM Multimedia*, 2008.
- [133] Min Xu, Changsheng Xu, Xiangjian He, Jesse S Jin, Suhuai Luo, and Yong Rui. Hierarchical affective content analysis in arousal and valence dimensions. *Signal Processing*, 93(8):2140–2150, 2013.
- [134] Yi-Hsuan Yang. Towards real-time music auto-tagging using sparse features. In *ICME*, 2013.
- [135] Chin-Chia Michael Yeh, Li Su, and Yi-Hsuan Yang. Dual-layer bag-of-frames model for music genre classification. In *ICASSP*, 2013.

- [136] Robert J Zatorre, Pascal Belin, and Virginia B Penhune. Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, 6(1):37–46, 2002.
- [137] Shijie Zhao, Xi Jiang, Junwei Han, Xintao Hu, Dajiang Zhu, Jinglei Lv, Tuo Zhang, Lei Guo, and Tianming Liu. Decoding auditory saliency from fmri brain imaging. In *ACM Multimedia*, 2014.
- [138] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.
- [139] Howard Zhou, Tucker Hermans, Asmita V Karandikar, and James M Rehg. Movie genre classification via scene categorization. In *ACM Multimedia*, 2010.