KU LEUVEN

FACULTY OF
ENGINEERING SCIENCE

Towards unsupervised detection of the auditory attention

Arnout Roebben

Thesis submitted for the degree of
Master of Science in
Electrical Engineering, option
Information Systems and Signal
Processing

Thesis supervisor:
Prof. dr. ir. A. Bertrand
Prof. dr. ir. T. Francart

Assessor:
Dr. ir. J. Vanthornhout
Dr. ir. S. Geirnaert

Mentor:
Ir. N. Heintz

Academic year 2021 – 2022

Preface

Dear reader, welcome to this work about unsupervised detection of the auditory attention, wherein approaches are investigated to decode whether a listener is attentive to an auditory stream. This work serves as the thesis dissertation for my Master's degree in Electrical Engineering and encapsulates the research I have been conducting past year.

During this research, I have been introduced to the wonderful field of biomedical signal processing and have been invited to implement my own ideas. Therefore, I would like to thank everyone involved.

More specifically, I would like to thank my mentor, ir. Nicolas Heintz: Thank you for guiding me through the process of (the art of) performing research and academic text writing. Our meetings really helped me understand the matter and boosted the quality of my research. Moreover, thank you for all the invested time in reading through the initial versions of the thesis text and for providing feedback about the presentations.

I would also like to thank assessor dr. ir. Jonas Vanthornhout for the explanation about his dataset, on top of which a great part of this work is built. The same goes for Elly Brouckmans and Linsey Dewit-Vanhaelen for granting me access to their dataset. Furthermore, I would like to thank thesis supervisor prof. dr. ir. Alexander Bertrand, and the assessors dr. ir. Jonas Vanthornhout and dr. ir. Simon Geirnaert for providing feedback after the midterm presentation. This resulted in a fresh look at my research results and presentation. Also my gratitude towards thesis supervisor prof. dr. ir. T. Francart for co-supervising this work. Finally, I would like to thank my friends, family and especially my girlfriend Ellen for their support during this work in particular, and my studies in general.

There is still a lot to be explored in the field of biomedical signal processing and certainly in the conversion of these algorithms into unsupervised ones. I hope that this work may serve as a building block in future research. In addition, I hope that this work persuades you, the reader, about the importance of research in this domain. However, foremost, I hope you enjoy reading this thesis dissertation.

Arnout Roebben

Contents

Abstract

Hearing-impaired people lack the ability to tune towards a speaker of interest in a multispeaker environment. For this purpose, neurosteered hearing devices are designed, wherein an auditory attention decoding (AAD) algorithm selects the attended speaker. Subjects could nevertheless be inattentive to any auditory stream, interfering with the setup. To this end, this work focuses on auditory attention measures. In addition, state-of-the-art attention detection requires subject-specific data, posing practical constraints, such that unsupervised implementations are pursued.

A first objective is the selection of the auditory attention features themselves. A second objective is the conversion of these features into unsupervised ones. A third objective consists in the application of these features in an AAD framework. A dataset concerning auditory and visual attention is used, as well as a dataset concerning attention towards audio, mathematical exercises and texts for validating final performance.

Regarding the first objective, least squares and canonical correlation analysis (CCA) prove discriminating between the attention cases. Furthermore, spectral entropy proves discriminating, whereas the band-power does not. In addition, a novel, Kullback-Leibler divergence (KLD) based feature attains a higher mean accuracy than the aforementioned features on low decision lengths. Combining the KLD and entropy features seems beneficial in both datasets. Finally, common spatial pattern attains the highest performance, yet suffers from a lack of interpretability.

Regarding the second objective, results are focused on the least squares feature. CCA and principal component analysis (PCA) are applied to the EEG of both a source subject (not under study) and target subject (under study) in a domain adaptation manner. CCA proves to outperform the unadapted feature, whereas PCA does not. In addition, a novel discriminator-based approach is proposed. This method proves to be inferior to the unadapted feature, possibly due to the limited flexibility of said feature. Neither method does achieve correlation levels equal to the state-of-the-art iterative design, although the CCA method yields comparable differences in mean correlation. In fact, combining CCA and the iterative procedure seems to boost the mean difference in correlation significantly with respect to the individual methods.

Regarding the third objective, auditory attention detection seems to allow for a reduction in the AAD training set size without compromising on performance.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

A	Anterior
AAD	Auditory attention decoding
ANOVA	Analysis of variance
BCI	Brain computer interface
BP	Band-power
C	Central
CB	Central-back
CC	Central-central
CCA	Canonical correlation analysis
CCA-DA	Canonical correlation analysis domain adaptation
CF	Central-front
CSP	Common spatial pattern
DA	Domain adaptation
D-DA	Discriminator domain adaptation
EEG	Electroencephalogram
F	Frontal
FDR	False discovery rate
GEVD	Generalised eigenvalue decomposition
GEVD-MWF	Generalised eigenvalue decomposition multichannel Wiener filter
I	Inion
ILS	Iterative least squares
KKT	Karush-Kuhn-Tucker
KLD	Kullback-Leibler divergence
LASSO	Least absolute shrinkage and selection operator
LB	Left-back
LC	Left-central
LDA	Linear discriminant analysis
LF	Left-front

LME	Linear mixed-effects model
LOSO	Leave-one-subject-out
LS	Least squares
LT	Left-temporal
MISO	Multiple input single output
MWF	Multichannel Wiener Filter
N.a.	Not applicable
O	Occipital
P	Parietal
PCA	Principal component analysis
PCA-DA	Principal component analysis domain adaptation
PDF	Probability density function
PSD	Power spectral density
QCQP	Quadratically constrained quadratic problem
RB	Right-back
RC	Right-central
RF	Right-front
RT	Right-temporal
SA-DA	Subspace alignment domain adaptation
SI-CV	Subject independent cross validation
SNR	Signal-to-noise-ratio
SS-CV	Subject specific cross validation
SS-V	Subject specific validation
SVD	Singular value decomposition
SVM	Support vector machine
T	Temporal
TPCA-DA	Target principal component analysis domain adaptation

Symbols

Matrix

a	Scalar
$\mathbf{a}$	Vector
A	Matrix
$A^\top$	Transpose of matrix A
A^{-1}	Inverse of matrix A
$\mathbb{I}_{m\times n}$	$m \times n$ identity matrix
$\mathbf{0}_{m\times 1}$	m-dimensional column vector, wherein each entry equals 0
$0_{m\times n}$	$m \times n$ matrix, wherein each entry equals 0
$\mathbf{1}_{m\times 1}$	m-dimensional column vector, wherein each entry equals 1
$1_{m\times n}$	$m \times n$ matrix, wherein each entry equals 1

A_{ij}	Element of matrix A at the ith row and the jth column
$A_{i,:}$	Vector of the elements at row i of matrix A
$A_{:,i}$	Vector of the elements at column i of matrix A
$det(.)$	Determinant
$diag(.)$	Diagonal matrix
$Tr(.)$	Trace
$\|\|.\|\|_F$	Frobenius norm

Mathematics

$\|\|.\|\|_p$	p-norm, $p \in \mathbb{N}_0$
$min(.)$	Minimum
$max(.)$	Maximum
$log(.)$	Natural logarithm
$Var(.)$	Variance
$E\{.\}$	Expected value operator
$\sigma(.)$	Logistic function
$O(.)$	Big O notation
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$F(m, n)$	F-distribution with m and n degrees of freedom
H_0	Null-hypothesis
$\rho(.)$	Pearson correlation

Units

ms	Millisecond
s	Second

Electroencephalography-audio

$\mathbf{x}(t)$	EEG signal at time t, concatenated over the channels
$y(t)$	Audio envelope at time t
c	EEG channel index
L_d	Lag difference at decoder: $L_{d-s} - L_{d-e}$
L_{d-e}	Ending lag value at decoder
L_{d-s}	Starting lag value at decoder
L_e	Lag difference at encoder: $L_{e-s} - L_{e-e}$
L_{e-e}	Ending lag value at encoder
L_{e-s}	Starting lag value at encoder
f_s	Sampling frequency
$\mathcal{P}(f)$	Power spectral density
$\bar{\mathcal{P}}(f)$	Power spectral density estimate
$\bar{.}$	Estimated variable

T	Time
$\cdot_{\mathcal{S}}$	Source domain
$\cdot_{\mathcal{T}}$	Target domain

Chapter 1

Introduction

1.1 Introduction

This chapter aims to outline the goal of this thesis dissertation, namely ***detecting*** *the* ***(in)attention*** *of a subject with respect to an* ***auditory stream***. To tackle this problem, electroencephalogram (EEG) signals are utilised. These EEG signals are detailed in section 1.2: What these EEG signals are, what the associated properties are, by what artefacts these signals can be corrupted and how these signals relate to auditory stimuli. Using this EEG concept, the thesis is motivated and the corresponding aims are elaborated in section 1.3. Finally, section 1.4 concludes by means of a chapter-by-chapter overview.

1.2 The electroencephalogram

In short, the electroencephalogram (EEG) measures the electrical activity of the brain [1]. The brain indeed consists of fundamental cells, called neurons, which can be modelled as dipoles [1]. Synchronised activity of large groups of these 'dipoles' generates electrical activity/potentials, measurable in a noninvasive way by placing sensors across the scalp. Since this electrical activity manifests itself as a potential, the sensor signals need to be measured with respect to a reference sensor [1]. An example of an EEG cap consisting of 64 electrodes, also called electrodes or channels, can be seen in figure 1.2.

Now, since the EEG sensors capture the electrical 'image' of the brain in a realtime way, EEG has a good temporal resolution [1]. On the other hand, the spatial resolution of this electrical 'image' is poor because the electrical signals get filtered by the structures in the brain on their path from source to electrode [1]. In addition, neighbouring electrodes pick up similar signals originating from the same sources.

These EEG signals can nevertheless be subdivided based on their sensor position. Indeed, the sensors are grouped based on the cortical area they belong to and their distance from the midline of the head [2]:

- The character prefix links the EEG sensor to a specific cortical area: frontal (F), anterior (A), parietal (P), central (C), temporal (T), occipital (O) and inion (I). Sensor names in between these areas are made up of combinations of these prefixes.
- The character postfix is negative for sensors in the left hemisphere and vice versa for positive postfixes. Larger values denote larger distances from the midline and z denotes a sensor on the midline.

Similarly, the EEG signals can be subdivided in a spectral manner into 6 different frequency bands: delta ($[0.5, 4]$ Hz), theta ($[4, 8]$ Hz), alpha ($[8, 13]$ Hz), beta ($[13, 30]$ Hz), gamma ($[30, 90]$ Hz) and high gamma ($>$90 Hz) [1].

However, in addition to electrical brain activity, the EEG sensors also record electrical activity from other sources, either from a physiological origin (e.g. muscle activity and eye movement) or a technological origin (e.g. electrode pop and power grid interference) [3]. As an example, figure 1.3[1] shows an EEG signal corrupted by eye blink (red) and muscle (lime) artefacts. Eyeblink artefacts manifest themselves as high amplitude peaks in the frontal channels. Spectral-wise these eye blinks overlap with the delta and theta band [4]. Muscle artefacts, e.g., caused by ear and neck movement, mainly manifest themselves as fast-varying signals, overlapping with the beta and gamma frequency bands (typically $[20, 60]$ Hz) [4].

In addition, **the neural response of the brain tracks the slow variations of a speech signal, the so-called auditory envelope** [5]. Figure 1.1 shows such an envelope over an accompanying auditory signal. This envelope tracking process in the brain is referred to as neural phase locking and starts as soon as the speech signal enters the early auditory areas in the brain [6].

1.3 Aims

The overall goal of this thesis dissertation is to ***detect*** *whether a subject is* ***attentively listening*** *to a speaker*, i.e., being able to differentiate between subjects who are attentive to an auditory stream and subjects who are not. To this end, auditory streams and accompanying EEG data are leveraged. In the next paragraph, we illustrate the potential relevance of this thesis by means of a possible application: enhancing auditory attention decoding (AAD) for the cocktail party problem [7, 8].

The human auditory system possesses the capability to tune towards a speaker of interest in a crowded room while filtering out competing speakers [7]. However, hearing-impaired subjects lack the ability to do so [9]. This problem is generally referred to as the cocktail party problem [7]. Indeed, while hearing aids are equipped to enhance source signals using beamformers [10], the question still remains which speaker to enhance. To tackle this problem, auditory attention decoding (AAD) is

[1] This plot has been generated using EEGLab, as available at https://github.com/sccn/eeglab on 28/10/2021.

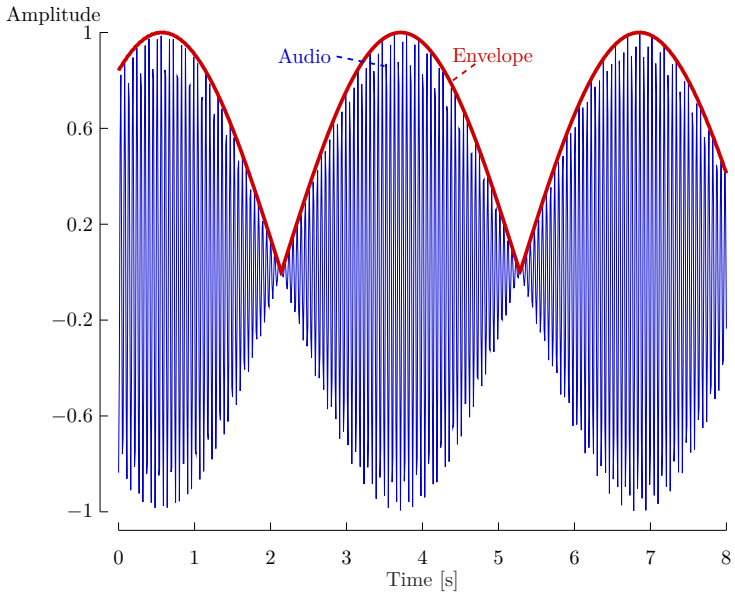


FIGURE 1.1: Audio signal and accompanying envelope.

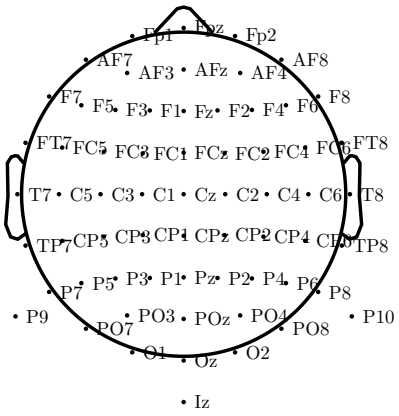


FIGURE 1.2: 64-channel EEG cap: placement of the sensors across the scalp.

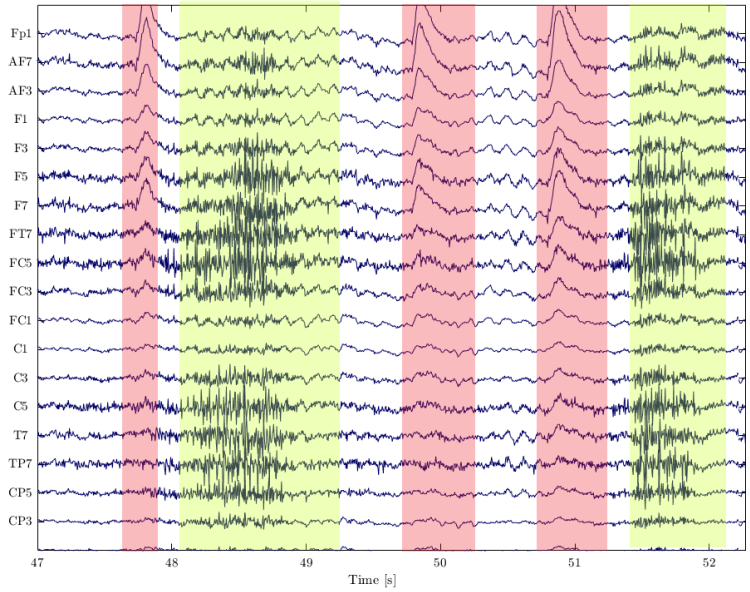


FIGURE 1.3: EEG signal with indicated muscle artefacts (lime) and eye blink artefacts (red). Eyeblink artefacts appear as high amplitude peaks in the frontal channels, whereas muscle artefacts appear as white noise-like signals.

introduced [8]. The goal of this AAD algorithm is to select the speaker of interest, i.e., to select the speaker that the user is currently paying attention to. A user is however not necessarily attentive to any auditory stream, which might interfere with the AAD setup. To this end, it might be useful to detect whether the user is being (in)attentive to any auditory stream and this is exactly where our research comes in.

To achieve this goal, we will follow a classic machine learning approach, as illustrated in figure 1.4. The audio and EEG data first pass through a preprocessing step to prepare the data for feature extraction. Thereafter, handcrafted features are extracted and a binary classifier categorises the data either as attentive to the auditory stream or not attentive to it.

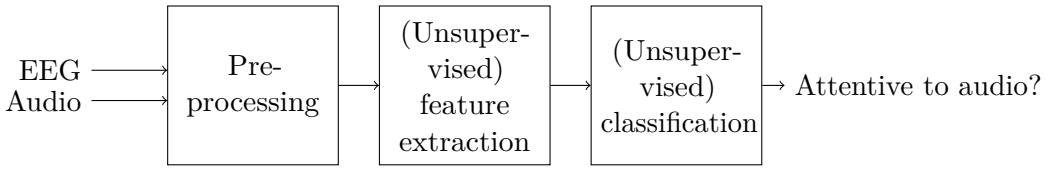


FIGURE 1.4: High-level overview of the framework attained in this work.

This approach, nonetheless, still suffers from two problems. Firstly, feature extractor and classifier design are typically performed in a subject-specific manner [8, 11, 12]. Therefore, training data need to be acquired for each subject, which poses practical constraints on the methods. Secondly, the design typically happens under stationarity assumptions. However, in practise these assumptions do not hold due

to environmental changes such as a displaced EEG cap, malfunctioning sensors and changes in background EEG activity [13]. This motivates us to look into unsupervised approaches, mitigating the above problems.

Taking the above considerations into account, we can formulate the following high-level objectives:

1. Which **features** can **discriminate** between **attention** and inattention to an **auditory stream**?
2. What **classification performance** can be attained by a feature extractor-classifier combination?
3. How can the classic machine learning framework be **converted** into an **unsupervised** one?
4. Does the **selection** of **high attention** time frames **improve** the **auditory attention decoding** (AAD) framework?

These objectives allow to take a step towards the ultimate goal of integrating an unsupervised auditory attention selection framework into an unsupervised AAD framework, as illustrated in figure 1.5, with practical applications in neurosteered hearing devices. This work indeed investigates the subparts of the auditory attention selection algorithms, the combination of these algorithms with AAD designs and *the conversion of these attention selection algorithms into unsupervised ones.*

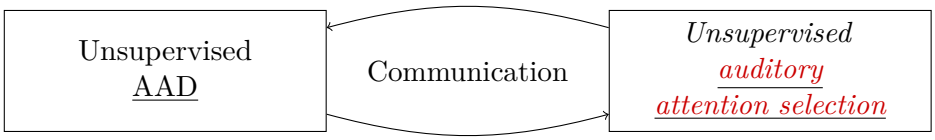


FIGURE 1.5: Connection between AAD and detection of the auditory attention.

1.4 Chapter-by-chapter overview

This text contains the following chapters:

Chapter 2: Preprocessing

This chapter describes the preprocessing methodology regarding EEG and audio data. Herein, two important concepts are artefact removal and envelope extraction. The artefact removal combats the influence of other electrical sources in the EEG and is based on a multichannel Wiener filtering (MWF) framework. The envelope extraction distils the envelope of the audio signal since the brain locks to these slow auditory variations. Thereafter, these blocks are incorporated into a joint preprocessing framework for EEG and audio.

Chapter 3: Feature extraction

A theoretical overview of the feature extraction methodologies is given in this chapter. Based on their operational principles, the features can be divided into two classes: neural envelope tracking based features and brain activity based features. The envelope tracking features relate the neural tracking of the speech envelope to the ground truth envelope and the brain activity features solely exploit properties of the EEG signals themselves. Within the class of neural envelope tracking based features a novel feature extractor, based on the concept of the Kullback-Leibler divergence, is presented as well. Furthermore, practical estimation of the features on finite length data and methodologies to avoid overfitting are discussed. In addition, there is a discussion about the neurological interpretation of the designed feature extractors.

Chapter 4: Classification

This chapter treats the classification step in the classic machine learning approach. In this work, we will focus on a particular classifier, namely the linear discriminant analysis (LDA) classifier. This chapter also looks into regularisation and input normalisation methods.

Chapter 5: Conversion to unsupervised algorithms

Due to the practical constraints of acquiring subject-specific data and due to changing statistics at inference time, the feature extraction and/or classification methods should be converted into unsupervised ones. To this end, the concept of domain adaptation is leveraged: the adaption of a feature extractor to combat the changing statistics at inference time. Herein, five such methods are described, including one novel expression (least squares-linear discriminant analysis discriminator domain adaptation (D-DA)). The other methods are novel in the sense of the application domain.

Chapter 6: Experimental procedures

Before being able to perform experiments, the experimental procedures need to be determined. This chapter describes the validation methodologies and presents the hyper-parameter choices, which will be adhered to in chapters 7-10. Furthermore, hypothesis testing, based on a framework of linear mixed-effects models and analysis of variance hypothesis tests, is described, as well as the correction of the resulting p-values. Subsequently, an overview of the dataset is given.

Chapter 7: Differentiating nature features

This chapter treats the first high-level objective: *Which **features** can **discriminate** between **attention** and inattention to an **auditory stream**?* To this end, the features as described chapter 3 are applied to the dataset.

Chapter 8: Feature extractor-classifier performance

This chapter treats the second high-level objective: *What **classification performance** can be attained by a feature extractor-classifier combination?* The features of chapter 3 are combined with the classifier of chapter 4 to study the combined performance and to gain additional insight into the feature extraction approaches. Moreover, the window length influence of this performance is investigated.

Chapter 9: Performance unsupervised algorithms

This chapter treats the third high-level objective: *How can the classic machine learning framework be converted into an **unsupervised** one?* Herein, the domain adaptation methodologies from chapter 5 are compared to the supervised ones.

Chapter 10: Boosting auditory attention decoding

This chapter treats the fourth and final high-level objective: *Does the **selection** of **high attention** time frames **improve** the **auditory attention decoding** (AAD) framework?* Herein, it is investigated whether training feature extractors solely on high attention segments leads to an increased performance of said auditory attention decoding framework.

Chapter 11: Conclusion and future work

This chapter winds up the thesis dissertation with a conclusion and an outlook to future work.

Appendices

Appendix A provides a non-convexity proof for the novel Kullback-Leibler divergence feature extractor and for the novel least squares-linear discriminant analysis discriminator domain adaptation (D-DA) method, as will be presented respectively in sections 3.3.5 and 5.3.2. Appendix B shows that there is evidence against the normality assumption of the feature space and the assumption that attention and inattention classes have the same covariance matrix, desirable for the linear discriminant analysis classifier. Subsequently, appendix C shows the topographic patterns, generated by yet another feature extractor of chapter 3, namely common spatial pattern. This appendix has to be seen in conjunction with the experiments of chapters 7 and 8. Finally, appendix D provides detailed numeric mean and standard deviations of the classification accuracies, according to the experiments of chapter 8.

Chapter 2

Preprocessing

2.1 Introduction

Before being able to reliably extract features, the EEG and audio data need to be preprocessed. Herein, two important building blocks are EEG artefact removal and auditory envelope extraction. As noted in section 1.2, artefact removal is important to diminish the impact of interfering electrical sources (e.g. eyes and muscles) in the feature values. Indeed, features could possibly exploit the artefacts, exaggerating their performance, or the other way around, features could fail due to the artefact influence. To this end, a general artefact removal procedure based on generalised eigenvalue decomposition (GEVD) multichannel Wiener filters (MWF) [14] is detailed in section 2.2. Furthermore, as noted in section 1.2, envelope extraction is necessary since the EEG signals track this envelope [5]. The procedure to achieve this envelope extraction is described in section 2.3 [15]. These two building blocks are integrated into a larger preprocessing framework, as is detailed in section 2.4 [16]. Finally, section 2.5 concludes this chapter.

2.2 EEG artefact removal

As detailed in section 1.2, **the EEG signal is polluted by interfering electrical sources, such as the eyes or muscles. These artefacts may be exploited by the features, overvaluing their performance. Alternatively, features might suffer from the influence of these artefacts.** Therefore, the artefacts need to be removed before issuing a feature extraction phase. To this end, in [14], a generic EEG artefact removal framework is presented. The general idea is to reconstruct the artefact from the EEG data and thereafter to subtract this artefact from the recorded EEG signal. Firstly, the mathematical solution is described. Thereafter, automated eye and muscle artefact detection procedures are detailed.

Generalised eigenvalue decomposition based multichannel Wiener filter

Let $X_c(t) \in \mathbb{R}$ represent the EEG signal at time $t \in \mathbb{N}$ in channel $c \in \mathbb{N}_0$, $\mathbf{x}(t) \in \mathbb{R}^{C\times 1}$ the channel-stacked EEG signal at time t and $\underline{\mathbf{x}}(t) \in \mathbb{R}^{L_dC\times 1}$ the channel-stacked and time-lagged EEG signal at time t with starting lag $L_{d-s} \in \mathbb{N}$ and ending lag $L_{d-e} \in \mathbb{N}$:

$$\begin{aligned}
\mathbf{x}(t) &= \begin{bmatrix} X_1(t) & X_2(t) & \dots & X_C(t) \end{bmatrix}^\top; \\
\underline{\mathbf{x}}(t) &= \begin{bmatrix} \underline{\mathbf{x}}_1(t)^\top & \underline{\mathbf{x}}_2(t)^\top & \dots & \underline{\mathbf{x}}_C(t)^\top \end{bmatrix}^\top; \\
\underline{\mathbf{x}}_c(t) &= \begin{bmatrix} X_c(t+L_{d-s}) & X_c(t+L_{d-s}+1) & \dots & X_c(t+L_{d-e}-1) \end{bmatrix}^\top, \qquad c = 1..C.
\end{aligned} \tag{2.1}$$

This artefact-corrupted EEG signal $\mathbf{x}(t)$ can be seen as the combination of artefact-free EEG $\mathbf{v}(t) \in \mathbb{R}^{C\times 1}$ and the artefact $\mathbf{a}(t) \in \mathbb{R}^{C\times 1}$ itself, i.e., $\mathbf{x}(t) = \mathbf{v}(t) + \mathbf{a}(t)$. All signals are assumed to be zero-mean. The goal is to reconstruct the artefact $\mathbf{a}(t)$ and to subsequently subtract it from the artefact-corrupted EEG $\mathbf{x}(t)$. This artefact reconstruction is achieved by designing a filter $W \in \mathbb{R}^{L_dC\times C}$, called a multichannel Wiener filter (MWF), according to the following cost function [14]:

$$\min_W \quad \mathbb{E}\{\|\mathbf{a}(t) - W^\top \underline{\mathbf{x}}(t)\|_2^2\}. \tag{2.2}$$

This cost function can be interpreted as retrieving the artefact EEG by minimising the mean squared error between the artefact data and the filtered artefact-corrupted data. Define $R_{\underline{xx}} = E\{\underline{\mathbf{x}}(t)\underline{\mathbf{x}}(t)^\top\}$, $R_{\underline{vv}} = E\{\underline{\mathbf{v}}(t)\underline{\mathbf{v}}(t)^\top\}$ and $R_{\underline{aa}} = E\{\underline{\mathbf{a}}(t)\underline{\mathbf{a}}(t)^\top\} \in \mathbb{R}^{L_dC\times L_dC}$ respectively as the artefact-corrupted, artefact-free and artefact correlation matrices. Then, it can be shown that the following expression minimises equation 2.2 [14]:

$$\begin{aligned}
R_{\underline{xx}} &= Q diag\left(\lambda_{\underline{x},1}, \dots, \lambda_{\underline{x},L_dC}\right) Q^\top; \\
R_{\underline{vv}} &= Q diag\left(\lambda_{\underline{v},1}, \dots, \lambda_{\underline{v},L_dC}\right) Q^\top; \\
W &= R_{\underline{xx}}^{-1} Q diag\left(\lambda_{\underline{x},1} - \lambda_{\underline{v},1}, \dots, \lambda_{\underline{x},L_dC} - \lambda_{\underline{v},L_dC}\right) Q^\top \begin{bmatrix} \mathbf{e}_1 & \dots & \mathbf{e}_C \end{bmatrix}.
\end{aligned} \tag{2.3}$$

Herein, the columns of $Q \in \mathbb{R}^{L_dC\times L_dC}$ are called the generalised eigenvectors of the matrix pencil $(R_{\underline{xx}}; R_{\underline{vv}})$ corresponding to eigenvalues $(\lambda_{\underline{x},1}, \dots, \lambda_{\underline{x},L_dC}; \lambda_{\underline{v},1}, \dots, \lambda_{\underline{v},L_dC}) \in \mathbb{R}$. In addition, $\mathbf{e}_c \in \mathbb{R}^{L_dC\times 1}$, $c = 1..C$ denotes the vector with element 1 at the position $(c-1)L_d+1$ and 0 elsewhere. Given the GEVD nature of the solution, filter W is referred to as a generalised eigenvalue decomposition multichannel Wiener filter (GEVD-MWF). The artefact-free EEG $\mathbf{v}(t)$ can subsequently be approximated as follows:

$$\begin{aligned}
\mathbf{v}(t) &= \mathbf{x}(t) - \mathbf{a}(t); \\
&\approx \mathbf{x}(t) - W^\top \underline{\mathbf{x}}(t).
\end{aligned} \tag{2.4}$$

Automated eye and muscle artefact detection

The question remains how to annotate the EEG data as being artefact-free or artefact-corrupted. To this end, we opt for automated artefact detection techniques. These techniques leverage the spatial and spectral properties of the specific artefacts, as detailed in section 1.2 [4, 16].

As mentioned in section 1.2, eye blink artefacts appear as high amplitude peaks in the frontal channels. Therefore, the segments are labelled as artefact segments whenever the power in the frontal channels[1] is 5 times larger than the average power in those channels [4, 16].

Similarly, muscle artefacts, caused by ear and neck movement, situate themselves around the sensors at the side of the head. Furthermore, these muscle artefacts typically live within the $[20, 60]$ Hz frequency range [4]. Therefore, to detect these artefacts, the EEG data are firstly bandpass filtered using a zero-phase Chebyshev type 2 filter that attains an 80 dB attenuation at the frequency 10% out of the $[20, 60]$ Hz passband. Segments are subsequently labelled as muscle artefacts if the power in the channels at the side of the head[2] have a power 60 times larger than the average power in those channels.

2.3 Envelope extraction

Before extracting features, the envelope of the speech signal needs to be extracted since **the EEG signal tracks this speech envelope** [5, 8, 15]. In this way, features can exploit this neural tracking property. In [15], a framework to extract this auditory envelope is proposed using three steps: filter bank filtering, power law application and subband combining.

Firstly, the audio signal is split into 28 subbands using a gamma-tone filter-bank consisting of 28 filters. Within this filter-bank, the filters have centre frequencies ranging from 50 Hz to 5 kHz, spaced according to 1 equivalent rectangular bandwidth [17, 18]. This filter-bank models the frequency-selectivity of the human auditory system. Indeed, the human auditory range is partitioned into critical bands which increase in width with increasing frequency, and within such a critical band, humans cannot distinguish sounds whenever other sounds are present [15, 19].

Secondly, a power law is applied to each subband, i.e., $|y_k(t)|^{0.6}$, $k = 1..28$, with $y_k(t) \in \mathbb{R}$ the kth subband signal. This operation extracts the stimulus envelope per subband and compensates for the fact that the connection between the intensity of

[1] More specifically, the available subset of following channels is used: Fp1, AF7, AF3, Fpz, Fp2, AF8, AF4 and AFz.

[2] More specifically, the available subset of following channels is used: AF7, F7, F5, FT7, FC5, T7, C5, TP7, CP5, P7, P5, P9, PO7, AF8, F6, F8, FC6, FT8, C6, T8, CP6, TP8, P6, P8, PO8 and P10.

the auditory signal and the perceived loudness is nonlinear [15, 20].

Finally, the 28 subbands are recombined into a single auditory envelope using equal weights per subband [15].

2.4 Preprocessing framework

The previously presented blocks (artefact removal and envelope extraction) fit into a general framework of jointly preprocessing EEG and audio data[3,4]. This scheme is proposed in [16] and summarised in figure 2.1. Only the muscle artefact removal block (shaded in red) is additionally included with respect to [16]. Moreover, the blocks which are shared in the preprocessing of audio and EEG are shaded in lime. The goal of this preprocessing is fourfold: extracting the auditory envelope, selecting 'interesting' frequency ranges, removing artefacts and downsampling the signal. Indeed, the filtering operation is introduced to leverage the spectral characteristics of the EEG and the downsampling operation allows to reduce the computational cost of the feature extractor-classifier framework.

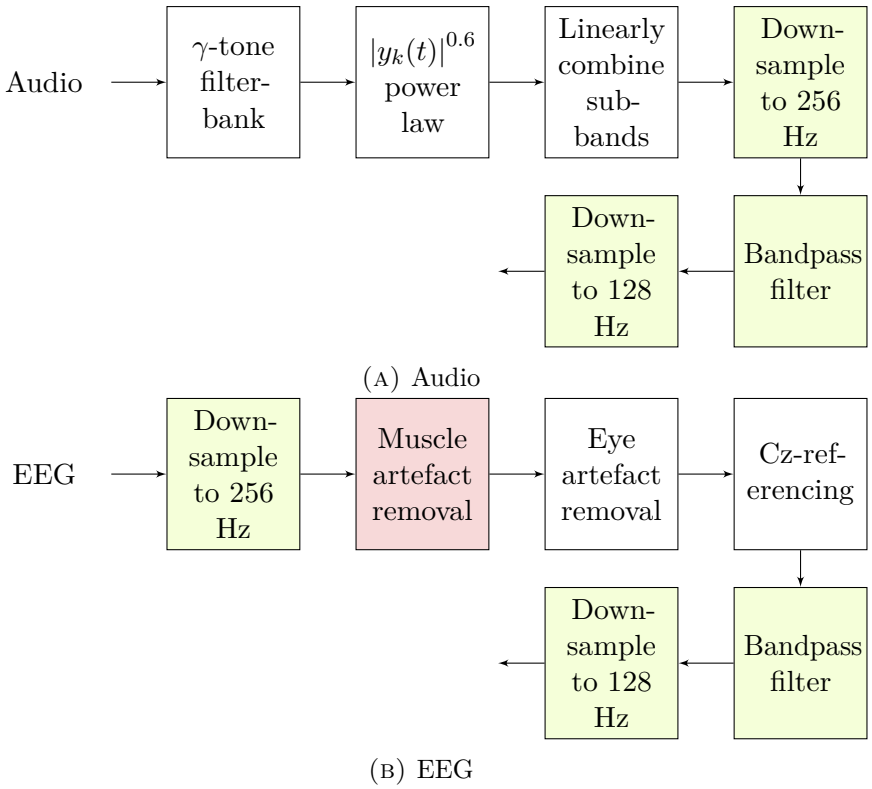
Regarding the auditory data stream, the envelope is extracted following the procedure of section 2.3. Thereafter a linear-phase, antialiasing filter is applied and the envelope is downsampled to 256 Hz. Following, the signal is bandpass filtered using a zero-phase Chebyshev type 2 filter that attains an 80 dB attenuation at the frequency 10% out of the passband. In this work, the passband is chosen to correspond to the following frequency ranges: $[0.5, 4]$ Hz (delta band), $[4, 8]$ Hz (theta band), $[8, 13]$ Hz (alpha band), $[13, 30]$ Hz (beta band) and $[0, 128]$ Hz (Unfiltered). Afterwards, another linear-phase, antialiasing filter is applied and the envelope is further downsampled to 128 Hz.

The EEG signal is first filtered with a linear-phase, antialiasing filter and downsampled to 256 Hz. In order to mitigate the effect of muscle and eye blink artefacts, the GEVD-MWF procedure of section 2.2 is adhered to, wherein an EEG time-lag of 3 samples, both positive and negative, is used. To this end, data are grouped per subject and the MWFs are computed in a subject-specific manner on the first 10 minutes of the associated EEG data. Succeeding, the EEG signals are referenced with respect to channel Cz by subtracting the Cz channel content from all other channels, and the signal is bandpass filtered using the same filter as for the audio preprocessing. Finally, after passing through another linear-phase, antialiasing filter, the signal is downsampled to 128 Hz.

[3]The code of the preprocessing framework has been generated based on the OMSI coding framework, developed by ExpORL KU Leuven, as available on 10/10/2021. This framework is not publicly accessible.

[4]For MWF based artefact removal, an MWF framework, developed by B. Somers as available on 10/10/2021 at https://github.com/exporl/mwf-artifact-removal, is utilised.

Periods of longer silence (> 0.25 s) are not considered in further design. To this end, segments are labelled as silence whenever the normalised amplitude of the audio signal is smaller than 0.05.



(A) Audio

(B) EEG

FIGURE 2.1: Joint EEG-audio preprocessing schematic. Except for the addition of the shaded block in red, this framework is implemented as proposed in [16]. Blocks, shaded in lime, are shared between the audio and EEG preprocessing.

2.5 Conclusion

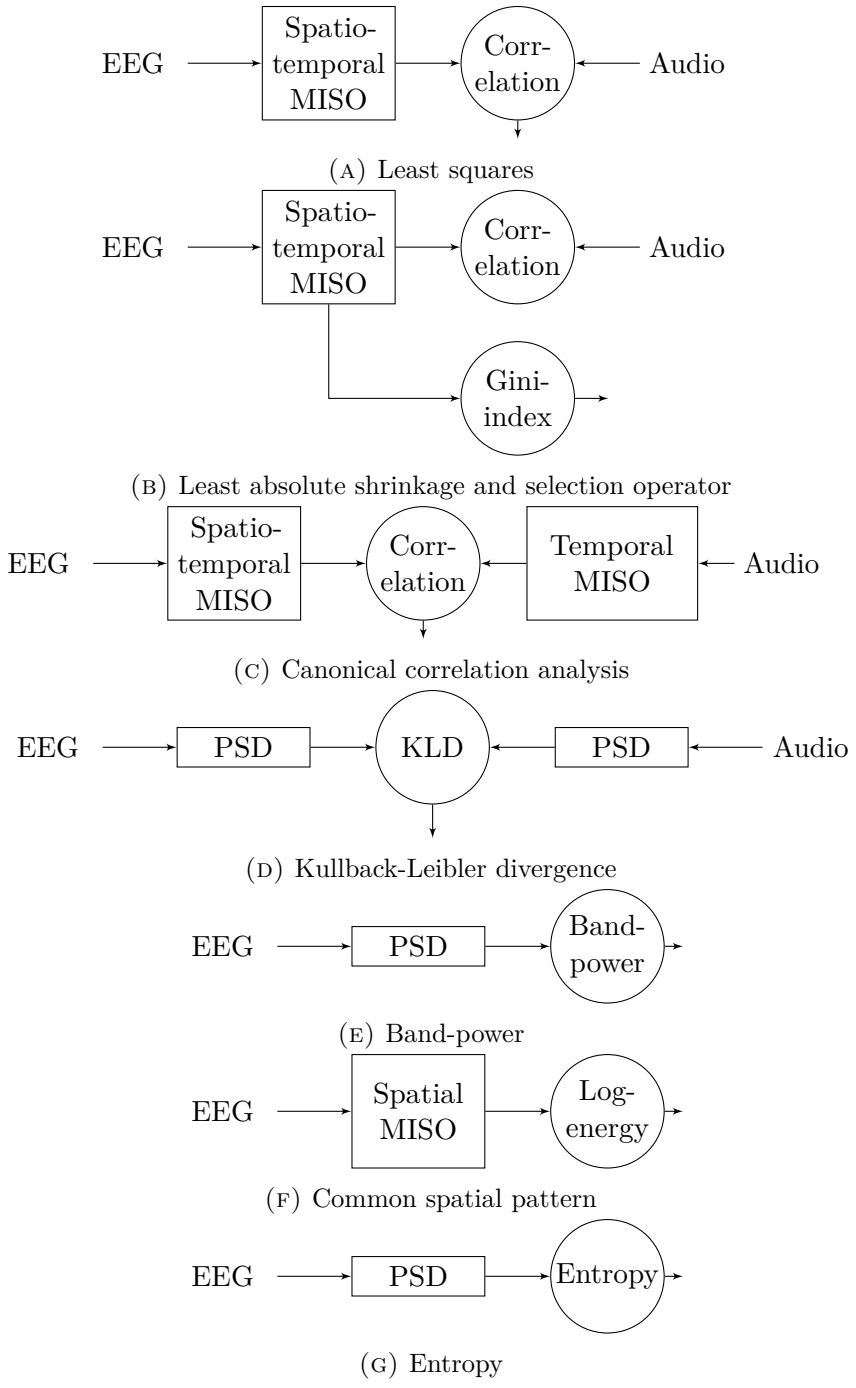
Before feature extraction, the auditory stream and EEG data pass through a preprocessing step as introduced in figure 2.1. This framework consists of an envelope extraction block since the EEG signal tracks the speech envelope. The framework also consists of an EEG artefact removal block to remove signals from interfering electrical sources. Furthermore, bandpass filters select 'interesting' frequency ranges to exploit the spectral EEG characteristics, and downsampling blocks are introduced to reduce computational cost at the feature extraction and classification level.

Chapter 3

Feature extraction

3.1 Introduction

This chapter gives a theoretical overview of features that show conceptual potential to discriminate between auditory attention and inattention states. This feature extraction phase forms the first step in the classic machine learning approach as illustrated in figure 1.4. Since some of these features are based on the concept of power spectral density (PSD), its definition and estimation are discussed first (section 3.2). Nevertheless, the main part of this chapter consists in introducing the features themselves, which are split into two groups: neural tracking based features, exploiting the envelope tracking property of the brain, and brain activity based features, exploiting solely EEG properties. The neural envelope tracking based features are described in section 3.3 and can be subdivided into a least squares (LS), a least absolute shrinkage and selection operator (LASSO), a canonical correlation analysis (CCA) and a novel Kullback-Leibler divergence (KLD) feature. Similarly, the brain activity based features are described in section 3.4 and can be subdivided into common spatial pattern (CSP), band-power (BP) and entropy. Figure 3.1 provides a schematic overview of each of these studied features and serves as visual support throughout this chapter. Moreover, both EEG and audio signals are assumed to be zero-mean. After this theoretical analysis, the estimation on real data and regularisation procedures are discussed respectively in sections 3.5 and 3.6. Indeed, in practise, only limited and possibly badly conditioned data are available. Thereafter, the visualisation procedure of the data-driven features is detailed in section 3.7. Finally, section 3.8 concludes this chapter.



(A) Least squares

(B) Least absolute shrinkage and selection operator

(C) Canonical correlation analysis

(D) Kullback-Leibler divergence

(E) Band-power

(F) Common spatial pattern

(G) Entropy

FIGURE 3.1: Schematic overview of the features.

3.2 Power spectral density estimation

The power spectral density (PSD) provides a way to assess the signal power per frequency bin, i.e., it provides a frequency domain representation of the signal power [21, 22]. The PSD $\mathcal{P}_U(f) \in \mathbb{R}$ of an infinite-length discrete time random process $u(t) \in \mathbb{R}$, with sampling frequency $f_s \in \mathbb{R}$ is defined as follows [21, 22]:

$$\begin{aligned} \mathcal{P}_U(f) &= \lim_{T\to\infty} \frac{1}{f_s^2 T} \mathbb{E}\left\{|u_T(f)|^2\right\}; \\ u_T(f) &= \sum_{t=0}^{T-1} u(t) e^{-j2\pi \frac{f}{f_s} t}. \end{aligned} \tag{3.1}$$

Herein, $u_T(f) \in \mathbb{R}$ corresponds to the discrete Fourier transform of the truncated random process consisting of $T \in \mathbb{N}_0$ samples. Furthermore, j represents the imaginary unit, and t and f respectively denote time and frequency. Since the power spectral density $\mathcal{P}_U(f)$ reveals the frequency distribution of the signal power, the band-power $P \in \mathbb{R}$ of $u(t)$ between frequencies $f_1 \in \mathbb{R}$ and $f_2 \in \mathbb{R}$ can be calculated as follows [21, 22]:

$$P = \sum_{f=f_1}^{f_2} \mathcal{P}_U(f). \tag{3.2}$$

However, equation 3.1 only holds for ideal, infinite-length random processes, whereas real life signals are finite length and noisy. Therefore, an estimation procedure needs to be implemented using finite length data. This estimation procedure furthermore need to cope with the following problems [21, 23]:

- **Bias:** Real life data are finite length, such that the time domain signals can be seen as a multiplication of an infinitely long signal by a rectangular function. Equivalently, in the frequency domain, this corresponds to a convolution between the Fourier transformed infinite-length signal and a sinc function, which distorts the frequency domain representation due to its main and side lobes. Indeed, the main lobe determines the spectral resolution ($\delta f \in \mathbb{R}$) of the frequency domain signal (narrowband bias) and the side lobes result in spectral leakage (broadband bias).
- **Variance:** Real life data are corrupted by noise, which results in noisy spectral estimations.

In order to limit the bias and variance in the spectral estimation, multitaper spectral analysis is utilised [21, 24]. In this method, $L \in \mathbb{N}_0$ separate PSD estimations are performed by multiplying the time domain signal, of length $T \in \mathbb{N}_0$, by L distinct window functions (tapers) and subsequently Fourier transforming the windowed time domain signals. This windowing smoothens the time domain signal, hence limiting the broadband bias. Thereafter, the L PSD estimations are averaged to limit the noise influence and to arrive at one estimation.

These tapers are chosen to be orthogonal to one another and to have optimal time-frequency concentration. This allows for uncorrelated PSD estimates and reduced

narrowband bias. In practise, a good choice for the number of tapers $L \in \mathbb{N}_0$ for a desired frequency resolution δf, corresponding to a signal of T samples, is computed according to the following heuristic [21, 23, 25]:

$$L \leq \lfloor f_s T \delta f \rfloor - 1, \tag{3.3}$$

wherein $f_s T \delta f \geq 2$. In what follows, $\bar{\mathcal{P}}_U(f)$ (note the overbar) will be used to denote the PSD **estimate** of a finite-length and noisy signal $u(t) \in \mathbb{R}$ using multitaper spectral analysis.

3.3 Neural envelope tracking based features

Firstly, this section expands the concept of neural envelope tracking and its implications. Thereafter, a description of four features, based on this envelope tracking concept, is given: least squares (LS), least absolute shrinkage and selection operator (LASSO), canonical correlation analysis (CCA) and the novel Kullback-Leibler divergence (KLD) based approach.

3.3.1 Neural envelope tracking

As noted in section 1.2, the brain tracks the envelope of speech signals [5, 8, 26, 27]. In fact, a higher level of auditory attention leads to improved envelope tracking [8, 16]. In [16], it has indeed been shown that subjects attentive to an auditory stream track the auditory envelope to a higher extent than subjects ignoring that auditory stream while focusing on a visual stream instead. In addition, in the AAD domain (see section 1.3), it has been proven that the degree of envelope tracking lies higher for attended than for unattended streams whenever multiple auditory streams are present at the same time [8, 11, 15, 28]. To characterise this envelope tracking, the time-lag between stimulus onset and EEG response, and the associated spectral EEG bands are leveraged [8, 16].

Regarding the time-lags, the EEG data track the auditory envelope approximately until 500 ms after stimulus onset [16, 27]. Indeed, due to causality, the EEG response always lags with respect to the auditory stimulus, explaining the positive delay. Further, both ignored and attended auditory streams seem to be tracked equally well in the early neural responses ($[0, 85]$ ms), while the attended stream seems to be tracked to a higher degree in the late neural responses ($[85, 500]$ ms) [27]. In [16] however, already a difference in the responses at lags $[0, 75]$ ms has been found between subjects attending or ignoring an auditory stream.

Regarding the spectral EEG distribution, the delta and theta bands appear to be most important in this envelope tracking [5, 8, 16]. These frequency bands indeed enclose the word and phrase rate ($[1, 4]$ Hz), and syllable rate ($[4, 8]$ Hz) [29, 30]. In addition, the high gamma band seems to encode differences in neural tracking [31]. Regarding the alpha band, debate exists in literature: In [32], envelope tracking has

been observed in the alpha band, although in [33], this effect is devoted to the alpha power and not to envelope tracking properties.

In order to reveal the level of envelope tracking, based on the time-lag and spectral properties, the EEG and/or audio signals get filtered by the feature extractor. There exist multiple approaches to do so and we will discuss four of them in the following subsections.

3.3.2 Least squares

This procedure consists in **relating the EEG data to the ground-truth auditory envelope**. As illustrated in figure 3.1a, this can be **achieved by introducing a spatiotemporal multiple input single output (MISO) filter (called decoder) at the EEG side, with the goal to reconstruct the auditory envelope from the EEG data**. In order to design this decoder, a least squares procedure is adhered to. To subsequently measure the degree of envelope tracking, the correlation between the reconstructed envelope and the ground truth auditory envelope is calculated. Higher correlation levels indeed denote higher attention levels: In [16], it has been shown that these correlations reach higher levels in subjects attending an auditory stream than in subjects ignoring this stream while focusing on a visual task. Furthermore, in [8], it has been shown that higher correlation levels are reached for attended streams than for unattended streams in a multiple speaker AAD scenario.

As in section 2.2 (equation 2.1), let $\underline{\mathbf{x}}(t) \in \mathbb{R}^{L_d C \times 1}$ represent a concatenation of EEG signals across C channels and across time-lags $L_{d-s}..L_{d-e}$ ($L_d = L_{d-e} - L_{d-s}$). Furthermore, let $y(t) \in \mathbb{R}$ represent the stimulus envelope at time t. Both EEG and audio are assumed to be zero-mean, in this and all following sections. The spatiotemporal MISO filter or decoder $\mathbf{d} \in \mathbb{R}^{L_d C \times 1}$ is then retrieved by minimising the mean squared error between the reconstructed ($\mathbf{d}^\top \underline{\mathbf{x}}(t) \in \mathbb{R}$) and original stimulus ($y(t) \in \mathbb{R}$) [8]:

$$\min_{\mathbf{d}} \quad \mathbb{E}\{(\mathbf{d}^\top \underline{\mathbf{x}}(t) - y(t))^2\}. \tag{3.4}$$

Setting the derivative of equation 3.4 to 0, yields the solution [8, 15]:

$$\begin{aligned} \mathbf{d} &= \mathbb{E}\{\underline{\mathbf{x}}(t)\underline{\mathbf{x}}(t)^\top\}^{-1}\mathbb{E}\{\underline{\mathbf{x}}(t)y(t)\} \\ &= R_{\underline{x}\underline{x}}^{-1}\mathbf{r}_{\underline{x}y}. \end{aligned} \tag{3.5}$$

In this equation, $R_{\underline{x}\underline{x}} = \mathbb{E}\{\underline{\mathbf{x}}(t)\underline{\mathbf{x}}(t)^\top\} \in \mathbb{R}^{L_d C \times L_d C}$ denotes the autocorrelation matrix of concatenated EEG channels across time-lags $L_{d-e}..L_{d-s}$ and $\mathbf{r}_{\underline{x}y} = \mathbb{E}\{\underline{\mathbf{x}}(t)y(t)\} \in \mathbb{R}^{L_d C \times 1}$ denotes the cross-correlation vector between the time-lagged and channel-concatenated EEG and the stimulus envelope.

From another viewpoint, these least squares decoders can also be retrieved (up to a scaling) by maximising the following benefit function [15]:

$$\begin{aligned} &\max_{\mathbf{d}} \quad \rho(\mathbf{d}^\top \underline{\mathbf{x}}(t), y(t)); \\ &\max_{\mathbf{d}} \quad \frac{\mathbb{E}\{\mathbf{d}^\top \underline{\mathbf{x}}(t) y(t)\}}{\sqrt{\mathbb{E}\{(\mathbf{d}^\top \underline{\mathbf{x}}(t))^2\}\mathbb{E}\{y(t)^2\}}}, \end{aligned} \tag{3.6}$$

i.e., the decoder $\mathbf{d}$ also represents the spatiotemporal filter that maximises the Pearson correlation coefficient $\rho(.)$ between the filtered EEG data and the stimulus envelope [15]. Therefore, this scale-invariant correlation measure can be used to quantify the degree of neural tracking. In what follows, we will utilise another flavour of correlation coefficient, namely the Spearman correlation coefficient [34], as measure of envelope tracking. This Spearman correlation coefficient corresponds to the Pearson correlation where the data are replaced by their rank values. As such, the Spearman correlation coefficient assesses monotonic relations, whereas the Pearson correlation assesses linear relations [34].

3.3.3 Least absolute shrinkage and selection operator

In addition to investigating the correlation using least squares decoders, in [35], it is proposed to **promote the sparsity of said decoders and to use this sparsity as a measure for envelope tracking**. They claim decoders, related to subjects tracking the neural envelope well, to consist of a mixture of large nonzero and zero weights. On the contrary, they claim decoders, related to subjects tracking the neural envelope less well, to have smaller, more randomly distributed weights. The main idea is thus to train decoders on data at inference time and to assess the sparsity of these trained decoders. A comparative study did however not find the same differentiating nature [28].

It is important to note that the studies of [28] and [35] took place in an AAD setting. The introduction of this sparsity intuition into our domain of differentiating between auditory (in)attention is novel. In addition, the combination of least squares decoders and the Gini-index sparsity measure, introduced below ,is novel as well.

In addition, it is important to understand the inherent difference in setup between this sparsity feature and the LS feature. In the LS setting, a decoder is calculated on a **training** set. At inference, this decoder is applied to the EEG data and the correlation between the reconstructed and ground truth envelope delivers the actual feature value. On the contrary, in this sparsity setting, a decoder is calculated at **inference** time and the sparsity of this decoder corresponds to the actual feature value. Figure 3.1b yields a visual representation of this sparsity feature.

The sparsity induction can be achieved by adding a sparsity-promoting least absolute shrinkage and selection operator (LASSO) term with weight $\gamma \in \mathbb{R}_0^+$ to the LS cost

function of equation 3.4 [36]:

$$\min_{\mathbf{d}} \quad \mathbb{E}\{(\mathbf{d}^\top \underline{\mathbf{x}}(t) - y(t))^2\} + \gamma ||\mathbf{d}||_1. \tag{3.7}$$

This problem is well-defined since it can be shown to correspond to a convex problem (more specifically a convex quadratically constrained quadratic problem (QCQP)), wherein each minimum corresponds to a global minimum [36, 37].

To assess the degree of sparsity, an appropriate measure is required. There can be formulated 6 properties, which are ideally satisfied by such a sparsity measure [38, 39, 40]:

- **Robin hood:** Subtracting a value from a large coefficient and redistributing it over the other coefficients decreases sparsity.
- **Scaling:** Scaling does not influence sparsity.
- **Rising Tide:** Adding the same value to each coefficient decreases sparsity.
- **Cloning:** The sparsity of a vector $\mathbf{d}$ or any positive number of concatenations of $\mathbf{d}$ is the same.
- **Bill Gates:** Increasing the value of the highest coefficient increases sparsity.
- **Babies:** Extending a vector with zero values increases sparsity.

The Gini-index is a sparsity measure that satisfies all of these properties, although it can only be applied if all elements of $\mathbf{d}$ are nonnegative [38]. Therefore, the Gini-index can be applied to the least squares decoder, wherein all elements are replaced by their absolute value. This Gini-index can subsequently be formulated as follows [38, 39, 41][1]:

$$S(\mathbf{d}) = 1 - 2\sum_{n=1}^{L_d C} \left(\frac{|d(n)|}{||\,|\mathbf{d}|\,||_1} \left(\frac{L_d C - n + \frac{1}{2}}{L_d C} \right) \right). \tag{3.8}$$

Note that this metric allows to compare the sparsity of vectors of different lengths since the vector length does not influence this sparsity measure [38]. This property allows, e.g., to still use the Gini-index whenever multiple sensors are malfunctioning.

3.3.4 Canonical correlation analysis

As with the LS feature, canonical correlation analysis (CCA) tries to assess the level of envelope tracking by measuring the correlation between (a filtered version of) the EEG signal and (a filtered version of) the auditory envelope, as illustrated in figure 3.1c [11, 42, 43]. To this end, again, a spatiotemporal MISO filter (decoder) is inserted at the EEG side. However, CCA also introduces a temporal MISO filter (encoder) at the auditory side. In other words, **both EEG and audio are transformed to a latent space, where a correlation coefficient measures the level of envelope tracking** [11, 28, 42]. As before, both EEG and audio are assumed to be zero-mean. CCA has been seen to lead to higher accuracies

[1] $|\mathbf{d}|$ represents the elementwise absolute value operation of vector $\mathbf{d}$.

in the AAD domain [11, 42], but has not yet been introduced in our domain of interest.

Mathematically speaking, CCA finds a decoder $\mathbf{d} \in \mathbb{R}^{L_d C \times 1}$ and an encoder $\mathbf{e} \in \mathbb{R}^{L_e \times 1}$ jointly such that the filtered EEG and filtered audio data are maximally correlated according to the Pearson correlation coefficient $\rho(.)$, i.e.[11, 42, 43]:

$$\max_{\mathbf{d},\mathbf{e}} \quad \rho(\mathbf{d}^\top \underline{\mathbf{x}}(t), \mathbf{e}^\top \underline{\mathbf{y}}(t)). \tag{3.9}$$

Herein, $\underline{\mathbf{x}}(t)$ represents the channel-concatenated and time-lagged EEG, as defined in equation 2.1, and $\underline{\mathbf{y}}(t) \in \mathbb{R}^{L_e \times 1}$ represents the time-lagged stimulus envelope corresponding to time-lags $L_{e-s}..L_{e-e}$ ($L_e = L_{e-e} - L_{e-s}$):

$$\underline{\mathbf{y}}(t) = \begin{bmatrix} y(t + L_{e-s}) & \dots & y(t + L_{e-e} - 1) \end{bmatrix}^\top. \tag{3.10}$$

Utilising the definition of the Pearson correlation coefficient and substituting correlation matrices $R_{\underline{xx}} = \mathbb{E}\{\underline{\mathbf{x}}(t)\underline{\mathbf{x}}(t)^\top\} \in \mathbb{R}^{L_d C \times L_d C}$, $R_{\underline{yy}} = \mathbb{E}\{\underline{\mathbf{y}}(t)\underline{\mathbf{y}}(t)^\top\} \in \mathbb{R}^{L_e \times L_e}$ and $R_{\underline{xy}} = \mathbb{E}\{\underline{\mathbf{x}}(t)\underline{\mathbf{y}}(t)^\top\} \in \mathbb{R}^{L_d C \times L_e}$, equation 3.9 can be rewritten as follows [11, 42, 43]:

$$\max_{\mathbf{d},\mathbf{e}} \quad \frac{\mathbb{E}\{(\mathbf{d}^\top \underline{\mathbf{x}}(t)\underline{\mathbf{y}}(t)^\top \mathbf{e})\}}{\sqrt{\mathbb{E}\{\mathbf{d}^\top \underline{\mathbf{x}}(t)\underline{\mathbf{x}}(t)^\top \mathbf{d}\}}\sqrt{\mathbb{E}\{\mathbf{e}^\top \underline{\mathbf{y}}(t)\underline{\mathbf{y}}(t)^\top \mathbf{e}\}}}; \tag{3.11}$$

$$\max_{\mathbf{d},\mathbf{e}} \quad \frac{\mathbf{d}^\top R_{\underline{xy}} \mathbf{e}}{\sqrt{\mathbf{d}^\top R_{\underline{xx}} \mathbf{d}}\sqrt{\mathbf{e}^\top R_{\underline{yy}} \mathbf{e}}}. \tag{3.12}$$

Since $\mathbf{d}$ and $\mathbf{e}$ are only defined up to a nonzero scalar, $\mathbf{d}^\top R_{\underline{xx}} \mathbf{d}$ and $\mathbf{e}^\top R_{\underline{yy}} \mathbf{e}$ are constrained to 1, i.e., the variances of the projected data are constrained to 1 [11, 42]:

$$\begin{aligned} \max_{\mathbf{d},\mathbf{e}} \quad & \mathbf{d}^\top R_{\underline{xy}} \mathbf{e} \\ \text{s.t.} \quad & \mathbf{d}^\top R_{\underline{xx}} \mathbf{d} = 1 \\ & \mathbf{e}^\top R_{\underline{yy}} \mathbf{e} = 1. \end{aligned} \tag{3.13}$$

By writing down the necessary conditions for first order optimality (so-called Karush-Kuhn-Tucker (KKT) conditions) [44], it can be shown that equation 3.13 (a QCQP) boils down to a generalised eigenvalue decomposition (GEVD) [11, 42, 43]:

$$\begin{bmatrix} 0_{L_d C \times L_d C} & R_{\underline{xy}} \\ R_{\underline{xy}}^\top & 0_{L_e \times L_e} \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \mathbf{e} \end{bmatrix} = \lambda \begin{bmatrix} R_{\underline{xx}} & 0_{L_d C \times L_e} \\ 0_{L_e \times L_d C} & R_{\underline{yy}} \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \mathbf{e} \end{bmatrix}. \tag{3.14}$$

The resulting decoder $\mathbf{d}$ and encoder $\mathbf{e}$ match the eigenvectors corresponding to the largest eigenvalue [11, 42, 43]. This solution can be extended to include additional decoders $\mathbf{d}_j \in \mathbb{R}^{L_d C \times 1}$ and encoders $\mathbf{e}_j \in \mathbb{R}^{L_e \times 1}$, $j = 1..P$ with P maximally equal to $min(rank(R_{\underline{xx}}), rank(R_{\underline{yy}})) \in \mathbb{R}$. To this end, these additional decoders also maximise the correlation of the projected data, yet under additional uncorrelatedness constraints with respect to previous projections, i.e.,

$$\begin{aligned} \mathbf{d}_j R_{\underline{xx}} \mathbf{d}_i &= 0, \qquad i \neq j; \\ \mathbf{e}_j R_{\underline{yy}} \mathbf{e}_i &= 0, \qquad i \neq j, \ \forall i, j = 1..P. \end{aligned} \tag{3.15}$$

Defining $D = \begin{bmatrix} \mathbf{d}_1 & \dots \mathbf{d}_P \end{bmatrix} \in \mathbb{R}^{L_d C \times P}$ and $E = \begin{bmatrix} \mathbf{e}_1 & \dots \mathbf{e}_P \end{bmatrix} \in \mathbb{R}^{L_e \times P}$, these additional constraints lead to the following optimisation problem [11, 42, 43]:

$$\begin{aligned} \max_{D,E} \quad & Tr\{D^\top R_{\underline{xy}} E\} \\ \text{s.t.} \quad & D^\top R_{\underline{xx}} D = \mathbb{I}_{P \times P} \\ & E^\top R_{\underline{yy}} E = \mathbb{I}_{P \times P}. \end{aligned} \tag{3.16}$$

Indeed, given the similarity in structure between equations 3.9 and 3.16, it can be seen that equation 3.16 serves as a generalisation of equation 3.9. The solution of this equation 3.16 thus also boils down to a generalised eigenvalue decomposition, wherein $\Lambda \in \mathbb{R}^{2P \times 2P}$ contains the eigenvalues on its diagonal:

$$\begin{bmatrix} 0_{L_d C \times L_d C} & R_{\underline{xy}} \\ R_{\underline{xy}}^\top & 0_{L_e \times L_e} \end{bmatrix} \begin{bmatrix} D \\ E \end{bmatrix} = \begin{bmatrix} R_{\underline{xx}} & 0_{L_d C \times L_e} \\ 0_{L_e \times L_d C} & R_{\underline{yy}} \end{bmatrix} \begin{bmatrix} D \\ E \end{bmatrix} \Lambda. \tag{3.17}$$

The decoders and encoders correspond to the eigenvectors related with the subsequent P largest eigenvalues. In other words, by choosing decoders and encoders corresponding to the subsequent largest eigenvectors, a multidimensional latent space can be constructed of maximal dimension $P = min(rank(R_{\underline{xx}}), rank(R_{\underline{yy}}))$ [11, 42, 43].

Finally, note that the LS approach is actually a special case of the CCA approach [11, 42]. Both paradigms indeed optimise the Pearson correlation coefficient, yet, the CCA approach also filters the auditory envelope. Nevertheless, whenever the time-lag of the temporal filter is chosen to be 0, this filter reduces to a scaling. Since the filters are only defined up to some nonzero scaling, due to the scale-invariance of the Pearson correlation coefficient, the CCA approach is in that case fully equivalent to the LS approach.

3.3.5 Kullback-Leibler divergence

Finally, envelope tracking can be studied in the frequency domain, based on the concept of the PSD (see section 3.2). A normalised PSD sums to 1 and only contains values between 0 and 1, such that normalised PSDs satisfy all the properties of probability density functions (PDF) [45, 46]. This insight provides a gateway for utilising measures from probability theory to quantify the level of neural envelope tracking. To this end, we leverage a similarity measure between distributions, the Kullback-Leibler divergence (KLD), as a neural envelope tracking feature. **This approach plus-minus consists in a binwise comparison of the power distribution in the EEG signal and the audio envelope.**

Although the KLD has already been used to relate PSDs to one another (e.g. [47, 48]), it has not yet been used to relate EEG to auditory envelopes. Furthermore, we present late and early fusion methods, novel as such to our best of knowledge, to fuse the C EEG channels into one KLD based feature.

Firstly, the concept of the KLD as a measure from probability theory is expanded. Thereafter, the KLD calculation between one EEG channel and the auditory envelope is detailed. Finally, both late and early fusion approaches are introduced to relate multiple EEG channels to the auditory envelope.

Probabilistic view

The Kullback-Leibler divergence (KLD) ≥ 0 between two discrete probability density functions $f_X(x)$ and $g_X(x)$, each consisting of $P \in \mathbb{N}_0$ probability values, is defined as follows [49, 50]:

$$KLD\left(f_X(x)\,||g_X(x)\right) = \sum_{p=1}^{P} f_X(x_p)\, log_2\left(\frac{f_X(x_p)}{g_X(x_p)}\right), \tag{3.18}$$

and evaluates the difference between these probability distributions $f_X(x)$ and $g_X(x)$. Larger values denote less similarity in distribution. However, note that this metric does not correspond to a distance-metric since $KLD\left(f_X(x)\,||g_X(x)\right) \neq KLD\left(g_X(x)\,||f_X(x)\right)$. Indeed, in $KLD\left(f_X(x)\,||g_X(x)\right)$, whenever $\exists p : f_X(x_p) = 0$, there is no contribution to the KLD. Yet, whenever $\exists p : g_X(x_p) = 0$ the KLD becomes equal to ∞. The metric is thus typically utilised when a distribution $g_X(x)$ is known and one would like to quantify how close another distribution $f_X(x)$ is to that ground truth distribution $g_X(x)$.

Relating one EEG channel to the stimulus envelope

These KLD properties are leveraged in a neural tracking setting. After normalising the PSD estimate of the channel-c EEG signal $\bar{\mathcal{P}}_{X_c} \in \mathbb{R}$ and the PSD estimate of the auditory envelope $\bar{\mathcal{P}}_Y \in \mathbb{R}$ to their respective band-power, the KLD can be readily computed as follows [2]:

$$KLD_c = \sum_{f=f_1}^{f_2} \bar{\mathcal{P}}_{X_c,n}(f) log_2\left(\frac{\bar{\mathcal{P}}_{X_c,n}(f)}{\bar{\mathcal{P}}_{Y,n}(f)}\right). \tag{3.19}$$

Therefore, as illustrated in figure 3.1d, the Kullback-Leibler divergence (KLD) can be utilised as a novel neural envelope tracking feature. This metric roughly consists in a binwise comparison between the PSD of the EEG and the PSD of the audio data. This feature thus comes with an additional neurological interpretation: Due to the binwise comparison, the KLD inherently assumes that the frequency content of the

[2] Subscript n denotes normalisation of the PSD with respect to the band-power.

auditory envelope is stored at the same frequencies in the EEG data. This requirement is not straightforward due to nonlinear processing in the brain, although in [51] it has been shown that the brain locks to the modulation frequency of sinusoidal waves. Therefore, this KLD measure shows the potential of providing additional insights into the mechanisms underlying the neural envelope tracking.

This procedure can readily be repeated for each EEG channel, yielding C KLD values. Nonetheless, in order to transform them into one interpretable KLD metric, we present two approaches: a late fusion and an early fusion one.

Late fusion

Regarding the late fusion approach, **the C KLD values are first computed and only thereafter combined**. This can, e.g., be achieved by a classifier, by using equal weights or by solving a custom optimisation problem. The latter approach is opted in this work to optimally control the interpretation of the resulting, averaged value. To this end, the KLD values are linearly combined according to the following cost function:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \mathbf{a}^\top \mathbf{k} + \gamma ||\mathbf{a}||_2^2 \\ \text{s.t.} \quad & \mathbf{0}_{C\times 1} \leq \mathbf{a} \\ & \mathbf{1}_{C\times 1}^\top \mathbf{a} = 1. \end{aligned} \tag{3.20}$$

Herein, $\mathbf{k} = [KLD_1 \dots KLD_C]^\top \in \mathbb{R}^{C\times 1}$ represents a stacking of the C KLD values, $\mathbf{a} \in \mathbb{R}^{C\times 1}$ represents the weights attributed to each channel and $\gamma \in \mathbb{R}$ denotes a nonzero weight to control the relative importance of each of the two cost terms. In other words, the linear combination is sought that minimises the overall combined KLD. In this light, the first constraint expresses that all elements in the weighting need to be larger than 0. This constraint guarantees that the overall feature is larger than 0, enabling interpretation as a KLD. The second constraint expresses that all elements must sum to 1, such that $\mathbf{a}$ can in fact be interpreted as a weighted sum. The term $\gamma||\mathbf{a}||_2^2$ moreover is added to the cost function, to favour solutions with smaller weights. There are two reasons for this: Firstly, in general, adding this term leads to better generalising solutions [52]. Secondly, without this term, the solution would just boil down to selecting the channel with minimal KLD, rendering the optimisation problem useless. After solving equation 3.20 at training time, the weighting vector $\mathbf{a}$ can readily be applied to PSD estimates at inference time.

Early fusion

Regarding the early fusion approach, **the PSD estimates of the C EEG channels are first combined and only thereafter the KLD is computed**. To this end, the normalised PSDs are first weighted by weighting vector $\mathbf{a} \in \mathbb{R}^{C\times 1}$, prior to

calculating the KLD, according to the following cost function:

$$\begin{aligned} \min_{\mathbf{a}} \quad & KLD\left(\frac{\sum_{c=1}^{C} a(c)\bar{\mathcal{P}}_{X_c,n}(f)}{\sum_{f=f_1}^{f_2}\sum_{c=1}^{C} a(c)\bar{\mathcal{P}}_{X_c,n}(f)} || \bar{\mathcal{P}}_{Y,n}(f)\right) \\ \text{s.t.} \quad & \mathbf{a} \geq \mathbf{0}_{C\times 1} \\ & ||\mathbf{a}||_2 = 1. \end{aligned} \tag{3.21}$$

The normalising denominator in the EEG weighting of the cost function and the first constraint ($\mathbf{a} \geq \mathbf{0}_{C\times 1}$) are added to enforce the PDF properties to the averaged PSD. Indeed, the combination of both makes sure that the resulting averaged PSD sums to 1 and that all its elements remain between 0 and 1. The normalisation factor nevertheless makes the cost function scale-invariant. To combat this scale-invariance, the second constraint ($||\mathbf{a}||_2 = 1$) is added to retrieve the solution with unit L2-norm. This constraint yields the interpretation of optimising the cost function over a hyper-sphere. A drawback of equation 3.21 is its non-convexity, which means that there may be local critical points present, which do not correspond to a global minimum. The interested reader can find a proof of this statement in appendix A.2.

3.4 Brain activity based features

In addition to relating the EEG signal to the auditory envelope, there are techniques to exploit properties of the EEG signal itself. Firstly, a brief overview of work related to brain activity (features) in the broad field of attention is given. Next, three brain activity based features are described: band-power (BP), common spatial pattern (CSP) and entropy.

3.4.1 Brain activity

In literature, the frontal and parietal-occipital regions (see figure 3.2) are associated with the concept of attention [53, 54, 55]. However these studies are not entirely the same as our domain of interest: In [53], visual sustained and working memory have been investigated and in [55], localisation, mathematical, spatial, verbal and multisource interference tasks have been studied. Nevertheless, the goal of this study is not focused on the broad concept of attention, but on attention towards an auditory stream and more specifically, discriminating this auditory attention state from auditory inattention states. **Furthermore, subjects being inattentive to an auditory stream, could still be attentive, yet to another stimulus such as a visual or a sensory one.** In [54], the difference between auditory and visual attention has been studied: Subjects were displayed a signal with the instruction to attend either to a visual or an auditory unit of a subsequent compound audiovisual task. Specifically, these subjects needed to attend either a beep sound or a flash in a beep-flash combination. They found that subjects attending the beep had higher

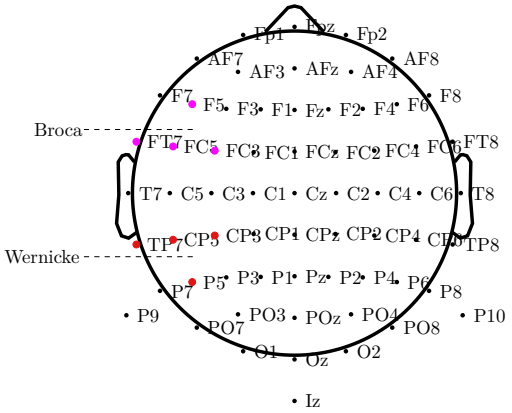


FIGURE 3.2: Broca's and Wernicke's areas mapped to a 64-channel EEG cap. The convention of [57] is adhered to.

power in the parietal-occipital region. However, our setup is still more general, using real life stimuli. Therefore, results will not necessarily coincide.

Moreover, two brain regions, Broca's and Wernicke's areas, are related to language processing and are correspondingly of interest [56, 57]. To relate these brain areas to specific EEG sensors we use the same mapping as proposed in [57] and as shown in figure 3.2. However, this mapping is not strict since the EEG sensors further away may still measure signals originating from these areas, albeit to a lesser extent. In addition, EEG has a poor spatial resolution, as indicated in section 1.2.

Frequency-wise, the alpha and beta bands are classically related to the concept of attention [12, 58, 59]. However, again, these studies mainly treat the general concept of attention, whereas our setup also needs to permit differentiation between different attention cases. Referring back to the study in [54], they have found differences in the alpha band-power, yet again, our setup is tailored towards real life stimuli and hence somewhat different.

3.4.2 Band-power

It is expected that different brain regions are less or more active whenever a subject is being (in)attentive to an auditory stream. This activity can, e.g., be characterised by the power, as illustrated in figure 3.1e. In [54], this approach has been applied in differentiating between attention to an auditory stream and attention to a visual stream. Nevertheless, whereas in [54] beep sound were utilised, we use real-life stimuli.

This band-power (BP) measures the power $P_{BP_c} \in \mathbb{R}$ of a channel-c EEG signal $X_c(t) \in \mathbb{R}$ in a certain frequency range $[f_1, f_2] \in \mathbb{R}$, e.g., by using its PSD estimate

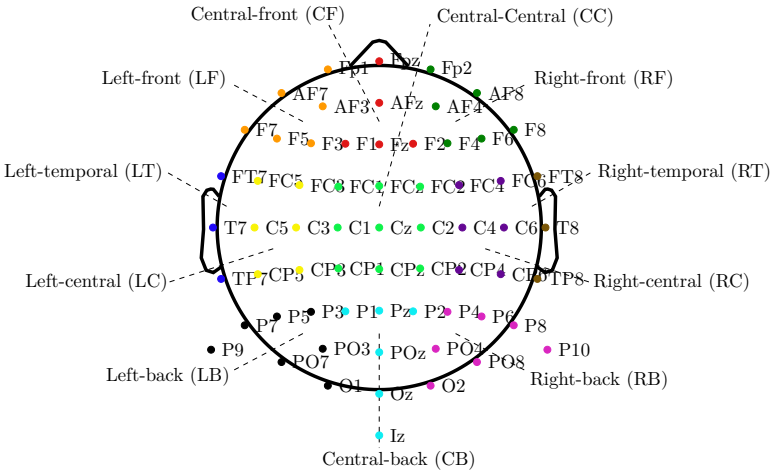


FIGURE 3.3: 64-channel EEG cap: division of the sensors into groups.

$\bar{\mathcal{P}}_{X_c}(f) \in \mathbb{R}$:

$$P_{BP_c} = \sum_{f=f_1}^{f_2} \bar{\mathcal{P}}_{X_c}(f). \tag{3.22}$$

This approach can be taken for every channel separately. Nonetheless, for visualisation and interpretation purposes, the band-power is aggregated into regions, as shown in figure 3.3. This is, the sensors are divided into a front-back and left-right grid, resulting in 11 groups. However, again, this subdivision does not result in a perfect separation due to neighbouring sensors picking up similar signals and due to the EEG's poor spatial resolution, as detailed in section 1.2.

3.4.3 Common spatial pattern

As illustrated in figure 3.1f, common spatial pattern (CSP) [60, 61] tries to differentiate between two classes, solely based on the EEG data, by applying a spatial MISO filter to the EEG data. The underlying idea is that the two classes (here attention and inattention to the auditory stimulus) result in different brain activity, which can be exploited by applying a spatial filter.

CSP has already been used in other brain computer interface (BCI) applications, e.g., in [62] and [60], and in [12] CSP was introduced into the AAD domain, where they found CSP to outperform LS- and CCA-approaches. However, it has not yet been applied in our domain of interest.

To exploit this spatial information, **CSP maximises the output energy of one**

class K_1, while minimising the output energy of another class K_2[3], i.e., CSP defines a filter $\mathbf{w} \in \mathbb{R}^{C\times 1}$ according to the following benefit function:

$$\max_{\mathbf{w}} \quad \frac{\mathbf{w}^\top R_1 \mathbf{w}}{\mathbf{w}^\top R_2 \mathbf{w}}. \tag{3.23}$$

Herein, the subscript 1 denotes instances belonging to class K_1 and, similarly, the subscript 2 denotes instances belonging to class K_2. As such, the channel-concatenated normalised EEG signals of classes K_1 and K_2 are respectively denoted by $\mathbf{x}_1(t) \in \mathbb{R}^{C\times 1}$ and $\mathbf{x}_2(t) \in \mathbb{R}^{C\times 1}$. The autocorrelation matrices $R_i \in \mathbb{R}^{C\times C}$, $i = 1, 2$ are subsequently defined as follows [60]:

$$\begin{aligned} R_1 &= \mathbb{E}\{\mathbf{x}_1(t)\mathbf{x}_1(t)^\top\}; \\ R_2 &= \mathbb{E}\{\mathbf{x}_2(t)\mathbf{x}_2(t)^\top\}. \end{aligned} \tag{3.24}$$

Equation 3.23 can be converted into a constrained optimisation problem (a QCQP), since $\mathbf{w}$ is only defined up to a nonzero scalar. In fact, constraining the denominator to 1 results in the following problem [60]:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^\top R_1 \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^\top R_2 \mathbf{w} = 1. \end{aligned} \tag{3.25}$$

Using the conditions for first order optimality (KKT-conditions), this problem subsequently boils down to solving the following equation [60]:

$$R_1 \mathbf{w} = \lambda R_2 \mathbf{w}. \tag{3.26}$$

Thus, the filter $\mathbf{w}$ can be retrieved as the solution of a generalised eigenvalue problem. In fact, $\mathbf{w}$ corresponds to the eigenvector with respect to the **largest** eigenvalue $\lambda \in \mathbb{R}$ [60].

Of course, one could also design a filter corresponding to the inverse optimisation problem of equation 3.23:

$$\max_{\mathbf{w}} \quad \frac{\mathbf{w}^\top R_2 \mathbf{w}}{\mathbf{w}^\top R_1 \mathbf{w}}. \tag{3.27}$$

Nonetheless, it turns out that this problem boils down to the same generalised eigenvalue decomposition of equation 3.26, with the only difference that now the eigenvector corresponding to the **smallest** eigenvalue needs to be retained [60]. Both problems therefore reduce to the same joint diagonalisation problem, although still two separate filters are obtained.

Common spatial pattern can also be extended to generate $P \in \mathbb{N}_0 \leq C$ filters by imposing additional uncorrelatedness constraints on the projections. For example, for

[3] Note the connection with the band-power feature, although the CSP feature firstly applies a filtering stage.

the filters that maximise the output energy of class K_1 $\left(W = \begin{bmatrix}\mathbf{w}_1 & \dots & \mathbf{w}_P\end{bmatrix} \in \mathbb{R}^{C\times P}\right)$, the following problem is solved [60]:

$$\begin{aligned} \max_{W} \quad & Tr\{W^\top R_1 W\} \\ \text{s.t.} \quad & W^\top R_2 W = \mathbb{I}_{P\times P}. \end{aligned} \tag{3.28}$$

Similar as before, this optimisation problem corresponds to a generalised eigenvalue decomposition, i.e., a joint diagonalisation of R_1 and R_2 with $\Lambda_1, \Lambda_2 \in \mathbb{R}^{P\times P}$ diagonal matrices containing the eigenvalues [60]:

$$\begin{aligned} W^\top R_1 W &= \Lambda_1; \\ W^\top R_2 W &= \Lambda_2. \end{aligned} \tag{3.29}$$

Similarly, the inverse optimisation problem of 3.28 corresponds to the same joint diagonalisation.

At inference, after the CSP-filtering step by P filters for maximising the output energy of class K_1 ($\mathbf{w}_{1,p} \in \mathbb{R}^{C\times 1}, p = 1..P$) and P filters for maximising the output energy of class K_2 ($\mathbf{w}_{2,p} \in \mathbb{R}^{C\times 1}, p = 1..P$), features $\mathbf{f} \in \mathbb{R}^{2P\times 1}$ are often extracted by computing the log-energy of the channel-concatenated EEG signal $\mathbf{x}(t) \in \mathbb{R}^{C\times 1}$ over some time window $T \in \mathbb{N}_0$ [12, 60]:

$$\begin{aligned} \mathbf{f} &= \begin{bmatrix} log\left(\sigma^2_{1,1}\right) & \dots & log\left(\sigma^2_{1,P}\right) & log\left(\sigma^2_{2,1}\right) & \dots & log\left(\sigma^2_{2,P}\right)\end{bmatrix}^\top; \\ \sigma^2_{i,p} &= \sum_{t=0}^{T-1} \left(\mathbf{w}^\top_{i,p}\mathbf{x}(t)\right)^2, \quad i = 1, 2,\ p = 1..P. \end{aligned} \tag{3.30}$$

3.4.4 Entropy

The entropy characterises the level of uncertainty in a distribution. Therefore, high attention levels connect to high entropy levels (and thus high uncertainty in distribution). The reverse holds true for low attention levels [6, 63]. In [6], the spectral entropy in alpha and beta band, and in frontal and parietal-occipital sensors has been used as a metric to discriminate between active and passive segments within the same EEG stream. In addition, in [63], the entropy allowed to distinguish between subjects attentive to a flashing screen from passive subjects. However, again, these setups are different from our application domain.

Reusing the idea of normalised PSDs satisfying all properties of a PDF, the spectral entropy $S_c\left[f_1, f_2\right] \in \mathbb{R}$ in a certain frequency band $[f_1, f_2] \in \mathbb{R}$ can be calculated using the PSD estimate $\bar{\mathcal{P}}_{X_c}(f) \in \mathbb{R}$ of a channel-c EEG signal $X_c(t) \in \mathbb{R}$ (see also figure 3.1g) [45, 46] [4]:

$$S_c\left[f_1, f_2\right] = -\sum_{f=f_1}^{f_2} \bar{\mathcal{P}}_{X_c,n}(f) \log_2\left(\bar{\mathcal{P}}_{X_c,n}\left(f\right)\right). \tag{3.31}$$

[4] Subscript n denotes normalisation with respect to the band-power

This entropy is an appropriate measure for the uncertainty in distribution since it satisfies the following properties; the entropy is [50]:

- non-negative;
- zero when the outcome is deterministic;
- maximal when all $P \in \mathbb{N}_0$ outcomes are equiprobable, i.e., when the distribution is uniform the entropy equals $log_2(P)$;
- smaller for a uniform distribution consisting of P outcomes than for a uniform distribution consisting of $Q \in \mathbb{N}_0$ outcomes, whenever $P < Q$;
- a continuous function of the probabilities.

As with the BP feature, the entropy values are aggregated over the groups as shown in figure 3.3 for visualisation and interpretation purposes.

3.5 Estimation on real data

The LS, LASSO, CCA and CSP feature extraction methodologies utilise the auto- and/or cross-correlation matrices/vectors of the EEG and/or audio input data. In sections 3.3 and 3.4, these matrices were defined using the expected value operator $\mathbb{E}\{.\}$. However, in practise, EEG and audio signals are finite length and only one observable is available, such that the expected value operator cannot be calculated as is. Therefore, stationarity [50] and ergodicity [50] are assumed and the auto- and cross-correlation matrices and vectors are estimated using sample averages (e.g. see [8] and [15]). Assuming EEG and audio signals of length $T \in \mathbb{N}_0$, the estimate $\bar{R}_{\underline{x}\underline{y}} \in \mathbb{R}^{L_d C \times L_e C}$ (note the overbar) of cross-correlation matrix $R_{\underline{x}\underline{y}}$ can be calculated as follows:

$$\bar{R}_{\underline{x}\underline{y}} = \frac{1}{T} \sum_{t=0}^{T-1} \underline{\mathbf{x}}(t) \underline{\mathbf{y}}(t)^\top. \tag{3.32}$$

Other correlation matrices/vectors can be estimated in a similar fashion. Nonetheless, note that these stationarity and ergodicity assumptions are rarely met in practise, e.g., due to changing brain activity [13].

3.6 Regularisation

In the estimation of the correlation matrices on this finite data, mainly two problems occur: Firstly, (too) large models tend to result in a poor fit on unseen validation data, despite almost fully fitting the training data; A concept known as overfitting [52, 64]. Secondly, the correlation matrices may be badly conditioned, introducing numerical errors in the design of the feature extractors [65].

In order to avoid this overfitting and cope with these badly conditioned correlation matrices, regularisation is introduced. This regularisation denotes a variety of methods, with the shared goal of decreasing the degrees of freedom in the system, in order to reduce the model's complexity. In this way, variance is diminished at the expense of introducing bias in the design [52]. Popular approaches to achieve

this regularisation, are adding a weighted norm term to the cost functions of the feature extractors and/or applying a prior dimensionality reduction on the EEG data [15, 42, 52]. In fact, there are underlying relations between these two methods, as discussed in [52]. In this text, we will detail this norm-weighted regularisation and the accompanying closed-form solution and we will also detail principal component analysis (PCA) as a dimensionality reduction approach [52, 66, 67]. For the sake of clarity, but without loss of generality, we will tailor this section to the least squares cost function of equation 3.4.

Norm-weight regularisation

In norm-weighted regularisation, a weighted norm-term is added to the feature extractor's cost function in order to drive the feature extractor weights to 0 and as such reduce the complexity of the model. Regarding the LS cost function, this looks as follows ($\gamma \in \mathbb{R}^+$) [52]:

$$\min_{\mathbf{d}} \quad E\{(\mathbf{d}^\top \underline{\mathbf{x}}(t) - y(t))^2\} + \gamma ||\Gamma \mathbf{d}||_p^2. \tag{3.33}$$

Herein, $\Gamma \in \mathbb{R}^{L_d C \times L_d C}$ can be chosen freely to incorporate prior model beliefs. A popular choice is nevertheless to equate Γ to the identity matrix $\mathbb{I}_{L_d C \times L_d C}$ [16, 65]. Moreover, the notation $||\Gamma \mathbf{d}||_p = \sum_{l=1}^{L_d C} |\Gamma_{l,:}\mathbf{d}|^p$, $p \geq 1$ represents the p-norm of the vector $\Gamma \mathbf{d} \in \mathbb{R}^{L_d C \times 1}$ [5].

One popular option is to choose $p = 1$, which is referred to as LASSO- or L1-regularisation [35, 52]. As already detailed in section 3.3, this results in a convex LS cost function (namely a convex QCQP) that induces sparsity [36]. Another popular choice is to set $p = 2$. This method is referred to as Tikhonov-regularisation, L2-regularisation or ridge regression [52, 68]. Setting the gradient of equation 3.33 to 0, yields the solution:

$$\mathbf{d} = (R_{\underline{xx}} + \gamma \Gamma^\top \Gamma)^{-1} \mathbf{r}_{\underline{xy}}. \tag{3.34}$$

Regularising least squares using L2-regularisation thus corresponds to adding a matrix with weight γ to the autocorrelation matrix. Similarly, CCA can be regularised by adding a weighted matrix to both autocorrelation matrices $R_{\underline{xx}}$ and $R_{\underline{yy}}$ [69]. Regarding CSP, adding a L2-penalty in the denominator of the cost function (equations 3.23 and 3.27), results in adding a weighted matrix to that denominator autocorrelation matrix [70].

Ledoit-Wolf closed-form for L2 correlation matrix regularisation

Concerning the L2-regularisation term γ, one could tune this parameter using a parameter sweep or, alternatively, one could find an estimated correlation matrix

[5] M_{kl} represents the element of matrix M at row k and column l. Herein, $k = :$ and $l = :$ refer respectively to selecting all elements in that respective row/column.

$\bar{R}^*_{\underline{xx}} = \alpha \mathbb{I}_{L_d C \times L_d C} + \beta \bar{R}_{\underline{xx}} \in \mathbb{R}^{L_d C \times L_d C}$ using weighting parameter $\alpha \in \mathbb{R}$ for the identity matrix $\mathbb{I}_{L_d C \times L_d C}$ and weighting parameter $\beta \in \mathbb{R}$ for the sample correlation matrix $\bar{R}_{\underline{xx}} \in \mathbb{R}^{L_d C \times L_d C}$ in an analytic fashion, optimal with respect to some optimisation problem [65].

Indeed, in [65], such analytic values α and β are calculated, optimal with respect to the Frobenius norm ($||.||_F$). This is, the resulting correlation matrix estimate $\bar{R}^*_{\underline{xx}}$ asymptotically converges to the ground-truth correlation matrix $R_{\underline{xx}} \in \mathbb{R}^{L_d C \times L_d C}$, with respect to the Frobenius norm, as the number of observations and variables go to infinity. Assume $N \in \mathbb{N}_0$ observations of a time-lagged EEG vector $\underline{\mathbf{x}}(n) \in \mathbb{R}^{L_d C \times 1}$, $n = 1..N$ are available. The correlation matrix estimate $\bar{R}^*_{\underline{xx}}$ can be found by first solving the following optimisation problem [65]:

$$\begin{aligned} \min_{\alpha,\beta} \quad & E\{||R^*_{\underline{xx}} - R_{\underline{xx}}||_F^2\} \\ s.t. \quad & R^*_{\underline{xx}} = \alpha \mathbb{I}_{L_d C \times L_d C} + \beta \bar{R}_{\underline{xx}}, \end{aligned} \tag{3.35}$$

and thereafter replacing the variables in the solution $R^*_{\underline{xx}} \in \mathbb{R}^{L_d C \times L_d C}$ by consistent estimators to arrive at $\bar{R}^*_{\underline{xx}}$. In this way, the strategy to calculate $\bar{R}^*_{\underline{xx}}$ is as follows [65]:

$$\begin{aligned} \bar{R}^*_{\underline{xx}} &= \frac{b^2}{d^2} m \mathbb{I}_{L_d C \times L_d C} + \frac{a^2}{d^2} \bar{R}_{\underline{xx}}; \\ m &= \frac{Tr\left(\bar{R}_{\underline{xx}}\right)}{L_d C}; \\ d^2 &= \frac{Tr\left(\left(\bar{R}_{\underline{xx}} - m \mathbb{I}_{L_d C \times L_d C}\right)\left(\bar{R}_{\underline{xx}} - m \mathbb{I}_{L_d C \times L_d C}\right)^\top\right)}{L_d C}; \\ c^2 &= \frac{1}{N^2} \sum_{n=1}^{N} \frac{Tr\left(\left(\underline{\mathbf{x}}(n)\underline{\mathbf{x}}(n)^\top - \bar{R}_{\underline{xx}}\right)\left(\underline{\mathbf{x}}(n)\underline{\mathbf{x}}(n)^\top - \bar{R}_{\underline{xx}}\right)^\top\right)}{L_d C}; \\ b^2 &= min(c^2, d^2); \\ a^2 &= d^2 - b^2. \end{aligned} \tag{3.36}$$

Dimensionality reduction: principal component analysis

Another way to regularise the cost functions of feature extractors is to perform a dimensionality reduction on the input EEG data. Indeed, by reducing the number of input channels, the models become less complex and hence regularisation is achieved. Principal component analysis (PCA) is a popular method to perform such a dimensionality reduction [52, 66, 67]. In this PCA method, the zero-mean data are first projected onto the vector (called principal component) that captures most of the variance of said data. In addition, subsequent $P \in \mathbb{N}_0 \leq C$ principal components also try to maximise the variance of the projected data, however, with the additional

constraints of being orthogonal to previous principal components. Mathematically speaking, this can be expressed as follows [52]:

$$\begin{aligned} \min_{W} \quad & Tr\left(W^\top R_{xx} W\right) \\ \text{s.t.} \quad & W^\top W = \mathbb{I}_{P\times P}. \end{aligned} \tag{3.37}$$

Herein, $W \in \mathbb{R}^{L_d C \times P}$ holds the $P \leq C$ principal components in its columns. In fact, this problem boils down to finding the best rank-P approximation of the EEG data [67].

After applying W to the EEG data, filter design can be performed identically as described in sections 3.3 and 3.4, albeit with a reduced amount of EEG channels.

3.7 Visualisation: a neurological interpretation

The LS, LASSO, CCA, CSP and KLD feature extraction design approaches, as described in sections 3.3 and 3.4, have in common that filters are designed and applied to the EEG data. It is important to note that the accompanying **filter weights do not allow for a neurological interpretation as such** [71]. To clarify this, assume the EEG data satisfy the following data model [71]:

$$\mathbf{x}(t) = A\mathbf{s}(t) + \mathbf{n}(t), \tag{3.38}$$

wherein $\mathbf{s}(t) \in \mathbb{R}^{P\times 1}$ and $\mathbf{n}(t) \in \mathbb{R}^{C\times 1}$ represent respectively the $P \in \mathbb{N}_0 \leq C$ EEG source signals of interest and the additive noise signals. Furthermore, $A \in \mathbb{R}^{C\times P}$ denotes the mixing matrix from source to sensor. **The columns of A can be neurologically interpreted**, i.e., the columns represent the transfer functions from the source to the sensor (up to a nonzero scaling), whereas this same property does not hold for the designed filters $W \in \mathbb{R}^{C\times P}$. Nevertheless, in [71], a transformation is proposed to transform the spatial filters W into a mixing matrix estimate $\bar{A} \in \mathbb{R}^{C\times P}$ as follows:

$$\bar{A} = \bar{R}_{xx} W \bar{R}_{ss}^{-1}, \tag{3.39}$$

wherein $\bar{R}_{xx} \in \mathbb{R}^{C\times C}$ denotes the measured EEG autocorrelation matrix and $\bar{R}_{ss} \in \mathbb{R}^{P\times P}$ denotes the source signal EEG correlation matrix. Moreover, when W corresponds to a square matrix, it can be shown that this equation reduces to [71]:

$$\bar{A} = \left(W^\top\right)^{-1}. \tag{3.40}$$

Note however that this result still needs to be interpreted with caution since the resulting mixing matrix might not be optimally regularised [72].

3.8 Conclusion

In this section both existing (LS, LASSO and CCA) and novel (KLD) neural envelope tracking based features were described (section 3.3). In addition, existing brain activity based methods (CSP, BP and entropy) were detailed (section 3.4). Except for the LS estimator, the application domain with real life stimuli is novel for all feature extractors, to our best of knowledge. Furthermore, due to finite data set length, the parameters of these methods need to be estimated on a finite-length training set (section 3.5). This might introduce overfitting effects and badly conditioned correlation matrices, which may be tackled using regularisation methods such as norm-weight regularisation and dimensionality reduction mechanisms (section 3.6). Finally, it is important to note that one cannot interpret the filter weights of the described methods as such, although transformations exist to convert the filter to neurologically interpretable variants (section 3.7).

Chapter 4

Classification

4.1 Introduction

This chapter treats the classification step in the classic machine learning approach, as illustrated in figure 1.4. Indeed, after a feature extraction phase, one leverages these features to decide whether the subject is being attentive to an auditory stream. Mathematically speaking, this corresponds to a binary classification problem.

These binary classifiers are based on the concept of a discriminant function, as will be detailed in section 4.2. Thereafter, a popular classifier is described in section 4.3, namely the linear discriminant analysis (LDA) classifier [12, 52, 64]. After presenting this LDA method, section 4.4 treats its regularisation and describes an input normalisation strategy. Finally, section 4.5 concludes this chapter.

4.2 Discriminant function

A discriminant function characterises the decision boundary between classes [52, 64]. Let $\underline{\mathbf{f}} = \begin{bmatrix} 1 & \mathbf{f}^\top \end{bmatrix}^\top \in \mathbb{R}^{F+1\times 1}$ be constructed as the stacking of feature vector $\mathbf{f} \in \mathbb{R}^{F\times 1}$ of dimension $F \in \mathbb{N}_0$ and a bias term 1. For the binary classification case, this stacked feature vector $\underline{\mathbf{f}}$ is subsequently assigned to class K_1 whenever the discriminant function $y(\underline{\mathbf{f}}) \in \mathbb{R}$ is positive and to class K_2 whenever the discriminant function is negative:

$$\begin{aligned} y(\underline{\mathbf{f}}) &= h(\boldsymbol{\theta}, \underline{\mathbf{f}}); \\ sign(y(\underline{\mathbf{f}})) &= +1, \qquad \underline{\mathbf{f}} \in K_1; \\ sign(y(\underline{\mathbf{f}})) &= -1, \qquad \underline{\mathbf{f}} \in K_2. \end{aligned} \tag{4.1}$$

Therefore, in its most general form, the discriminant function appears as a function $h(.)$ with $U \in \mathbb{N}_0$ parameters, stacked in parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{U\times 1}$. Nonetheless, a common choice of this discriminant function is a linear form [52, 64]:

$$y(\mathbf{f}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{f}. \tag{4.2}$$

Herein, $\boldsymbol{\beta} \in \mathbb{R}^{F\times 1}$ denotes the slope and $\beta_0 \in \mathbb{R}$ the bias term, i.e., this discriminant function tries to linearly separate the feature space. Therefore, it readily follows that: $\theta = \begin{bmatrix}\beta_0 & \boldsymbol{\beta}^\top\end{bmatrix}^\top$ and $h(.)$ equals a unit transform. Note that the separation by this linear discriminant in fact poses no hard constraints on the classification. Indeed, prior to classification, the features can be nonlinearly transformed, such that the transformed feature space is linearly separable [52, 64].

From a probabilistic perspective, the discriminant function can be related to the posterior class distributions (probability of a class given a feature vector) [52]. In order to reveal this relation, let $\mathbf{X} \in \mathbb{R}^{F\times 1}$ and $K \in \mathbb{R}$ represent respectively the random variable for the feature vector and the random variable for the class label. The posterior class distributions can now be expanded using Bayes rule and the definition of the logistic function $\left(\sigma(a) = \frac{1}{1+e^{-a}}\right)$ [52, 64]:

$$\begin{aligned} &P(K = K_1|\mathbf{X} = \mathbf{f}) \\ &\qquad = \frac{P(\mathbf{X} = \mathbf{f}|K = K_1)P(K = K_1)}{P(\mathbf{X} = \mathbf{f}|K = K_1)P(K = K_1) + P(\mathbf{X} = \mathbf{f}|K = K_2)P(K = K_2)}; \\ &\qquad = \frac{1}{1 + exp(-a)}; \\ &\qquad = \sigma(a). \end{aligned} \tag{4.3}$$

Herein, $a \in \mathbb{R}$ can be equated to $y(\mathbf{f})$ of equation 4.1, i.e., one could assign the feature vector to the class with the largest posterior class probability. Indeed, a takes the form of the log-ratio of both class posteriors ($P(K = K_2|\mathbf{X} = \mathbf{f}) = 1 - P(K = K_1|\mathbf{X} = \mathbf{f})$) [52, 64]:

$$a = log\left(\frac{P(K = K_1|\mathbf{X} = \mathbf{f})}{P(K = K_2|\mathbf{X} = \mathbf{f})}\right). \tag{4.4}$$

4.3 Linear discriminant analysis

Linear discriminant analysis (LDA) leverages this linear discriminant model by constructing a linear discriminant function according to equation 4.2 [52]. This is achieved by utilising the expression of equation 4.4, under the following assumptions [52]:

- The class conditional distribution ($P(\mathbf{X} = \mathbf{f}|K = K_i),\ i = 1, 2$) is multivariate normal distributed.
- The class conditional distributions for both classes have equal covariance matrices.
- Both classes have equal class prior distributions ($P(K = K_1) = P(K = K_2)$).

Let $\boldsymbol{\mu}_i \in \mathbb{R}^{F\times 1}$ $(i = 1, 2)$ denote the mean of the class conditional distributions and $\Sigma \in \mathbb{R}^{F\times F}$ the accompanying covariance matrix. It can subsequently be shown that,

under these assumptions, the slope $\boldsymbol{\beta}$ and bias term β_0 take the following form [52]:

$$\begin{aligned} \boldsymbol{\beta} &= \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1); \\ \beta_0 &= \frac{-1}{2}\boldsymbol{\beta}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \end{aligned} \tag{4.5}$$

Note that, in practise, the means and covariance matrix have to be estimated using sample averages, corresponding to the procedure of section 3.5.

4.4 Regularisation and normalisation

The curse of dimensionality [52, 73] denotes the effect that a space is sampled more sparsely if its dimensionality increases while the amount of data points is kept the same. This manifests itself, e.g., in the training of classifiers. Indeed, if the feature vector dimensionality is increased and the training data samples the space sparsely, there are many ways to separate the training data. However, that does not necessarily mean that this separation generalises well. The classification performance might subsequently go down when adding additional features to the feature vector, although this expanded feature space inherently encodes more/the same information as the unextended feature vector.

Thus, as in section 3.6, the classification methods may suffer from overfitting. Again, one can add additional norm-weighted terms to combat this problem. In the case of the LDA classifier, this boils down to adding a weighted matrix to the covariance matrix, which is consistent with the results of section 3.6 [52].

Further, features may operate at different scales [74]. This interferes with the norm-weighted regularisation, where penalties are applied according to the magnitude of the feature values. Therefore, the features are normalised using a z-score in advance of being fed into the classifier [74]:

$$z(i) = \frac{f(i) - \mu_f(i)}{\Sigma_f(i,i)}, \qquad \forall i = 1..F. \tag{4.6}$$

Herein, $\mu_f \in \mathbb{R}^{F\times 1}$ and $\Sigma_f \in \mathbb{R}^{F\times F}$ respectively correspond to the mean and covariance matrix of the training set features.

4.5 Conclusion

Classification methods, such as the LDA classifier, are based on the concept of a discriminant function, which mathematically defines the decision regions in a feature space. To this end, the LDA classifier defines a linear decision boundary. As with the feature extraction methods, the classification methods may overfit the input data, such that regularisation methods are desired. However, due to this regularisation, input normalisation techniques are required as well. Indeed, different features might operate at different scales, which is alleviated by this normalisation.

Chapter 5

Conversion to unsupervised algorithms

5.1 Introduction

This chapter discusses the conversion of the classic machine learning approach, as illustrated in figure 1.4, towards unsupervised algorithms. Indeed, the supervised framework poses practical problems: Firstly, the feature extractor-classifier design is typically performed in a subject-specific manner, requiring more resources than subject-independent ones [75]. Secondly, the stationarity assumptions, on which the designs are based, rarely hold in practise, e.g., due to EEG cap displacements and changes in EEG background activity [13]. Introducing unsupervised algorithms mitigates the above problems in order to arrive at practical implementations.

Formally, the initial training set is referred to as the source domain, consisting of an L-dimensional source data space $\boldsymbol{\mathcal{X}}_{\mathcal{S}} \in \mathbb{R}^L$ and accompanying source label space $\mathcal{Y}_{\mathcal{S}} \in \mathbb{R}$ [76]. Designs trained on this source domain cannot be readily applied to other subjects or to the same subject in a nonstationary environment, i.e., to the target domain, due to a mismatch between both domains. Similarly to the source domain, this target domain consists of an L-dimensional target data space $\boldsymbol{\mathcal{X}}_{\mathcal{T}} \in \mathbb{R}^L$ and accompanying target label space $\mathcal{Y}_{\mathcal{T}} \in \mathbb{R}$. The mismatch between both domains is characterised by a shift in distribution, i.e., $P(Y_{\mathcal{S}}, \mathbf{X}_{\mathcal{S}}) \neq P(Y_{\mathcal{T}}, \mathbf{X}_{\mathcal{T}})$, $Y_{\mathcal{S}} \in \mathcal{Y}_{\mathcal{S}}$, $Y_{\mathcal{T}} \in \mathcal{Y}_{\mathcal{T}}$, $\mathbf{X}_{\mathcal{S}} \in \boldsymbol{\mathcal{X}}_{\mathcal{S}}$, $\mathbf{X}_{\mathcal{T}} \in \boldsymbol{\mathcal{X}}_{\mathcal{T}}$ [76].

To combat this distribution shift problem, we look into the conversion of the **feature extractors** of chapter 3 into unsupervised ones. The goal of these unsupervised algorithms is to exploit labelled source domain data and unlabelled target domain data to tailor the source trained feature extractors to the target domain[1]. This

[1]Since we leverage methods from the field of computer vision, we adhere to the definition of unsupervised methods prevalent in that field [77, 78]. Algorithms are referred to as unsupervised if solely unlabelled data is assumed for the target domain. No restrictions are imposed on the availability of labels in the source domain data. Nevertheless, one might argue that methods requiring

procedure allows to arrive at practical algorithms since labelled source domain datasets are readily available and unlabelled target domain data can be collected at inference time. **More specifically, we will tailor our research towards the least squares feature extractor.** Although most discussed procedures are more generally applicable to any data-driven feature extractor, using least squares as a running example improves readability. Moreover, the least squares feature extractor is popular in other domains, such as in the AAD field [8, 15, 28].

State-of-the-art unsupervised least squares design involves an iterative procedure and serves as a benchmark for subsequent methods [75]. To this end, the iterative procedure is briefly detailed in section 5.2. Nevertheless, we take a different approach to tackle the domain shift, namely using unsupervised domain adaptation (DA) [78]. Section 5.3 introduces this DA concept and discusses accompanying procedures in sections 5.3.1 and 5.3.2. Finally, section 5.4 concludes the chapter.

5.2 Iterative least squares

In [75], an unsupervised least squares algorithm is proposed by iteratively self-labelling the data and updating the decoder correspondingly. We will thus subsequently refer to this procedure as iterative least squares (ILS). Although this algorithm is originally tailored to the AAD domain, it can be applied to our domain of interest, as pseudocodewise summarised in algorithm 1[2].

Referring to equation 3.5, a least squares decoder consists of the combination of an EEG autocorrelation matrix and an EEG-audio cross-correlation vector. Regarding the EEG autocorrelation matrix, source and target domain information are fused by making a weighted average (with weight $\delta \in \mathbb{R}$) between the source attention EEG autocorrelation matrix[3] $R_{\underline{x}_{\mathcal{S},A}\underline{x}_{\mathcal{S},A}} \in \mathbb{R}^{L_dC \times L_dC}$ and the target EEG autocorrelation matrix $R_{\underline{x}_{\mathcal{T}}\underline{x}_{\mathcal{T}}} \in \mathbb{R}^{L_dC \times L_dC}$. This autocorrelation matrix remains fixed throughout the remainder of the algorithm. Regarding the EEG-audio cross-correlation vector, the source attention cross-correlation vector $\mathbf{r}_{\underline{x}_{\mathcal{S},A}y_{\mathcal{S},A}} \in \mathbb{R}^{L_dC \times 1}$ is used as an initial estimate and is updated throughout the algorithm. The resulting autocorrelation matrix and cross-correlation vector are combined, according to equation 3.5, to yield the initial decoder $\mathbf{d} \in \mathbb{R}^{L_dC}$.

What follows is an iterative procedure, wherein the current decoder is applied to the target data to retrieve an estimate of the target attention segments. Using

the availability of labelled source domain data are effectively semi-supervised methods [79].

[2]Nevertheless, (nontrivial) adaptations are required to adjust this iterative procedure to our domain of interest: In the AAD domain, the EEG data of only 1 subject is present, such that the EEG data remain fixed throughout the algorithm, whereas this assumption does not hold in our domain of interest. We, however, justify our approach since we are only interested in utilising this method as a benchmark. A study of the convergence effects of this change in EEG data utilisation is beyond the scope of this text.

[3]Subscript A refers to 'attention' and subscript AE refers to 'attention estimation'.

these estimated labels, an estimate of the target attention cross-correlation vector $\mathbf{r}_{\underline{x}_{\mathcal{T},AE}y_{\mathcal{T},AE}} \in \mathbb{R}^{L_d C \times 1}$ is computed. Thereafter, the decoder is updated by combining the fixed autocorrelation matrix and and an updated cross-correlation vector, calculated as the weighted average between $\mathbf{r}_{\underline{x}_{\mathcal{T},AE}y_{\mathcal{T},AE}}$ and $\mathbf{r}_{\underline{x}_{\mathcal{S},A}y_{\mathcal{S},A}}$ (with weight $\eta \in \mathbb{R}$). This procedure is repeated until convergence.

Algorithm 1 Iterative least squares (ILS)

$R_{\underline{x}\underline{x}} \leftarrow \delta R_{\underline{x}_{\mathcal{S},A}\underline{x}_{\mathcal{S},A}} + (1-\delta) R_{\underline{x}_{\mathcal{T}}\underline{x}_{\mathcal{T}}}$;
$\mathbf{r}_{\underline{x}y} \leftarrow \mathbf{r}_{\underline{x}_{\mathcal{S},A}y_{\mathcal{S},A}}$;
$\mathbf{d} \leftarrow R_{\underline{x}\underline{x}}^{-1}\mathbf{r}_{\underline{x}y}$;
for $i = 1$; $i \leq i_{max}$; i++ **do**
    Apply decoder $\mathbf{d}$ to the target data and retrieve
    the target attention estimation (AE) segments;
    $\mathbf{r}_{\underline{x}y} \leftarrow \eta \mathbf{r}_{\underline{x}_{\mathcal{S},A}y_{\mathcal{S},A}} + (1-\eta)\mathbf{r}_{\underline{x}_{\mathcal{T},AE}y_{\mathcal{T},AE}}$;
    $\mathbf{d} \leftarrow R_{\underline{x}\underline{x}}^{-1}\mathbf{r}_{\underline{x}y}$;
end for

5.3 Domain adaptation

Domain adaptation (DA) takes another approach to combat the distribution shift between source and target domain by leveraging the information of the unlabelled target data in the calculation of the feature extractors. More specifically, we will look into unsupervised DA methodologies using feature alignment [76]. This feature alignment DA theoretically enforces a joint embedding of source and target domain features, effectively mitigating the need for unsupervised classification methods. Indeed, at inference time, one may perform DA using the training set and newly encountered target domain data to create said joint embedding. Since the features of both domains now 'live in the same space', one can issue the adapted feature extractor on the source data and train a regular, supervised classifier on the adapted source domain features. Thereafter, the feature extractor and classifier can theoretically be applied to the target domain data as is.

We will consider two classes of DA methodologies. **A first class, EEG basis transformation DA, consists in finding a joint basis between the source and target EEG.** In this work, we will focus on PCA and CCA based transformations, as is detailed in section 5.3.1. The general idea of CCA and PCA based DA is not novel (e.g. [80, 81, 82]), nevertheless, the application to EEG data in order to relate EEG to audio data is novel to our best of knowledge. **A second class, discriminator DA, consists in introducing a source-target classifier in the feature design, i.e., the features are designed such that it is hard for the source-target classifier to discriminate between source and target domain.** Whereas discriminator domain adaptation has already been explored in other domains (e.g.

[78]), the specific expression, as detailed in section 5.3.2, is novel as such to our best of knowledge.

5.3.1 EEG basis transformation

The general goal of the EEG basis transformation is to find a common basis between source and target EEG. This allows to train a decoder on the transformed attention source EEG in a supervised manner, and to apply this decoder to the target data. In order to achieve this joint EEG basis, CCA and PCA are leveraged. Since different approaches are possible, we detail four EEG basis transformation flavours: canonical correlation analysis DA (CCA-DA), principal component analysis DA (PCA-DA), subspace alignment DA (SA-DA) and target principal component analysis DA (TPCA-DA). Whereas these CCA and PCA mechanisms have been extensively used in the field of domain adaptation (e.g. [80, 81, 83]), their application in our domain of interest is novel to our best of knowledge.

Formally, let $X_{\mathcal{S}}, X_{\mathcal{T}} \in \mathbb{R}^{T \times C}$ contain $T \in \mathbb{N}_0$ samples of respectively source and target EEG data (both attention and inattention) and let $\mathbf{y}_{\mathcal{S}}, \mathbf{y}_{\mathcal{T}} \in \mathbb{R}^{T \times 1}$ contain the accompanying audio envelopes. The goal of the EEG basis transformation is to design $P \in \mathbb{N}_0 \leq C$ spatial source filters $W_{\mathcal{S}} \in \mathbb{R}^{C \times P}$ and P spatial target filters $W_{\mathcal{T}} \in \mathbb{R}^{C \times P}$, to transform source and target EEG data as follows: $X_{\mathcal{S}} W_{\mathcal{S}} \in \mathbb{R}^{T \times P}$ and $X_{\mathcal{T}} W_{\mathcal{T}} \in \mathbb{R}^{T \times P}$. Subsequently, the labelled attention portion of the transformed source data, $X_{\mathcal{S},A} W_{\mathcal{S}} \in \mathbb{R}^{T_2 \times C}$, $T_2 \in \mathbb{N}_0 \leq T$ with corresponding envelope $\mathbf{y}_{\mathcal{S},A} \in \mathbb{R}^{T_2 \times 1}$, can be utilised to train a least squares decoder $\mathbf{d} \in \mathbb{R}^{L_d C \times 1}$. This decoder $\mathbf{d}$ can thereafter be applied to the transformed target data $X_{\mathcal{T}} W_{\mathcal{T}}$ to deduce the (in)attention state.

Note that these procedures inherently regularise the problem, since only $P \leq C$ vectors are used to construct a basis, such that a dimensionality reduction is effectively realised.

CCA- and PCA-based domain adaptation can be performed as follows:

Canonical correlation analysis

Canonical correlation analysis domain adaptation (CCA-DA) applies the CCA procedure of section 7.2.5 to source and target EEG, i.e., by maximising the correlations between source and target EEG, a joint EEG space is realised. A pseudocode representation of this CCA-DA procedure can be found in algorithm 2[4].

[4] In algorithm 2, *LS* denotes least squares decoder design according to the procedure of section 3.3.2, and *CCA* and *PCA* respectively denote canonical correlation analysis and principal component analysis design according to the procedures of sections 3.3.4 and 3.6 .

Algorithm 2 CCA and PCA based domain adaptation

Collect and preprocess EEG and audio;

if CCA-DA **then**
    $W_{\mathcal{S}}, W_{\mathcal{T}} \leftarrow CCA(X_{\mathcal{S}}, X_{\mathcal{T}})$;
else if PCA-DA **then**
    $W_{\mathcal{S}} = W_{\mathcal{T}} \leftarrow PCA([X_{\mathcal{S}}^{\top}\ X_{\mathcal{T}}^{\top}]^{\top})$;
else if SA-DA **then**
    $W'_{\mathcal{S}} \leftarrow PCA(X_{\mathcal{S}})$; $W_{\mathcal{T}} \leftarrow PCA(X_{\mathcal{T}})$;
    $W_{\mathcal{S}} \leftarrow W'_{\mathcal{S}} W'^{\top}_{\mathcal{S}} W_{\mathcal{T}}$;
else if TPCA-DA **then**
    $W_{\mathcal{S}} = W_{\mathcal{T}} \leftarrow PCA(X_{\mathcal{T}})$;
end if

$\mathbf{d} \leftarrow LS(X_{\mathcal{S},A} W_{\mathcal{S}}, \mathbf{y}_{\mathcal{S},A})$;
correlation$(X_{\mathcal{T}} W_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}}, \mathbf{d})$;

Principal component analysis

Principal component analysis domain adaptation (PCA-DA) takes a similar route as CCA-DA, except for exchanging the CCA method with a PCA procedure on the concatenation of source and target EEG data, as illustrated in algorithm 2.

Subspace alignment

Subspace alignment domain adaption (SA-DA) [82] also takes a PCA-based approach, nevertheless, the subspace construction differs from PCA-DA. In SA-DA, firstly, PCA is applied separately to source EEG data $X_{\mathcal{S}}$ and target EEG data $X_{\mathcal{T}}$, to respectively obtain filters $W_{\mathcal{S}}$ and $W_{\mathcal{T}}$. Secondly, the source filters $W_{\mathcal{S}}$ are linearly transformed to be close to $W_{\mathcal{T}}$ with respect to the Frobenius norm. In other words, a matrix $\bar{M} \in \mathbb{R}^{P \times P}$ of coefficients is designed by minimising the Bregman matrix divergence [82]:

$$\bar{M} = \underset{M}{\operatorname{argmin}} \, ||W_{\mathcal{S}} M - W_{\mathcal{T}}||_F^2. \tag{5.1}$$

The solution of this expression equals $\bar{M} = W_{\mathcal{S}}^{\top} W_{\mathcal{T}}$ [82]. Using this matrix $\bar{M}$, the source domain basis is thus modified to $W_{\mathcal{S}} \bar{M} = W_{\mathcal{S}}^{\top} W_{\mathcal{T}} \in \mathbb{R}^{C \times P}$. Algorithmically, this procedure can be summarised as shown in algorithm 2.

Target principal component analysis

The idea of subspace alignment can now be expanded: Instead of transforming the source EEG to an intermediate basis close to the target domain, the target domain PCA basis could be utilised as is. The principal components resulting from the application of the PCA procedure to the target domain EEG can indeed be utilised

as a basis for the source domain EEG. This procedure is pseudocodewise illustrated in algorithm 2.

5.3.2 Discriminator

The discriminator domain adaptation (D-DA) methodology is of another type than previous EEG transformation methods. **The main idea is to include a source-target classifier in the feature extractor design. By maximising the loss of said classifier, in conjunction with regular feature design (chapter 3), a joint feature embedding is theoretically realised.** In other words, if it is hard for a classifier to discriminate between the source domain features and the target domain features, features of both domains should be more similar. Thus, by concurrently pushing for this similarity between source and target domain, and optimising a regular feature extractor cost function on the labelled source domain data, domain adaptation is achieved.

The idea of such a discriminator is presented in [78], where this principle is applied to neural networks. However, due to the constraints in the data-driven feature extractors of chapter 3, the mathematical procedure, as presented here, will prove to differ from [78], effectively creating a new methodology. Furthermore, the application of this discriminator procedure to the domain of interest is novel to our best of knowledge.

Schematically, the approach operates as follows:

$$\begin{aligned} \max_{\mathbf{d}} \quad & \textit{Source attention correlation} + \textit{Binary cross entropy penalty} \\ s.t. \quad & \textit{Source} - \textit{target classifier design}, \end{aligned} \tag{5.2}$$

i.e., the regular source attention correlation maximisation of equation 3.6 is biased by a binary cross entropy (BCE) term to penalise discriminability of source (denoted by $z_{\mathcal{S}} = 1, z_{\mathcal{T}} = 0$) and target domain (denoted by $z_{\mathcal{S}} = 0, z_{\mathcal{T}} = 1$) based on the correlation values $c_l(n) \in \mathbb{R}$ ($n = 1..N_l$, $l = \{\mathcal{S}, \mathcal{T}\}$). Indeed, the BCE is a measure for classification performance of the source-target classifier. Assume availability of $N_{\mathcal{S}} \in \mathbb{N}_0$ source and $N_{\mathcal{T}} \in \mathbb{N}_0$ target data frames of a predefined length of $T \in \mathbb{N}_0$ samples, this BCE term takes the following form [52][5]:

$$\begin{aligned} penalty = \frac{-1}{N_{\mathcal{S}} + N_{\mathcal{T}}} \sum_{l=\{\mathcal{S},\mathcal{T}\}} \sum_{n=1}^{N_l} & z_l(n) log(P_{Source-Target}(c_l(n))) + \\ & (1 - z_l(n)) log(1 - P_{Source-Target}(c_l(n))). \end{aligned} \tag{5.3}$$

Herein, $P_{Source-Target}(c_l(n)) \in [0, 1]$ represents the probability of a correlation value belonging to the source domain using the source-target classifier. According to the

[5] This BCE expression implicitly assumes an equal amount of source and target domain data, although generalisations exist that take class imbalance into account [84].

probabilistic classification view of section 4.2, $P_{Source-Target}(c_l(n))$ can be expressed as follows for a linear classifier with parameters $\beta \in \mathbb{R}$ and $\beta_0 \in \mathbb{R}$ ($\sigma(x) = \frac{1}{1+e^{-x}}$) [52]:

$$P_{Source-Target}(c_l(n)) = \sigma(\beta_0 + \beta c_l(n)). \tag{5.4}$$

Referring to section 4.3, the LDA classifier computes a closed-form solution for β and β_0.

Bringing everything together, let the time-lagged and channel concatenated EEG data, in Hankel format, be denoted by $X_l(n) \in \mathbb{R}^{T \times L_d C}$, and the associated audio data be denoted by $\mathbf{y}_l(n) \in \mathbb{R}^{T \times 1}$. Moreover, let $\mathbf{r}_{\underline{x}_{\mathcal{S},A} y} \in \mathbb{R}^{L_d C \times 1}$ and $R_{\underline{x}_{\mathcal{S},A}\underline{x}_{\mathcal{S},A}} \in \mathbb{R}^{L_d C \times L_d C}$ respectively denote the audio-EEG cross-correlation vector and the EEG autocorrelation matrix of the source attention data. Then, the discriminator LS feature extractor cost function is expressed as follows ($\nu \in \mathbb{R}_0^+$):

$$\begin{aligned}
\min_{\mathbf{d}} \quad & -\mathbf{d}^\top \mathbf{r}_{\underline{x}_{\mathcal{S},A} y} + \frac{\nu}{N_\mathcal{S} + N_\mathcal{T}} \sum_{l=\{\mathcal{S},\mathcal{T}\}} \sum_{n=1}^{N_l} z_l(n) \sigma(\beta_0 + \beta c_l(n)) + \\
& (1 - z_l(n))(1 - \sigma(\beta_0 + \beta c_l(n))) \\
s.t. \quad & \mathbf{d}^\top R_{\underline{x}_{\mathcal{S},A}\underline{x}_{\mathcal{S},A}} \mathbf{d} = 1 \\
& c_l(n) = \rho(X_l(n)\mathbf{d}, \mathbf{y}_l(n)), \quad n = 1..N_l, \; l = \{\mathcal{S}, \mathcal{T}\} \\
& \bar{\Sigma} = \frac{1}{N_\mathcal{S}} \sum_{n=1}^{N_\mathcal{S}} (c_\mathcal{S}(n) - \bar{\mu}_\mathcal{S})^2 + \frac{1}{N_\mathcal{T}} \sum_{n=1}^{N_\mathcal{T}} (c_\mathcal{T}(n) - \bar{\mu}_\mathcal{T})^2 \\
& \bar{\mu}_\mathcal{S} = \frac{1}{N_\mathcal{S}} \sum_{n=1}^{N_\mathcal{S}} c_\mathcal{S}(n) \\
& \bar{\mu}_\mathcal{T} = \frac{1}{N_\mathcal{T}} \sum_{n=1}^{N_\mathcal{T}} c_\mathcal{T}(n) \\
& \beta = \bar{\Sigma}^{-1}(\bar{\mu}_\mathcal{T} - \bar{\mu}_\mathcal{S}) \\
& \beta_0 = -\frac{1}{2}\beta(\bar{\mu}_\mathcal{T} + \bar{\mu}_\mathcal{S}).
\end{aligned} \tag{5.5}$$

Note that the $log(.)$ operator in the BCE term has been dropped to ease the tuning of the weight parameter ν.

In summary, equation 5.5 simultaneously performs regular source domain least squares design and minimises the discriminability of source and target domain by maximising the binary cross-entropy between both domains using an LDA classifier. Algorithm 3 presents this summary in a pseudocode fashion. Finally, we note that equation 5.5 proves to be non-convex, as is described in appendix A.3.

Algorithm 3 Discriminator (D-DA)

1: Solve equation 5.5
2: correlation($X_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}}, \mathbf{d}$)

5.4 Conclusion

Unsupervised methodologies are introduced to cope with the changing statistics in the target domain. An iterative least squares based approach is currently state-of-the-art. Nevertheless, we propose to utilise domain adaptation feature space alignment techniques: Creating a joint embedding between target and source domain to theoretically mitigate the need for unsupervised classification methods. On one hand, EEG transformations, based on CCA and PCA procedures between source and target EEG data, may be leveraged to create said joint space at the EEG level. These methods are novel in our application domain to our best of knowledge. On the other hand, discriminators introduce a source-target classifier to the feature extractor design to create the joint embedding at the feature level. To this end, the loss of this source-target classifier is maximised, i.e., the feature extractor should make it difficult for the classifier to differentiate between source and target domain features. The specific expression of this approach in an LS-LDA setting is novel to our best of knowledge.

Chapter 6

Experimental procedures

6.1 Introduction

In order to perform rigorous experiments, experimental procedures need to be outlined. To this end, section 6.2 details the validation methodologies and section 6.3 the hyper-parameter choices. Section 6.4, thereafter, describes a hypothesis testing framework, based on linear mixed-effects models and analysis-of-variance hypothesis tests, combined with the Benjamini-Hochberg correction. Next, in section 6.5, two auditory attention based datasets are outlined: the Vanthornhout dataset [16] and the Brouckmans-Dewit-Vanhaelen dataset [85]. Finally, section 6.6 concludes the chapter.

6.2 Validation methodologies

As already described in sections 3.6 and 4.4, the feature extractor and classification designs might overfit the training data. Therefore, it is important to validate the performance of these feature extractors and classifiers on a separate set, not seen at training time. Such a set is called a validation set, as illustrated in figure 6.1. Different validation strategies exist, including which [86]:

- **Subject specific validation (SS-V):** In advance, the data are split into a fixed training set and a fixed validation set per subject. Subject-specific feature extractors/classifiers are subsequently trained on the training set and validated on the validation set.
- **Subject specific cross validation (SS-CV):** This procedure is similar to the SS-V approach, with the extension that the procedure is repeated for different choices of validation sets. Indeed, instead of fixing the training and validation sets, the data are split into $k \in \mathbb{N}_0$ folds and there are k accompanying training-validation rounds. In each round $(k-1)/k$th of the data are used to train the feature extractors/classifiers, after which the feature extractors/classifiers are validated on the left-out $1/k$th of the data [86]. After the k rounds, the results are averaged over the validation folds.

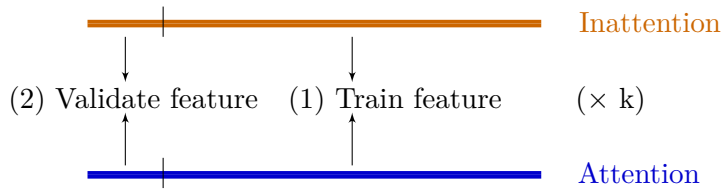


FIGURE 6.1: (k-fold cross) validation strategy. Different portions of the inattention (orange) and attention (blue) data are alternately labelled as training and validation set until each portion has been used once as a validation set.

- **Subject independent cross validation (SI-CV):** In this procedure, also referred to as leave-one-subject-out (LOSO) cross validation (CV), the data are utilised in a subject-independent way. To this end, the feature extractors/classifiers are trained on the data of all but one subject and validated on the left-out subject. This procedure is repeated until each subject has been used once as a validation set and, thereafter, results are averaged subjectwise [86]. Note that the preprocessing scheme of section 2.4 should be slightly altered in this approach: the artefact removal MWF filters are calculated on the first subject and readily applied to all other subjects to mitigate differences in preprocessing between subjects.

6.3 Hyper-parameter choice

Unless explicitly stated otherwise, the hyper-parameters, as presented in table 6.1[1], are utilised for the feature extractors, classifier and unsupervised methodologies.

The hyper-parameter choices are mainly based on the available literature, as described in chapters 3-5, and we refer the reader to these chapters. We will, however, zoom in on some choices that require special attention:

- **PSD estimation:** As detailed in section 3.2, multitaper spectral analysis is issued to estimate PSDs. To this end, 7 Slepian tapers are used to create a frequency spacing of 0.5 Hz.
- **LS-CCA-CSP-regularisation:** We favour Ledoit-Wolf L2 regularisation because of its closed-form solution and thus utilise this regularisation method in the LDA classifier and CSP feature extractor. However, due to the poor conditioning of the dataset after preprocessing, a QR-singular value decomposition (SVD) based approach is opted for as CCA implementation over the GEVD method [87, 88]. Nevertheless, this QR-SVD approach does not allow to incorporate L2 regularisation. CCA and LS designs are therefore regularised in a PCA manner, wherein a conservative approach of 32 channels is taken to mitigate the loss of neural envelope tracking information. The interested reader can find this QR-SVD CCA implementation in [87, 88].

[1] n.a. refers to not applicable.

- **LS-CCA lag values:** Regarding the LS lag values, a conservative choice of $[0, 500]$ ms captures the relevant neural response of the auditory stimuli [8, 16]. Subsequently, the CCA lag values (EEG: $[0, 400]$ ms, audio: $[-100, 0]$ ms) are chosen such that the total model lag for LS and CCA is of equal length. This makes the comparison between both methods fair in the sense that an equal amount of data is leveraged in both methods, although we acknowledge that this CCA model requires more parameters than its LS counterpart.
- **KLD frequency band:** Since we do not want to make prior assumptions on the frequency band with respect to this novel feature, we will utilise the unfiltered frequency band. Nevertheless, only the $[0.5, 64]$ Hz frequency range is effectively utilised after PSD estimation in order to remove the DC component, as well as low frequency noise.
- **LASSO-KLD regularisation:** After performing a prior hyper-parameter sweep (LASSO: $\gamma \in [10^{-10}, 10^{-5}]$ and KLD late fusion: $\gamma \in [10^{-10}, 10^{0}]$) using a 2 fold CV procedure, the average over all folds and subjects is taken as regularisation parameter (LASSO: $1.41 \cdot 10^{-7}$ and KLD late fusion: $7.91 \cdot 10^{-5}$). For the sake of completeness, we mention that this hyper-parameter sweep has been performed using a Bayesian optimisation framework with 30 iterations [89]. We refer the interested reader to [89] for a discussion about this Bayesian optimisation procedure.
- **CSP frequency band:** We will utilise CSP using delta, theta, alpha and beta frequency bands, both separately and in conjunction with one another.
- **BP-Entropy frequency band:** As with the KLD, the unfiltered frequency band is used to obtain a PSD estimation. Delta, theta, alpha and beta frequency ranges are selected after PSD estimation.
- **Classifier choice:** The classifier can also be seen as a hyper-parameter. To this end, we utilise the LDA classifier. This LDA classifier assumes normality of the feature space and equal covariance matrices for the class conditional distributions. Although this LDA classifier could still prove its worth if these assumptions do not hold [52], it is insightful to assess them. Using Mardia's test [90], evidence is found against this normality assumption, although this deviation from normality seems rather limited for the mean cases. Similarly, using Box's M test [91], evidence is found against the equal covariance assumption. Nonetheless, the mean ratio between the covariance matrices does seem to have the proper order of magnitude $O(1)$. The interested reader can find these experiments in appendix B.
- **Unsupervised regularisation:** To allow for a fair comparison between the unsupervised methodologies and the PCA-regularised LS supervised feature extractors, wherein 32 channels are kept, CCA-DA, PCA-DA, SA-DA and TPCA-DA all utilise 32 bands as well.
- **D-DA discriminator weight:** The weight ν of the discriminator term is fixed to 10^{-2}. Indeed, a prior sweep (5 log-spaced values between $[10^{-4}, 10^{2}]$) has shown that this weight value corresponds to the lowest weight to yield a significant difference between D-DA and regular supervised designs, i.e., this weight is large enough to reflect the discriminator term reliably.

- **D-DA initial value:** Since the D-DA problem formulation (see equation 5.5) is non-convex, an iterative strategy is required. To this end, an optimisation algorithm (interior point) is utilised, which is initialised with a supervised source trained LS decoder.
- **ILS weights:** The weights δ and η are both fixed to 0.5.

6.4 Hypothesis testing

In order to make statistically significant conclusions about any results, hypothesis tests are required. These tests assume a null-hypothesis (i.e. some belief one would like to examine, e.g., discriminability of a feature between the attention and inattention case) and return a p-value denoting the probability that the null-hypothesis holds true [92]. In this work, we will combine the concepts of linear mixed-effects models and analysis of variance tests for hypothesis testing purposes. These concepts are described, respectively in sections 6.4.1 and section 6.4.2. Furthermore, after performing simultaneous hypothesis tests, the resulting p-values need to be corrected, as will be explained in section 6.4.3 [92]. To this end, we leverage the Benjamini-Hochberg correction [93].

6.4.1 Linear mixed-effects models

A linear mixed-effects model (LME) essentially amounts to an extension of linear regression models to include both fixed and random effects [94, 95, 96]. Herein, coarsely put, a fixed effect represents the effect under study and a random effect represents the effect that influences the outcome variable, but which is not under study [95]. Mathematically speaking, the following data model is used:

$$\mathbf{v} = H\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\epsilon}. \tag{6.1}$$

Herein, $\mathbf{v} \in \mathbb{R}^{N\times 1}$ denotes the response vector, with $N \in \mathbb{N}_0$ the number of outcomes. The right hand side is constructed using three terms: one corresponding to the fixed effects ($H\boldsymbol{\beta}$), one corresponding to the random effects ($Z\mathbf{b}$) and one corresponding to the residuals $\boldsymbol{\epsilon}$. More specifically, $H \in \mathbb{R}^{N\times Q}$ corresponds to the design matrix of the fixed effects and $\boldsymbol{\beta} \in \mathbb{R}^{Q\times 1}$ contains the associated $Q \in \mathbb{N}_0$ fixed effect parameters. Correspondingly, $Z \in \mathbb{R}^{N\times P}$ and $\mathbf{b} \in \mathbb{R}^{P\times 1}$ denote respectively the random effect design matrix and the random effect parameter vector of dimension $P \in \mathbb{N}_0$. Finally, $\boldsymbol{\epsilon} \in \mathbb{R}^{N\times 1}$ denotes the residual vector.

Moreover, the following model assumptions are made in an LME [96, 97]: The random effect vector and residuals are assumed to have a normal distribution with covariance matrices respectively $\sigma^2 D(\theta) \in \mathbb{R}^{P\times P}$ and $\sigma^2 U \in \mathbb{R}^{N\times N}$, while being independent of one another [96, 97]:

$$\begin{aligned} \mathbf{b} &\sim \mathcal{N}(\mathbf{0}_{P\times 1}, \sigma^2 D(\theta)); \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}_{N\times 1}, \sigma^2 U); \\ \mathbf{b} &\textit{ and } \boldsymbol{\epsilon} \textit{ independent}. \end{aligned} \tag{6.2}$$

Method	Frequency band	Time-lags	Regularisation method	Regularisation strategy
LS	Delta	EEG: $[0, 500]$ ms audio: 0 ms	PCA	32 channels
LASSO	Delta	EEG: $[0, 500]$ ms audio: 0 ms	LASSO	$\gamma = 1.41 \cdot 10^{-7}$
CCA	Delta	EEG: $[0, 400]$ ms audio: $[-100, 0]$ ms	PCA	32 channels
KLD late fusion	Unfiltered ($[0.5, 64]$ Hz)	n.a.	L2	$\gamma = 7.91 \cdot 10^{-5}$
KLD early fusion	Unfiltered ($[0.5, 64]$ Hz)	n.a.	norm-constraint	$\|\|\mathbf{a}\|\|_2 = 1$
CSP	Delta, Theta, Alpha, Beta	EEG: 0 ms audio: n.a.	L2	Ledoit-Wolf
BP	Delta, Theta, Alpha, Beta	n.a.	n.a.	n.a.
Entropy	Delta, Theta, Alpha, Beta	n.a.	n.a.	n.a.
LDA	n.a.	n.a.	L2	Ledoit-Wolf
CCA-DA, PCA-DA SA-DA and TPCA-DA	Delta	EEG: $[0, 500]$ ms audio: 0 ms	n.a.	32 channels
D-DA, ILS	Delta	EEG: $[0, 500]$ ms audio: 0 ms	PCA	32 channels

TABLE 6.1: Hyper-parameter choices of the feature extractors, LDA classifier and unsupervised approaches.

The covariance matrices are also assumed to be positive semidefinite and parametrised by $\sigma \in \mathbb{R}$ and $\theta \in \mathbb{R}$. Yet, these constraints are not enough to provide a unique solution and an additional assumption on the covariance matrix of the residuals has to be made [96]. In this work, the following constraint is adhered to [96, 98]:

$$\sigma^2 U = \sigma^2 \mathbb{I}_{N\times N}. \tag{6.3}$$

In words, this choice of covariance matrix implies uncorrelatedness of the residuals. Given a response vector $\mathbf{v}$ and design matrices H and Z, the parameters $\boldsymbol{\beta}$, $\mathbf{b}$, σ and θ can be estimated using a maximum-likelihood framework, for which the interested reader is referred to [96, 97].

6.4.2 One-way analysis of variance

One-way analysis of variance (ANOVA) [99, 100] corresponds to a hypothesis test on $P \in \mathbb{N}_0$ groups, each with N_p, $p = 1..P$ observations, under the null-hypothesis of equal means:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_P, \tag{6.4}$$

wherein μ_p, $p = 1..P$ represents the mean of each group. Furthermore, the test operates under the following assumptions [99, 100]:

- independent observations;
- independent groups;
- data are drawn from a normal distribution;
- equal standard deviations within groups.

Intuitively, the corresponding F-statistic equals the ratio of the variation between groups to the variation within groups [100]. Formally, let $x_{p,n}$ refer to the nth ($n = 1..N_p$) data-sample of the pth group ($p = 1..P$). Then, the F-statistic can be computed as follows:

$$\begin{aligned}
\bar{\mu}_p &= \frac{1}{N_p}\sum_{n=1}^{N_p} x_{p,n};\\
N &= \sum_{p=1}^{P} N_p;\\
\bar{\mu} &= \frac{1}{N}\sum_{p=1}^{P}\sum_{n=1}^{N_p} x_{p,n};\\
F &= \frac{\frac{1}{P-1}\sum_{p=1}^{P}\sum_{n=1}^{N_p}\left(\bar{\mu}_p - \bar{\mu}\right)^2}{\frac{1}{N-P}\sum_{p=1}^{P}\sum_{n=1}^{N_p}\left(x_{p,n} - \bar{\mu}_p\right)^2}.
\end{aligned} \tag{6.5}$$

Note that this F-statistic follows an F-distribution $F(P-1, N-P)$, which is used to calculate the associated p-value, i.e., the probability that the null-hypothesis is true [100].

This ANOVA hypothesis test can subsequently be issued in conjunction with the LME, regarding the test whether the fixed effect vector estimate $\bar{\boldsymbol{\beta}} \in \mathbb{R}^{Q\times 1}$ significantly differs from zero [101]:

$$H_0 : \bar{\beta}(q) = 0, \forall q = 1..Q. \tag{6.6}$$

In other words, this test checks whether the fixed effect parameter is necessary to fit the data [101]: If the p-value is smaller than some significance level (in this work: 5%) the fixed effect parameter is assumed to be necessary to explain the response variable $\mathbf{v}$. Under this null-hypothesis of equation 6.6, the F-statistic can be computed as follows [101]:

$$F = \frac{\left(\bar{\beta}(q) - 0\right)^2}{Var(\bar{\beta}(q))},\ \forall q = 1..Q. \tag{6.7}$$

6.4.3 Correcting p-values

When performing a hypothesis test under significance level $\alpha \in \mathbb{R}$, the probability of falsely rejecting the null-hypothesis (so-called false positive or false discovery) equals α [92]. When performing $L \in \mathbb{N}_0$ simultaneous hypothesis tests under that same significance level α, the probability of having false positives increases to a level αL [92]. In order to deal with this multiple testing problem, a group of methods exist to ensure that the probability of falsely rejecting the null-hypothesis is kept below the significance level α [92].

One such method to control this false discovery rate (FDR) is the Benjamini-Hochberg correction, which is used in this work [93]. To achieve an FDR of α, the L p-values are sorted in decreasing order and corrected in the following way [93]:

$$\begin{aligned} &p_{corrected}(l) = p(l),\ l = 1; \\ &p_{corrected}(l) = min\left(p(l-1), \frac{L}{L-l+1}p(l)\right),\ l = 2..L. \end{aligned} \tag{6.8}$$

6.5 Datasets

Ideally, the goal is to find feature extractor-classifier combinations that can differentiate between attention to an auditory stream and any form of inattention to that stream. Unfortunately, no such dataset exists at present time and could no longer be recorded within the time span of this thesis. The following two datasets concerning auditory attention can nonetheless be leveraged:

- **Vanthornhout dataset:** This dataset is concerned about auditory versus visual attention [16].
- **Brouckmans-Dewit-Vanhaelen dataset:** This dataset contains auditory attention conditions, as well as conditions wherein subjects are instructed to solve mathematical exercises or to read a text while an auditory stimulus is present [85].

The Vanthornhout dataset originates from [16], wherein 20 normal-hearing and unpaid subjects participated in (one of) the following experiments:

- actively listen to a stimulus and answer some questions afterwards;
- watch a silent, subtitled cartoon movie while ignoring the stimulus.

These subjects were volunteers between 18 and 35 years old and were native Dutch (Flemish) speakers [16]. Regarding the stimulus, two different speech types were used: a story stimulus and concatenated sentences of the Flemish Matrix Test [102].

The story stimulus corresponds to the story 'Milan', as narrated by the Flemish male speaker Stijn Vrancken, and resembles a real life continuous auditory stream [16]. This stimulus consists of continuous speech, has a length of about 15 minutes and was presented using a signal-to-noise-ratio (SNR) of 100 dB.

On the other hand, the sentences of the Flemish Matrix test follow a fixed pattern of 'name verb numeral adjective object' (e.g. Ellen ziet twee beige fietsen), wherein for each grammatical category 10 options are available [16, 103]. The corresponding matrix-stimulus is subsequently created by concatenating 20 such sentences and adding a random silence, ranging from 0.8 s to 1.2 s, between these sentences. Each such matrix-stimulus was presented between 1 and 7 times, effectively creating an audio fragment between 2 minutes and 14 minutes. Moreover, these sentences were presented at various SNR-levels between -12dB and 100 dB. On the plus side, the phoneme occurrence count in this stimulus type is close to the distribution of the Dutch language [103]. On the minus side, these sentences may be dull to actively listen to, interfering with the experiment setup.

There are 7 subjects for which data for all cases (attention-story, attention-matrix, movie-story and movie-matrix) are available. For the union of the attention-story, attention-matrix and movie-matrix cases, 17 subjects are available. EEG has been recorded using a 64-channel BioSemi EEG cap[2] in a radio frequency shielded room.

Furthermore, a subset of the Brouckmans-Dewit-Vanhaelen dataset, originating from [85], is utilised. 10 Flemish speaking and unpaid subjects, between 21 and 27 years old, engaged in the following experiments:

- actively listen to a stimulus;
- solve as many mathematical exercises (e.g. $1012 - 448$) as possible while ignoring the auditory stimulus;
- read a text while ignoring the auditory stimulus.

In each of these conditions different story stimuli were utilised: 'Bianca en Nero' narrated by Luc Nuyens (male) in the auditory attention condition, 'Ver van het kleine paradijs' narrated by Iris Van Cauwenbergh (female) and 'Milan' narrated by Stijn Vrancken (male) in the mathematical exercises distraction condition, and

[2] https://www.biosemi.com/

'Eline' narrated by Luc Nuyens (male) in the text reading distraction condition.

In total, about 20 minutes of the auditory attention condition, about 5 minutes of the mathematical exercise distraction condition and about 13 minutes of the text reading distraction condition are available. Contrary to the Vanthornhout dataset, a smarting[3] mobile EEG cap, consisting of only 24 EEG channels, is utilised [4,5]. Furthermore, data were collected in a non-radio frequency shielded room.

Note that the datasets have a major downside: the setup permits solely to distinguish between subjects actively listening and subjects under dedicated distraction conditions (while passively listening), whereas the general goal is to detect any form of inattention to the auditory stimulus. Therefore, it is difficult to draw general conclusions from these datasets. We will nevertheless refer to auditory attention as 'attention' and visual attention as 'inattention' or 'movie'. The mathematical exercises and text reading distraction conditions will also be referred to as 'inattention' or respectively 'mathematics' and 'text'.

Throughout this work, mainly the Vanthornhout dataset will be leveraged for experimental explorations, whereas the Brouckmans-Dewit-Vanhaelen dataset will be used to validate performance of the final feature extractor-classifier proposal. Thus, unless specifically stated otherwise, usage of the Vanthornhout dataset is assumed.

6.6 Conclusion

In order to assess the performance of feature extractor-classifier combinations SS-V, SS-CV and SI-CV approaches are used. Moreover, the corresponding hyperparameters can be found in table 6.1. In order to perform statistical evaluation, hypothesis testing using a framework of LME, ANOVA and Benjamini-Hochberg corrections is utilised. Finally, the Vanthornhout dataset and the Brouckmans-Dewit-Vanhaelen dataset will be used in the remainder of this work. The Vanthornhout dataset consists of two states: attention to an auditory stimulus and attention to a visual stimulus while ignoring the auditory stimulus. To this end, two types of stimuli are leveraged: a continuous story and concatenated sentences of the Flemish matrix test. The Brouckmans-Dewit-Vanhaelen dataset consists of an auditory attention condition and two distractors, namely mathematical exercise solving and text reading. In this dataset, only story stimuli are utilised. Note that the limited number of attention conditions might hamper the drawing of general conclusions.

[3] https://mbraintrain.com/

[4] The 24-channel EEG cap consists of channels Fp1, Fp2, Fz, F7, F8, FC1, FC2, Cz, C3, C4, T7, T8, CPz, CP1, CP2, CP5, CP6, TP9, TP10, Pz, P3, P4, O1 and O2. These channels are further treated exactly the same as the 64-channel EEG cap. Only the TP9 and TP10 channels are not available in the 64-channel cap and are added respectively to the left-temporal (LT) and right-temporal (RT) groups.

[5] Raw EEG data are converted into a .mat format using OpenVibe software available at http://openvibe.inria.fr/converting-ov-files-to-matlab/.

Chapter 7

Differentiating nature features

7.1 Introduction

This chapter presents the experiments, linked to the first high-level objective: *Which **features** can **discriminate** between **attention** and inattention to an **auditory stream**?* We will inspect this differentiating nature for each of the features, as introduced in chapter 3.

Section 7.2 presents the low-level research questions and corresponding experiments. Thereafter, section 7.3 concludes this chapter.

7.2 Experiments

This section presents the low-level research questions and corresponding experimental results in subsections 7.2.1-7.2.10. Referring to section 6.2, we will utilise the Vanthornhout dataset in the following ways:

- **SS-CV:** A 10 fold cross validation approach is taken, wherein the **attention-story and movie-story** data are partitioned into 10 folds. To this end, data of 7 subjects are available.
- **SS-V:** The **attention-story and movie-story data are used as a training set** and **the attention-matrix and movie-matrix data as a validation set**. The union of attention-matrix, attention-story and movie-story contains 17 subjects. Therefore, the LS, KLD, BP and entropy experiments are conducted using these 17 subjects. However, when adding the movie-story condition to the union, only 7 subjects remain. Therefore, the LASSO and CSP experiments are conducted using 7 subjects.

We reemphasise that the hyper-parameters remain fixed as shown in table 6.1, unless explicitly stated otherwise. Hypothesis tests are performed according to the procedures of section 6.4, where the subject identifier and SNR level are incorporated as a random effect, and the feature value and offset as a fixed effect. Furthermore, these hypothesis tests are performed using a significance level of 5%.

We also reemphasise that the features are either novel as such (KLD), novel in the domain (LASSO, CCA and CSP) or novel in the setup (BP and entropy) to our best of knowledge; except for the LS feature. Indeed, in [16], the LS feature has already been investigated on the same Vanthornhout dataset. Nevertheless, our research is complementary: Firstly, in [16], the effect of the different SNR levels was studied, whereas here we aggregate over SNR levels using an LME. Secondly, in [16], only an SS-V approach was adhered to, whereas we also leverage an SS-CV approach. Finally, in [16], only delta and theta band and EEG lags $[0, 75]$ ms and $[0, 500]$ ms were studied, whereas we issue a more fine-grained LS hyper-parameter evaluation.

7.2.1 Least squares frequency band influence

Research questions

1. Which frequency bands reflect the level of attention to an auditory stream when using an LS approach?
2. Which frequency band is most differentiating?

Hypothesis

Since, in [16], a significant difference was found in the delta band between the attention and inattention state, we hypothesise the delta band to be differentiating. This hypothesis is strengthened by the overlap of the delta band with the word and phrase rate [29, 30], as described in section 3.3. In [16], furthermore, no significance was found between attention and inattention in the theta band such that we do not expect to find significant differences in this theta band. Moreover, debate exists in the literature on whether the alpha band decodes attention to auditory stimuli [32, 33]. Finally, regarding the beta band, we expect no significant differences since it does not overlap with word, phrase or syllable rate.

We further hypothesise the delta band to be most differentiating, given the results of [16] and the classic relation of the delta band to auditory stimuli [8, 33].

Experiment

The frequency bands are varied between delta, theta, alpha and beta band. Firstly, to assess the differentiable nature of the frequency bands with respect to the LS feature extractor, hypothesis tests are performed between attention and inattention correlations per frequency band. Subsequently, hypothesis tests on the difference in correlation between attention and inattention (attention-inattention) per band with respect to the delta band are performed to assess which frequency band is most differentiating.

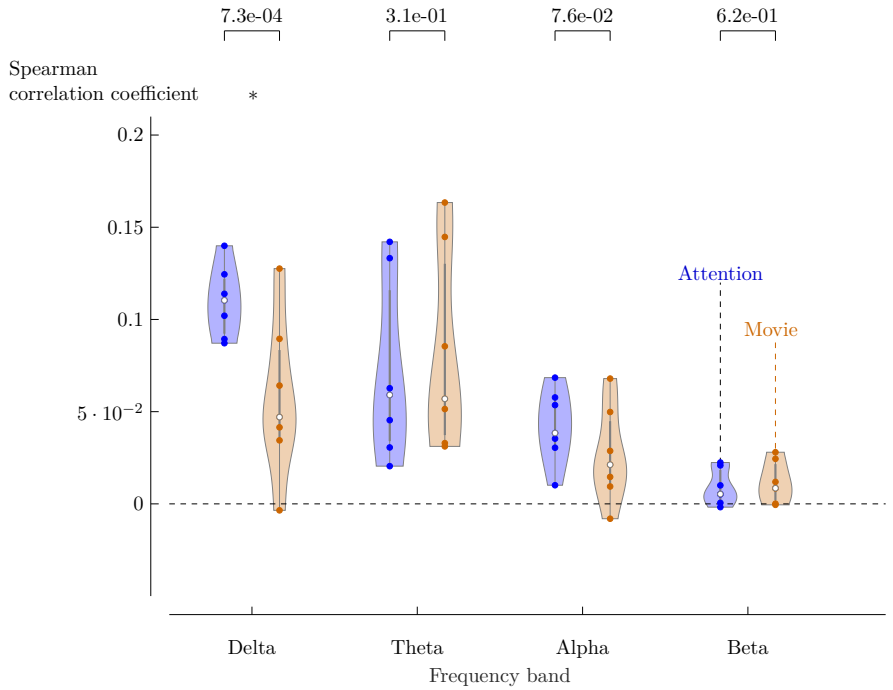


FIGURE 7.1: The spearman correlation, using an SS-CV approach, shows a significant difference in correlation regarding the delta band. In addition, the alpha band is almost differentiating. No significant differences are observable in the theta or beta band.

Result

Figure 7.1[1,2,3] displays the Spearman correlation coefficients, adhering to the SS-CV strategy. The delta band is differentiating between the attention and movie cases ($p=7.3 \cdot 10^{-4}$) and the alpha band is almost so ($p=7.6 \cdot 10^{-2}$). On the contrary, the theta and beta band do not result in any form of significance (respectively $p=3.1 \cdot 10^{-1}$ and $p=6.2 \cdot 10^{-1}$). Using the SS-V approach, different results are obtained: The alpha and beta bands prove to be differentiating (respectively $p=1.6 \cdot 10^{-2}$ and $p=1.3 \cdot 10^{-2}$), whereas the delta and theta band do not ($\min(p)=1.2 \cdot 10^{-1}$).

Furthermore, when comparing the attention-movie correlation difference for each band with respect to the delta band, significance is obtained in each case, when adhering to the SS-CV approach ($\max(p)=9.8 \cdot 10^{-4}$).

[1]This and all subsequent violin boxplots were generated using code developed by B. Bechtold and available at https://github.com/bastibe/Violinplot-Matlab. This framework was accessed on 19/10/2021.

[2]MATLAB figures were converted into tikz format using code mainly developed by N. Schlöme and available at https://github.com/matlab2tikz/matlab2tikz on 11/11/2021.

[3]Asterisks * denote significance with respect to a level of 5%.

Discussion

Surprisingly, using an SS-V approach, alpha and beta decoders yield significant differences, whereas delta and theta decoders do not. Nevertheless, these results need to be interpreted with care. Indeed, correlation values of both alpha and beta decoders lie close to zero. In fact, the mean correlation averaged over the SNR levels, is of order of magnitude $O(10^{-3})$, both for the attention and the movie case. Therefore, this significance may possibly be attributed to the finite dataset size and not (entirely) to the underlying differentiating nature of these bands. This statement is strengthened by the fact that delta and theta bands are not differentiating in this SS-V approach, despite these bands classically being related to envelope tracking [8, 33]. Furthermore, the results of the SS-CV approach seem to confirm this statement as well.

Indeed, using this SS-CV approach, **the delta band is (most) differentiating and seems to encode differences in neural envelope tracking between attention and inattention states**. As hypothesised, neural tracking is possibly induced due to overlap with word and phrase rate [29, 30]. Although the theta band achieves correlations that are significantly different from zero ($p=7.2 \cdot 10^{-4}$ for attention and $p=7.2 \cdot 10^{-4}$ for movie), theta band decoders do not result in a significant difference between attention and inattention. Therefore, the theta band appears to encode some level of envelope tracking, possibly due to overlap with the syllable rate [29, 30], albeit in a similar manner regarding the attention and inattention cases. The fact that the alpha band results in almost-significance is in line with [32], although according to [33], this result needs to be interpreted with caution since the dominant alpha power might interfere with the results. Also, as hypothesised, the beta band does not result in a significant difference between attention and inattention.

Taking these arguments into account, we may possibly state that the SS-V approach does not result in 'meaningful' significance. This 'insignificance' may be caused by interference in the setup: The sentences of the Flemish matrix test are dull to listen to, such that the actual attention state might have differed from the requested one [16].

7.2.2 Least squares starting lag influence

Research questions

Does the difference in correlation between attention and inattention (i.e. the difference in neural envelope tracking) increase when the early EEG lag values are excluded?

Hypothesis

It is argued that the brain tracks both attended and ignored auditory streams at early lag values $[0, 85]$ ms, while suppressing the ignored auditory streams with respect to the attended streams at later lag values > 85 ms [5, 27]. In [16], however, tracking

differences between attention and inattention states has already been observed at an ending lag value of 75 ms. Altogether, we hypothesise that excluding the early lag values will result in a larger difference in correlation between attention and inattention. However, we also hypothesise that excluding too much data will eventually result in a lower correlation difference since in that case (too much) envelope tracking information will be thrown away.

Experiment

The starting lag value is varied over the following values: 0, 75, 100, 200, 300 and 400 ms. A hypothesis test between attention and inattention for each pair of lag values $[x, 500]$ ms ($x \in \mathbb{N}_0$) is performed, as well as a hypothesis test on the difference between attention and inattention (attention-inattention) correlations for each lag value pair with respect to the $[0, 500]$ ms base case.

Result

Figure 7.2 presents the Spearman correlations using an SS-CV approach. Excluding EEG lag values results in decreasing correlation values, regarding both the attention and the movie case. Nevertheless, all correlation values are (almost) significantly different from 0, e.g. , the $[300, 500]$ ms movie case ($p=4.9 \cdot 10^{-3}$) and the $[400, 500]$ ms movie case ($p=6.0 \cdot 10^{-2}$). Moreover, all lag value pairs result in a significant difference between the attention and movie case ($\max(p)=3.8 \cdot 10^{-2}$).

The attention-movie correlation difference of each lag value pair is also compared with respect to the $[0, 500]$ ms base case: This base case proves to yield a significant higher correlation difference with respect to the $[200, 500]$ ms, $[300, 500]$ ms and $[400, 500]$ ms cases ($\max(p)=2.8 \cdot 10^{-2}$). However, the base case does not significantly outperform the $[75, 500]$ ms and $[100, 500]$ ms cases ($\min(p)=1.1 \cdot 10^{-1}$). Using an SS-V approach, no differences between attention and movie are observed for any of the lag value pairs ($\min(p)=1.3 \cdot 10^{-1}$).

Discussion

The results suggest that there is a **loss in envelope tracking whenever EEG time lags are excluded**. Indeed, the decreasing trend in correlation values, for both the attention and the inattention case, points to a loss in envelope tracking. However, more importantly, there also seems to be a loss in the difference in envelope tracking between the two cases. Indeed, the difference between attention and inattention correlation is significantly different, with respect to the $[0, 500]$ ms case, regarding the $[200, 500]$ ms, $[300, 500]$ ms and $[400, 500]$ ms cases. In addition, t**here does not seem to be any gain in excluding early lag values since the $[0, 500]$ ms case does not perform significantly different from the $[75, 500]$ ms and $[100, 500]$ ms cases**.

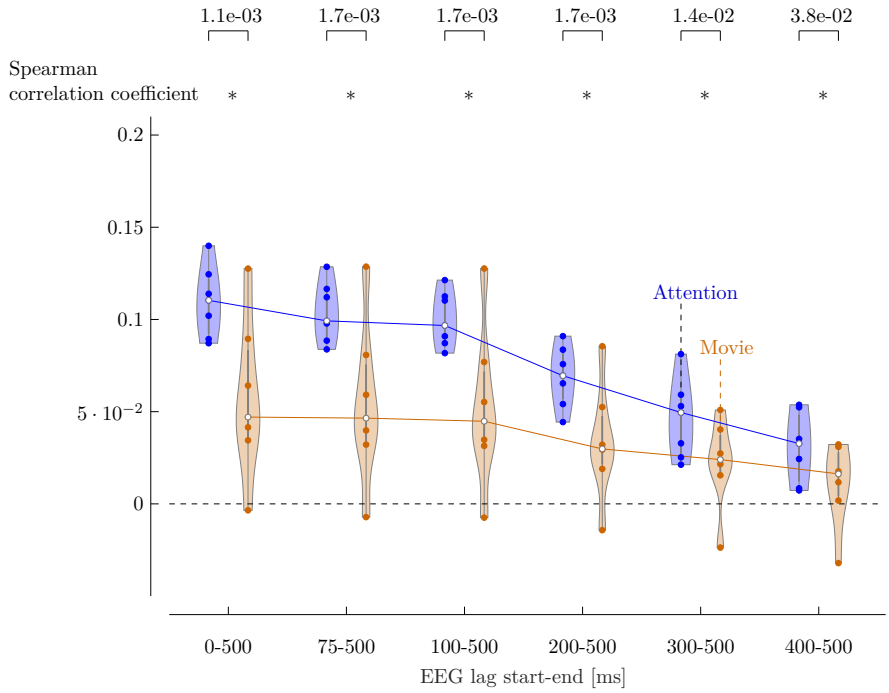


FIGURE 7.2: The spearman correlation, using an SS-CV approach, shows a significant difference in correlation regarding all shown lag values. However, the overall trend of correlation magnitude is declining.

7.2.3 Least squares ending lag influence

Research questions

To what extent can late neural responses be excluded, i.e., to what extent do late neural responses influence the envelope tracking?

Hypothesis

Given that in [16] a significant difference was found (at the -6.9 dB SNR level) using an SS-V approach with $[0, 75]$ ms decoders, we hypothesise that the late neural lag values can be excluded to $[0, 75]$ ms. However, we do not expect to find a significant difference between attention and inattention when using too low ending lag values ($<$ 75 ms). Indeed, in [27], it has been observed that simultaneously played, attended and ignored auditory streams are tracked similarly at these low lag values. Nonetheless, we note that our domain of interest is somewhat different from this AAD domain.

Experiment

The ending lag value is varied over the following values: $\frac{1}{128}$, 10, 20, 50, 75 and 500 ms. Note that an ending lag value of $\frac{1}{128}$ ms effectively corresponds to a spatial decoder since the sampling rate equals 128 Hz. Further, hypothesis tests between

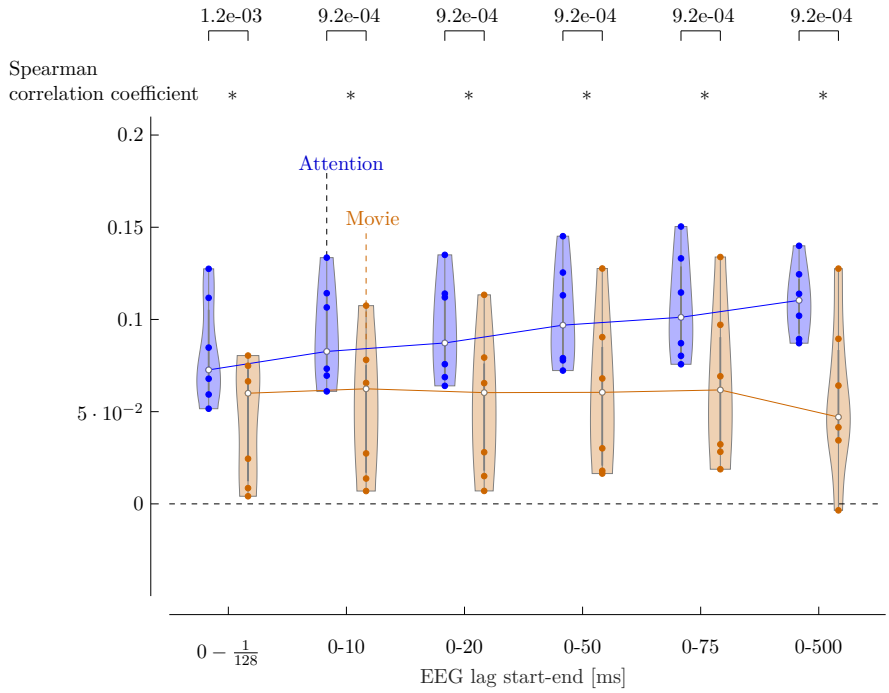


FIGURE 7.3: The spearman correlation, using an SS-CV approach, shows a significant difference in correlation regarding all shown lag values. The overall trend of correlation magnitude is increasing with increasing ending lag value regarding the attention case, and about stationary regarding the movie case.

attention and inattention are performed for each pair of lag values $[0, x]$ ms ($x \in \mathbb{N}_0$), as well as between the attention-inattention correlation differences for each lag value pair with respect to the $[0, 500]$ ms base case.

Result

Figure 7.3 displays the Spearman correlation coefficients, using an SS-CV approach. All ending lag values result in a significant difference between the attention and movie case ($\max(p) = 1.2 \cdot 10^{-3}$). Moreover, whereas the trend in correlations is nominally increasing with increasing ending lag value regarding the attention case, the same does not hold true regarding the movie case. Median movie correlations indeed remain about stationary with respect to this movie case. Furthermore, no attention-movie correlation difference significantly differs with respect to the $[0, 500]$ ms base case ($\min(p) = 1.5 \cdot 10^{-1}$). Using an SS-V approach, no significance is obtained for any lag value pair ($\min(p) = 5.4 \cdot 10^{-1}$).

Discussion

In line with our hypothesis, decoders whose ending lag value exceeds 75 ms, are differentiating between the attention and movie case using an SS-CV approach.

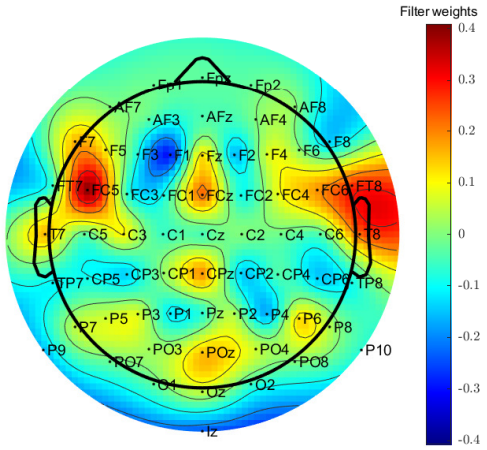


FIGURE 7.4: Topographic map of the filter weights when using a spatial decoder under an SI-CV approach. Structuring around the temporal lobes is clearly visible. Note that the filter weight structuring would be more smooth if regularisation would be added.

However, contrary to our hypothesis, decoders with even lower lag values are also differentiating between both cases. In fact, there is no significant difference between the lagged and spatial decoders. Since the EEG response always follows the stimulus onset, due to causality, it seems unlikely that these spatial decoders are effectively decoding the stimulus envelope. Yet, figure 7.4 may be utilised to assess our claims. This topographic plot shows the (normalised) filter weight distribution across the scalp when adhering to an SI-CV approach without any regularisation[4]. Referring to section 3.7, we acknowledge this topographic plot cannot be interpreted as such. Nevertheless, the highly structuring of the weights around the temporal sensors may be used to assess the spatial decoder's functionality [8]. Remains the question how this spatial decoder operates if it is only marginally causal. One hypothesis is that this spatial decoder is performing segmentation based on brain activity. However, this does not explain why the correlations at these lag values are significantly different from 0 ($p=1.1 \cdot 10^{-7}$ for attention and $p=9.8 \cdot 10^{-5}$ for movie). Moreover, the LS decoders do not process any inattention data at training time. Other hypotheses are mixing in acausal information due to MWF artefact removal, such that the decoder is not entirely acausal, and a small misalignment between EEG and stimulus, although **the bottom line still remains that lagged decoders do not seem to significantly outperform their spatial counterparts**.

[4]The SI-CV approach mitigates subject-specific influences for the sake of interpretability of the topographic plots. Preprocessing has been repeated by applying the MWF filters for the first subject to all subjects to limit differences in preprocessing between subjects.

7.2.4 Significance of least absolute shrinkage and selection operator

Research questions

Does the sparsity of decoders reflect the level of attention to the auditory stream?

Hypothesis

Despite contradictory results in the AAD domain [28, 35], we expect a significant difference in sparsity between LASSO regularised decoders trained on attention and inattention data. Decoders trained on attention data are hypothesised to be more sparse due to higher levels of envelope tracking, whereas decoders trained on the inattention data are expected to be less sparse [35].

Experiment

The sparsity of LASSO regularised decoders, trained on the attention and inattention data, is compared using the Gini-index.

Result

Figure 7.5 shows the Gini-indices of the LASSO regularised decoders, trained on attention and movie data, using an SS-CV approach. Visually, the Gini-indices (and thus sparsity) of the attention decoders exceed the Gini-indices of the movie decoders. Indeed, for 4 out of 7 subjects the sparsity decreases when going from the attention to the movie case. However, no significance is obtained ($p=2.9 \cdot 10^{-1}$). In addition, no significance is obtained using an SS-V approach ($\min(p)=6.3 \cdot 10^{-1}$).

Discussion

Contrary to our hypothesis, the sparsity of the LASSO regularised decoders does not result in a significant difference in the experimental setups. In fact, even using the entire Vanthornhout dataset (instead of a 10 fold SS-CV), no statistical significance ($p=2.9 \cdot 10^{-1}$) is obtained. **Therefore, it might be that the sparsity of LASSO regularised decoders does not (reliably) reflect the difference in (in)attention towards an auditory stream.**

7.2.5 Significance of canonical correlation analysis

Research questions

1. Does canonical correlation analysis reflect the level of attention to the auditory stream?
2. How does canonical correlation analysis perform with respect to least squares?

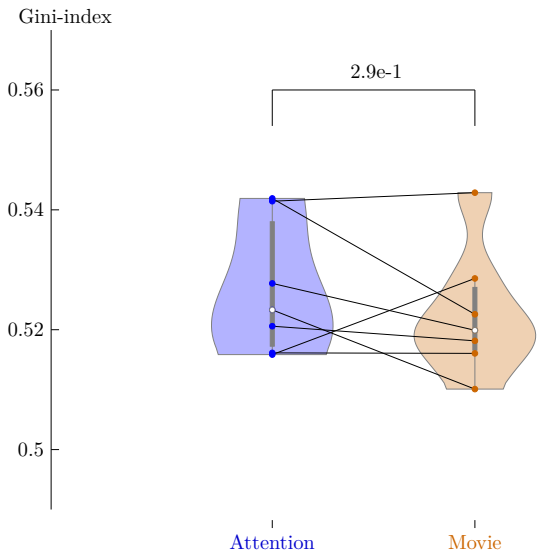


FIGURE 7.5: Gini-indices of LASSO-regularised LS decoders, when adhering to an SS-CV strategy. Larger Gini-indices denote sparser solutions.

Hypothesis

As mentioned in section 3.3, CCA corresponds to a generalisation of least squares due to the introduction of filters on the auditory side. Considering the significance of least squares decoders, we hypothesise to subsequently obtain significance using the CCA approach. Furthermore, we expect to find higher correlations in the first CCA filters with respect to the LS approach due to CCA's model being more flexible, although this flexibility does not necessarily imply that the difference in correlation will be larger. Indeed, CCA tries to maximise the correlation of the attention case and does not try to maximise the difference in correlation between the attention and inattention cases as such.

Experiment

The LS and CCA approaches are both applied to the dataset. As before, hypothesis tests are performed both within a setup (i.e. LS or CCA filter) and between attention-inattention correlation differences across setups with respect to the LS base case.

Result

As displayed in figure 7.6, the SS-CV approach on the Vanthornhout dataset results in a significant difference between the attention and movie case for the first ($p=3.6 \cdot 10^{-3}$), second ($p=8.4 \cdot 10^{-3}$) and fourth ($p=1.2 \cdot 10^{-2}$) CCA filter. Furthermore, the fifth filter results in almost significance ($p=7.1 \cdot 10^{-2}$). In line with the LS results of section 7.2.1, significance is obtained for the LS decoder ($p=1.8 \cdot 10^{-4}$).

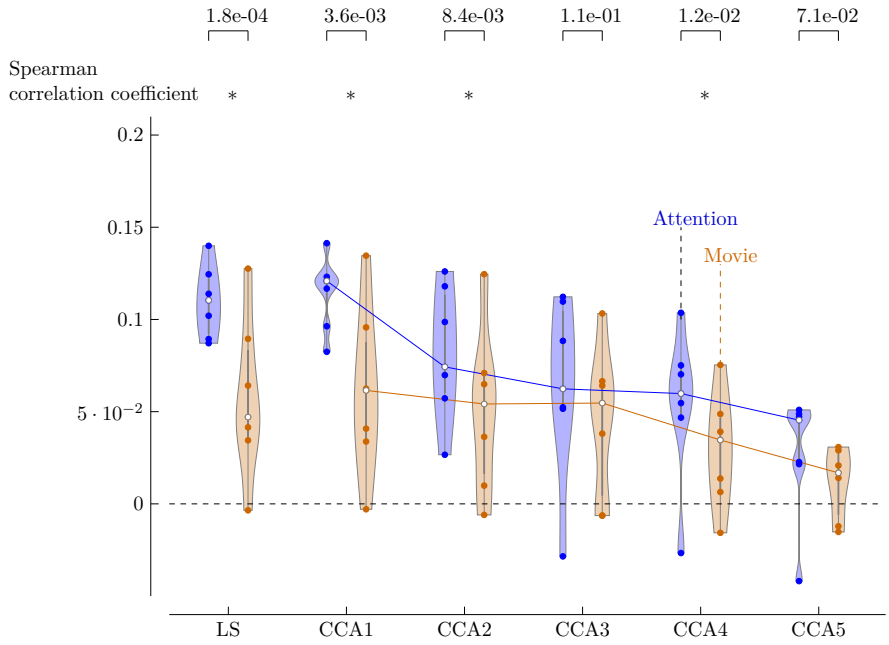


FIGURE 7.6: Spearman correlations of the LS decoder and CCA filters, when using an SS-CV approach. The 'CCA+number' notation reflects subsequent filters generated by the CCA algorithm. The correlation trend is decreasing for these subsequent CCA filters. Furthermore, the LS correlations do not significantly differ from the corresponding CCA1 correlations.

In addition, the overall trend for both attention and movie correlations is decreasing for increasing CCA filter. The correlations of the first CCA filter do visually not differ much from the LS correlations and, mathematically, neither attention ($p=4.9 \cdot 10^{-1}$) nor movie ($p=2.6 \cdot 10^{-1}$) differ significantly. In fact, the difference in correlation between attention and movie condition does not differ significantly between the LS decoder and the first CCA filter ($p=8.5 \cdot 10^{-1}$).

An SS-V approach does not result in any significance for any CCA filter ($\min(p)=3.1 \cdot 10^{-1}$).

Discussion

As hypothesised, **the CCA filters do result in significant differences between the attention and inattention states**. Note that the fact that the third filter is not significant, whereas the fourth filter is, does not violate any theory. As mentioned in the hypothesis, the CCA filters try to maximise the correlation of the attention case under orthogonality constraints. However, this does not imply anything about the difference in correlation between the attention and inattention cases.

Furthermore, the decreasing trend of CCA correlation values is as expected, due to the addition of the orthogonality constraints. **Yet, surprisingly, the correlations of the CCA filters are not significantly different from the LS correlations. Apparently, time-lagging the stimulus envelope does not result in a significant performance gain.** Furthermore, this time lagging comes at the expense of a more complex model, such that the individual CCA filters might not be beneficial with respect to the LS approach. However, the entire CCA approach can still be advantageous with respect to the LS approach: CCA designs multiple filters of which the correlations can be optimally combined (in some sense) in a classifier. On the contrary, LS only yields one correlation value. Therefore, we will have to resume this discussion in chapter 8.

7.2.6 Significance of Kullback-Leibler divergence

Research questions

1. Does the Kullback-Leibler divergence late fusion approach reflect the level of attention to the auditory stream?
2. Does the Kullback-Leibler divergence early fusion approach reflect the level of attention to the auditory stream?

Hypothesis

Since the neural envelope is tracked less well in the inattention case (section 7.2.1), it is hypothesised that the KLD is larger for the inattention than for the attention case. Indeed, the KLD is an envelope tracking feature extractor, and a larger KLD value reflects less correspondence in distribution. Furthermore, we hypothesise the early fusion approach to be more discriminating than the late fusion approach since, in the former, information is already aggregated prior to computing the KLD.

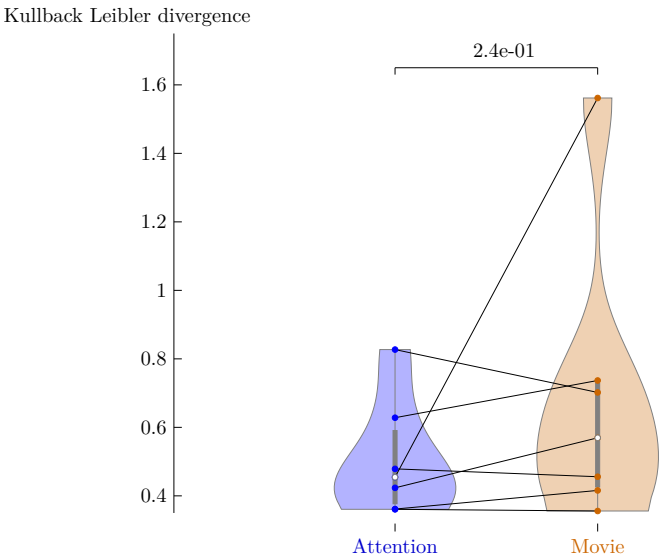
Experiment

Both the late and early fusion KLD approaches are applied to the dataset.

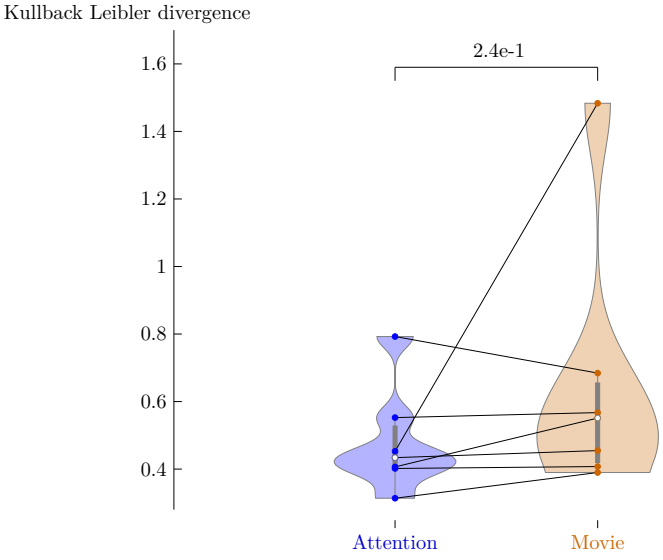
Result

As presented in figure 7.7, the median movie KLD visually exceeds the attention one, when using an SS-CV approach. This observation holds regarding both the late and early fusion approaches. However, despite 4/7 subjects attaining a larger KLD value in the late-fusion movie case, no significance is obtained ($p=2.4 \cdot 10^{-1}$). Similarly, KLD values of the early fusion approach increase for 6/7 subjects from the attention to the movie case. Yet, again, no significance is obtained ($p=2.4 \cdot 10^{-1}$) since most increases happen only marginally.

On the contrary, significance is obtained using an SS-V approach, both regarding the late fusion ($p=1.8 \cdot 10^{-5}$) and the early fusion ($p=7.3 \cdot 10^{-5}$) approaches.



(A) The KLD late fusion approach results in larger KLD values for 4 out of 7 subjects. Nevertheless, no significance is obtained.



(B) The KLD early fusion approach results in larger KLD values for 6 out of 7 subjects. Nevertheless, no significance is obtained.

FIGURE 7.7: KLD methodologies, using an SS-CV approach.

7.2.7 Discussion

Overall, the inattention median KLD values visually exceed the attention KLD values. **However, the differentiating nature of the results is inconsistent regarding the SS-CV and the SS-V approaches: The SS-V approach does result in differentiating results, whereas the SS-CV approach does not.** Undifferentiating results could originate from the noise component (i.e. background EEG) overwhelming the desired component (i.e. neural tracking information) in the PSD estimation, such that the neural tracking information might not be sufficiently encoded in this PSD estimate. Furthermore, the finite length of the data introduces edge effects in the PSD estimation, which are similar for all PSDs and therefore work as a similarity force. Nevertheless, the differentiating results in the SS-V approach might indicate that the method is viable.

To gain more insight into the operation of the novel KLD metric, the topographic maps of the filter weights, using an SI-CV approach, are displayed in figure 7.8. Mostly, large weights in the frontal(-right temporal) region are observed. We would have expected these high weight values to be structured around the temporal sensors at both sides of the head, as was the case for the LS decoders (section 7.2.3). Therefore, it is not entirely clear what data characteristics the novel KLD method is exactly exploiting.

Further, no proof is found for the hypothesis of a more discriminating early fusion approach. Indeed, a hypothesis test on the difference in attention-inattention KLD between both approaches does not result in significance ($p=5.5 \cdot 10^{-1}$).

Finally, we note that these results might come with an interesting byproduct: As explained in section 3.3, the KLD approach assumes that the frequency content of the stimulus envelope is preserved in the EEG signal. Since significance is obtained using the KLD approach in an SS-V evaluation setting, we can at least not reject this hypothesis. Furthermore, this assumption is in line with the brain locking to the modulation frequency of sine waves [51].

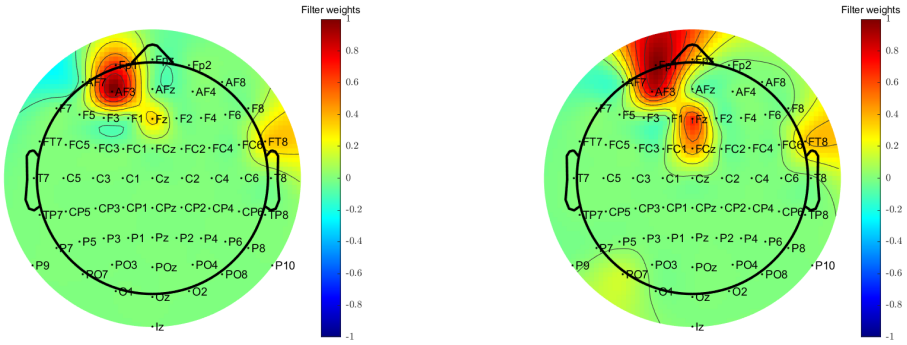
7.2.8 Significance of band-power

Research questions

1. Does the band-power reflect the level of attention to the auditory stream for any frequency band-sensor combination?
2. Does the band-power reflect the level of attention to the auditory stream for any frequency band in Broca's or Wernicke's areas?

Hypothesis

Given the discussion in section 3.4, we hypothesise to find a difference in parietal-occipital and frontal regions in the alpha and beta bands. Although, again, since



(A) KLD late fusion (B) KLD early fusion

FIGURE 7.8: Topographic map of the KLD weights, under an SI-CV approach. Both fusion methods result in large weight values in the right frontal-temporal region.

these features have not yet been applied to our setup, results might differ. The closest work is [54], where a subject was informed to either pay attention to a beep tone or to a flashing screen in a compound beep-flash task. In that work [54], a higher band-power has been observed in the parietal-occipital region in the alpha band regarding subjects attending the beep tone.

In addition, since Broca's and Wernicke's areas (figure 3.2) are related to language processing, we hypothesise to find a difference in band-power in these regions [56, 57].

Experiment

The band-power is calculated for each sensor region (see figure 3.3) and frequency band combination, and also in Broca's and Wernicke's areas (see figure 3.2).

Result

No significant difference in band-power is obtained for either sensor region-frequency band combination, both using the SS-CV ($\min(p)=1.4 \cdot 10^{-1}$) and using the SS-V ($\min(p)=5.9 \cdot 10^{-1}$) approaches. Similarly, no significance is observed in Broca's (SS-CV: $\min(p)=2.0 \cdot 10^{-1}$ and SS-V: $\min(p)=5.9 \cdot 10^{-1}$) or Wernicke's (SS-CV: $\min(p)=5.3 \cdot 10^{-1}$ and SS-V: $\min(p)=7.6 \cdot 10^{-1}$) areas.

Discussion

Contrary to our hypothesis, **no significance is obtained in any frequency band, for any sensor group.** However, this does not necessarily mean no significance can be found using a brain activity exploiting feature extractor. Firstly, neighbouring EEG sensors pick up similar signals and EEG signals have a poor spatial resolution, such that it is difficult to localise the origin of a source. Secondly, EEG signals are bipotential signals, such that the reference also influences these power measurements.

Thirdly, as with the KLD experiment, the PSD calculation introduces edge effects, which are similar for attention and inattention cases, interfering with the setup. Finally, the band-power could just not be an appropriate measure to exploit brain activity differences in the current setup. **However, since ideally we would like to find universal measures of attention to an auditory stream, the band-power does not seem a suitable feature extractor.**

The fact that **there is no significance found in Broca's and Wernicke's areas** might be due to similar reasons. Nevertheless, setup specifics could also be in play: The subjects inattentive to the stimulus were watching a **subtitled** movie, which means language processing was involved in the inattention task as well [56].

7.2.9 Significance of common spatial pattern

Research questions

Does common spatial pattern reflect the level of attention to the auditory stream for any frequency band?

Hypothesis

Since classically the alpha and beta band are related to the concept of attention [58, 59], and since the beta band has been proven profound for good CSP performance in the AAD domain [12], differences in the alpha and beta band might be hypothesised. Nonetheless, we reemphasise that the target problem is not exactly the same as this AAD domain since multiple states of attention exist.

Experiment

The common spatial pattern approach is conducted for the delta, theta, alpha and beta band separately.

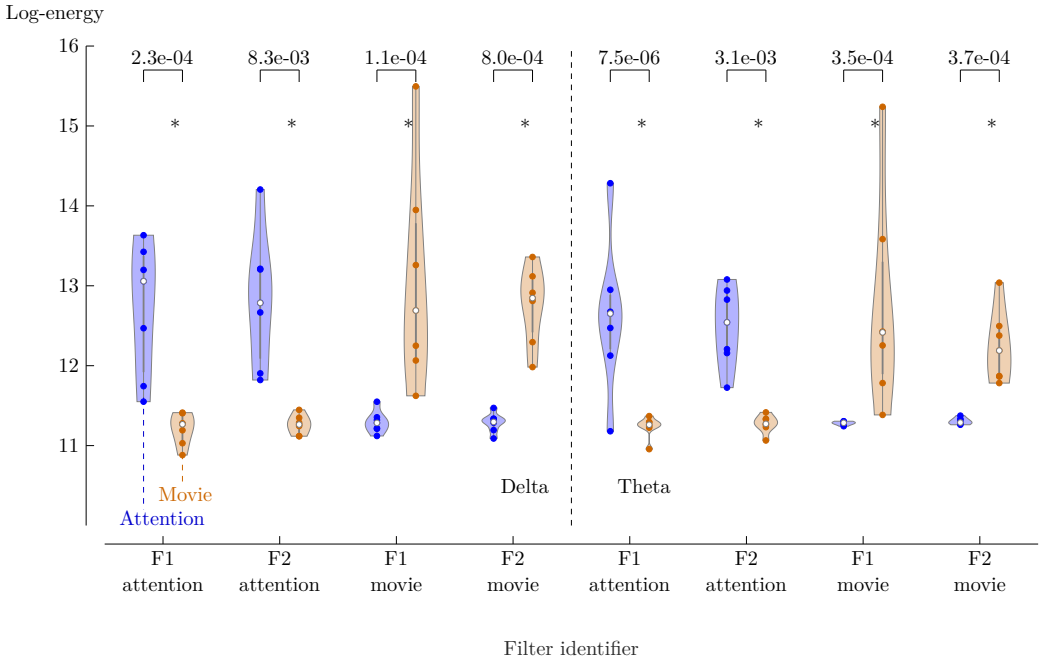
Result

Using an SS-CV approach, we obtain rather spectacular results: For the first 2 filters that maximise the output energy of the attention case and the first 2 filters that maximise the output energy of the movie case (see figure 7.9), significant results are obtained for all frequency bands ($\max(p)=8.3 \cdot 10^{-3}$). Also, visually, the attention and movie boxplots are well separated in figure 7.9.

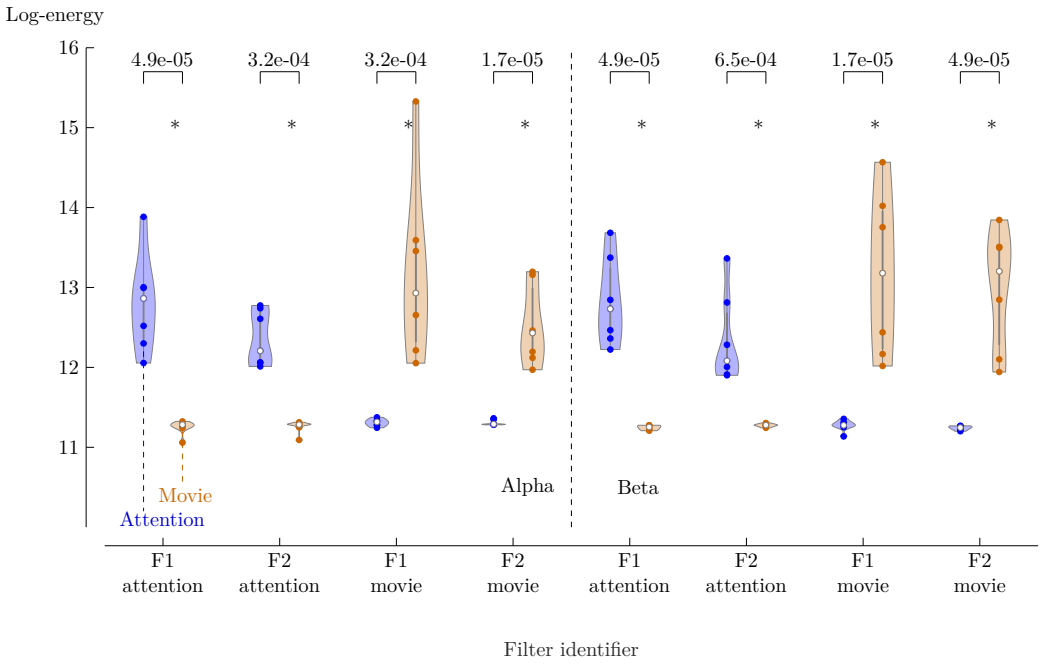
Nevertheless, no significance is obtained using an SS-V approach ($\min(p)=1.9 \cdot 10^{-1}$).

Discussion

CSP seems to be a promising feature extraction methodology since it results in major significance. Although results need to be interpreted



(A) Delta and theta band



(B) Alpha and beta band

FIGURE 7.9: The log-energy of the CSP filter output, using an SS-CV approach, shows significance for both the first two filters that maximise attention output energy and the first two filters that maximise the movie output energy. These filters are denoted by 'F + filter number + state of which the energy is maximised'.

with care:

Firstly, **an SI-CV approach does not result in significance**. Neither the first nor the second filter maximising the attention or movie output energy proves to be significant ($p=3.4 \cdot 10^{-1}$). This possibly denotes poor generalisation of CSP as a subject-independent feature extractor.

Secondly, we reemphasise the **'black-box' nature** of CSP: It is difficult to assess what data properties CSP is effectively exploiting, although muscle and eye blink artefacts were (largely) removed by the preprocessing framework. As a result, it is difficult for CSP to exploit these artefacts, yet not impossible since residual artefacts can still be present. Moreover, the pattern topographic plots, using an SI-CV approach (see section 3.7), may guide the interpretation. Regarding the delta band, structuring around the temporal lobes can be observed (see figure 7.10). On one hand, this may refer to the relationship between the delta band and the temporal lobes. On the other hand, this pattern is not entirely apparent in the filters maximising the movie output energy, and no clear patterns are observed for the other frequency bands. This means we still cannot give much interpretation to the CSP filters. The interested reader can find the first two topographic patterns, for both movie and attention, in appendix C, and this for each frequency band.

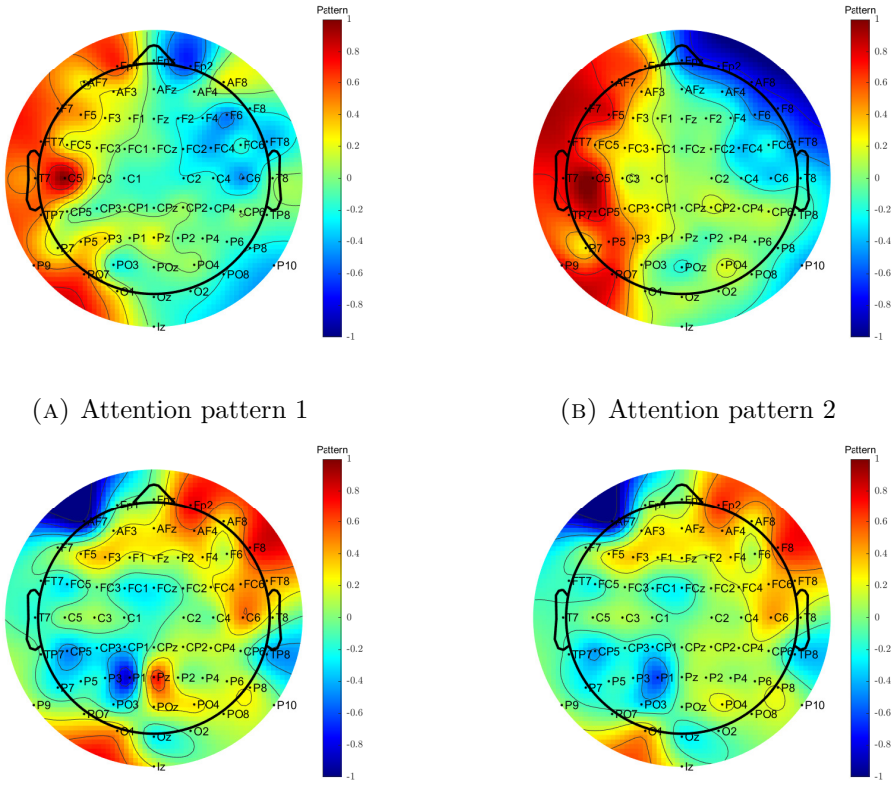
7.2.10 Significance of entropy

Research questions

1. Does the entropy reflect the level of attention to the auditory stream for any frequency band-sensor combination?
2. Does the entropy reflect the level of attention to the auditory stream for any frequency band in Broca's or Wernicke's areas?

Hypothesis

In [63], the entropy, calculated over the $[0.5, 25]$ Hz range, has been reported discriminatory between subjects, attentive to a screen projecting different colours, and subjects at rest. Likewise, the entropy in the $[8.5, 32]$ Hz range ($\approx$ alpha and beta band) has proven lower in passive than in auditory attention states [6]. However, since in the Vanthornhout dataset the distractor effectively corresponds to a visual **attention** condition, we will not necessarily find similar results [6]. Nevertheless, both alpha and beta bands are classically related to the concept of attention [12, 58, 59], such that we hypothesise to find differences in these frequency bands.



(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE 7.10: SI-CV pattern topographic plot of delta band CSP. Some structuring around the temporal lobes is visible.

Experiment

The entropy is calculated for each sensor region-frequency band combination and also in Broca's and Wernicke's areas[5].

Result

Significance is obtained in the RB ($p=4.1 \cdot 10^{-3}$) and RT ($p=3.3 \cdot 10^{-3}$) group in the beta band, when using an SS-CV approach. In both groups, the movie entropy values exceed the attention entropy values. Also, note that the RC, CC and LB group result in almost significance ($p=5.8 \cdot 10^{-2}$). No significant results are found in the other frequency band-sensor region combinations ($\min(p)=2.1 \cdot 10^{-1}$). Moreover, using an SS-V approach, no significance is obtained in any frequency band-sensor region combination ($\min(p)=6.7 \cdot 10^{-1}$).

[5]See figure 3.3 for the sensor group naming convention and figure 3.2 for the mapping of Broca's and Wernicke's areas to the EEG sensors.

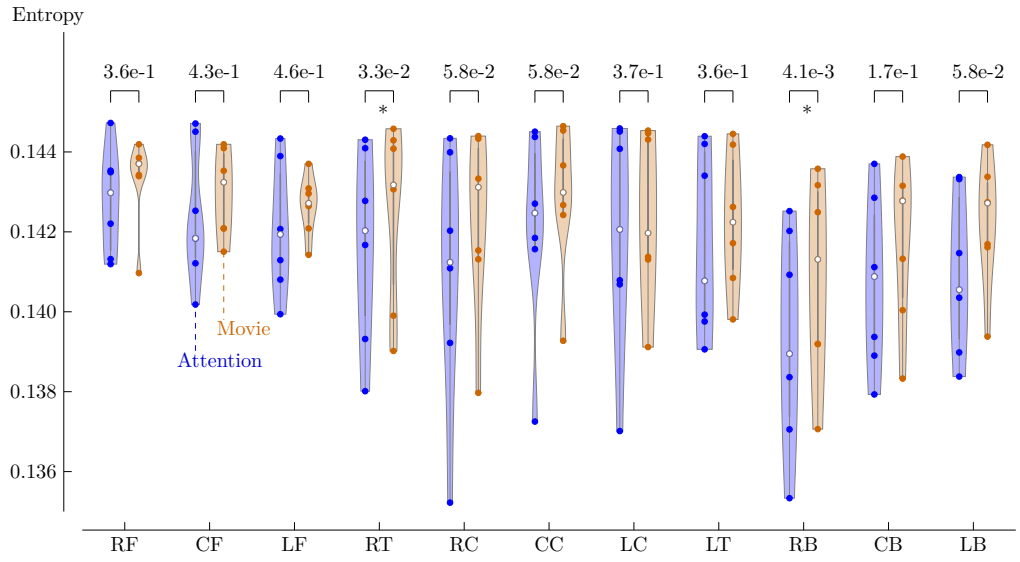


FIGURE 7.11: Entropy in the beta band, averaged over the sensors in each group, when using an SS-CV evaluation methodology. Significantly higher movie entropy values are obtained in the RT and RB group with respect to the attention condition. Group indices refer to the naming convention as introduced in figure 3.3.

In addition, no significance is obtained in Broca's (SS-CV: min(p)= $3.6 \cdot 10^{-1}$ and SS-V: min(p)=$6.7 \cdot 10^{-1}$)) or Wernicke's (SS-CV: min(p)=$5.8 \cdot 10^{-2}$ and SS-V: min(p)=$6.7 \cdot 10^{-1}$) areas, for both validation approaches.

Discussion

The entropy values are differentiating in the beta band. This result might not be so surprising since the beta band is classically related to the concept of attention [12, 58]. The inattention entropy values exceed the movie condition in the **RT and RB regions**, such that it seems that these sensor regions are affected by larger levels of distribution uncertainty in the movie condition. Yet, in [6], entropy values of auditory attentive subjects were found to exceed entropy values of subjects in a passive state. **Therefore, the relevant entropy values seem to increase from a passive state over the state of attention to an auditory stream to attention to a visual stream.**

Moreover, we note that the entropy in the RT and RB group cannot be directly mapped to cortical activity. Indeed, referring to section 7.2.8, EEG signals are bipotential and hence the mapping of sensor-regions to cortical activity is influenced by the reference signal. Furthermore, EEG has a poor spatial resolution.

As with the band-power feature, **no significance is obtained in Broca's or Wernicke's areas.** We refer to the poor spatial resolution of the EEG signal and the fact that there are **subtitles** present in the cartoon movie as possible explanations for this insignificance.

7.3 Conclusion

Table 7.1 provides a summary of the (in)significance of the feature extraction methodologies. Herein, CSP seems to be the most promising one, resulting in low p-values ($\max(p)=8.0\cdot10^{-4}$ for the first two attention and inattention filters). Nevertheless, the CSP performance needs to be interpreted with caution since CSP operates somewhat like a 'black-box' and does not seem to generalise well to subject-independent designs.

As can be seen in table 7.1, feature significance also depends on the validation strategy: Almost no significance is obtained on the sentences of the Flemish matrix test. Only the LS and KLD approaches prove to be differentiating. Yet, the LS significance might be explained by the finite length of the dataset (see section 7.2.1), such that only the KLD approaches truly seem to be differentiating. This general insignificance can possibly be explained due to the difficulty of subjects to perform the task as instructed. Indeed, these Flemish matrix-sentences do not provide a coherent story, which promotes interference in the setup.

The novel KLD early fusion approach is not yet fully understood. It comes with the interpretation of binwise linking of frequency information and seems mostly focused on the frontal-(right temporal) region. Nonetheless, no KLD significance is obtained using an SS-CV approach, whereas significance is obtained using an SS-V approach. Possibly, edge effects in the PSD estimation and poor SNR conditions, in terms of envelope encoding EEG versus background EEG, influence the results.

Furthermore, regarding the LS approach, lagged decoders do not significantly outperform spatial decoders. This paves the way for computationally cheap, spatial decoders.

Finally, note that the limited amount of attention conditions in the dataset hampers the drawing of general conclusions.

Significant	Insignificant
LS	LASSO
CCA	KLD late fusion
CSP	KLD early fusion
Entropy	BP

(A) SS-CV

Significant	Insignificant
LS	LASSO
KLD late fusion	CCA
KLD early fusion	BP
	CSP
	Entropy

(B) SS-V

TABLE 7.1: Summary of the feature extraction experiments: Features are classified according to their (in)significance between the attention and inattention states. Nevertheless, LS significance in the SS-V approach appears most likely due to finite dataset length effects, as explained in section 7.2.1.

Chapter 8

Feature extractor-classifier performance

8.1 Introduction

This chapter presents the experiments, linked to the second high-level objective: *What **classification performance** can be attained by a feature extractor-classifier combination?* To this end, we will study combinations of the chapter 3 features, in compound with the LDA classifier of chapter 4. Note that the LDA assumptions of normality and equal covariance matrices of the class conditional distributions do not seem to be entirely satisfied (see section 6.3 and appendix B.1), although this LDA classifier can still prove its worth [52]. Furthermore, classification performance will be evaluated for different window lengths. Larger window lengths indeed allow to incorporate more information in the decision making process, although this comes at the cost of a reduced latency of the decision itself.

Although not every feature has proven to yield significant differences between the attention and inattention states, classification performance-wise, it still might be useful to consider them in conjunction with other features. Indeed, features without classification power when considered alone might still add information in conjunction with other features [104]. In fact, only if two features fully correlate, no theoretical gain in classification performance is possible in comparison with the individual features, i.e., if features fully correlate, no complementary information can be exploited [104]. This motivates us to also include the LASSO and KLD features in this study.

Section 8.2 provides the low-level research questions and the accompanying experiments. Thereafter, section 8.3 concludes this chapter.

8.2 Experiments

Since an LDA classifier needs to be trained, the validation procedure of chapter 7 needs to be altered somewhat. To this end, figure 8.1 shows a graphical summary of

the feature extraction-classification strategy, as will be adhered to in this chapter.

Ignoring steps (2), (5) and (8) for now, the basic idea consists of the combination of an outer and inner 10 fold cross validation. The left-out outer fold serves as a validation set for the final trained feature extractor-classifier combination, and the left-in outer folds are utilised to train both the feature extractor and the classifier. To this end, the left-in outer folds can readily be leveraged as a training set for the feature extractor. This does however not hold true regarding the classifier, such that the inner-fold cross validation comes into play. Separate validation features are indeed required to train the classifier. For this purpose, the inner-fold cross validation consists of a 10 fold partitioning of the left-in outer folds. A feature extractor is trained on the left-in inner folds and applied to the left-out outer fold. By repeating this inner-fold cross validation, inner validation features are acquired that can subsequently be fed to the LDA classifier training algorithm. As a result, both a trained feature extractor and classifier are acquired.

In addition to this basic schematic, windowing is incorporated ((5) and (8)), as can be observed in figure 8.1. The validation folds are not utilised as a whole, but partitioned in non-overlapping windows. This step is required because the distribution of the validation feature values depends on the window size. In addition, normalisation using z-scoring (see section 4.4) proves necessary ((2), (5) and (8)). Indeed, due to the regularisation of the LDA classifier, the magnitude of the features becomes important. Not all features however operate at the same scale, hence the need for normalisation. This normalisation is implemented by prior windowing of the left-in outer folds to calculate the mean and standard deviation of the training data features. This mean and standard deviation are subsequently used to z-score both the windowed inner and outer validation folds, according to equation 4.6.

Using this evaluation strategy, the following experiments are conducted:

8.2.1 Feature extractor-classifier performance

Research questions[1]

1. To what extent does the number of CCA filters influence the classification performance?
2. Does CCA outperform the LS feature extractor in terms of classification performance?
3. To what extent does the frequency band choice influence the CSP classification performance?
4. Which combination of feature extractors results in the highest classification performance?
5. How do the feature extractor-LDA combinations perform in function of the window length?

[1] Of course all research questions only hold with respect to the LDA classifier.

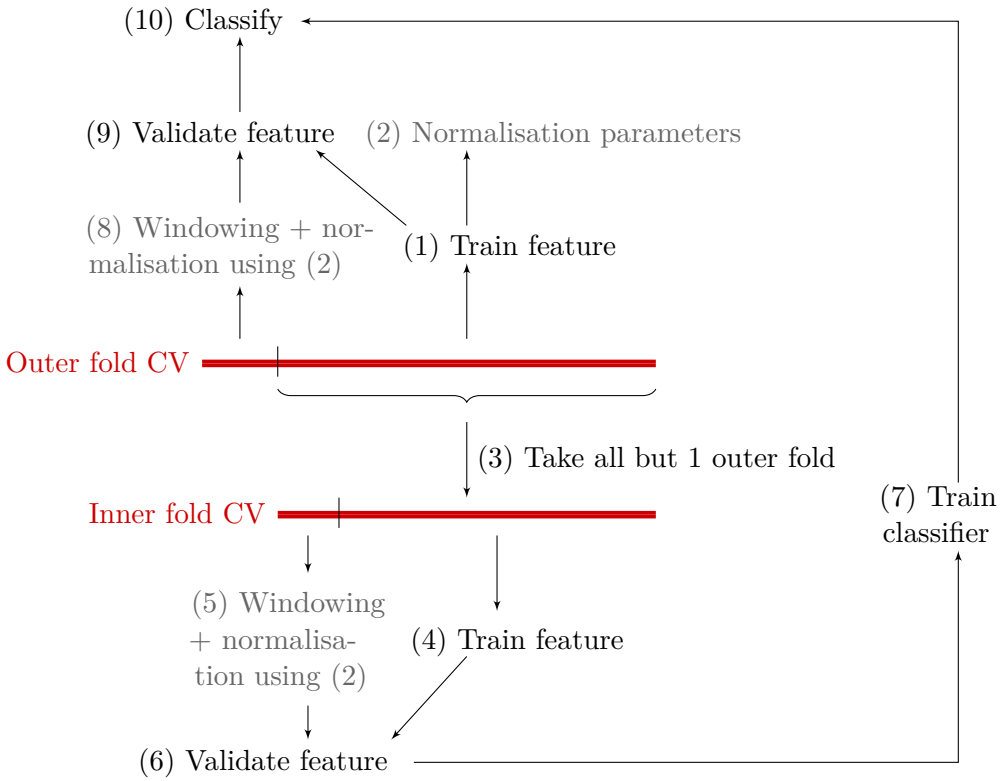


FIGURE 8.1: Schematic overview of the general classification strategy.

Hypothesis

Regarding the first research question, including more CCA filters does inherently result in, at least, an equal amount of information. However, an increase in the number of filters entails an increase in the dimensionality of the feature space, worsening the curse of dimensionality. Therefore, we hypothesise to not observe a monotonic increase in classification performance with an increasing number of CCA filters, but to find an optimal number of filters, in terms of classification performance, instead. Moreover, this curse of dimensionality has a bigger impact on the features with increasing window length since an increasing window length is associated with fewer feature data. Therefore, the optimal number of CCA filters might differ per window length.

Regarding the second research question and referring to section 7.2.5, CCA can be seen as a generalisation of LS due to CCA's more flexible model. In addition, multiple CCA filter correlations can be optimally combined (with respect to the LDA classifier). Therefore, we hypothesise the CCA approach to outperform the LS one, at least when the number of CCA filters is not too large to let the curse of dimensionality dominate performance.

Regarding the third research question and given the results in section 7.2.9, we hypothesise CSP's classification performance to exceed the performance of other feature extractors for each frequency band version of CSP. However, since the alpha and beta bands are classically related to attention [12, 58, 59], we hypothesise CSP to mostly rely on alpha and beta bands. Although, again, our setup is somewhat different from the setups in [12], [58] and [59], such that results might differ.

Regarding the fourth research question, again, the tradeoff between encoding more information and the curse of dimensionality needs to be taken into account. Nevertheless, we hypothesise feature combinations to outperform the individual features since different feature designs are based on different principles (e.g. neural envelope tracking versus brain activity) and are subsequently expected to encode complementary information with respect to the LDA classifier.

Finally, regarding the fifth research question, we hypothesise to find a tradeoff between the window length and the classification performance. Indeed, larger window lengths allow to incorporate more information and therefore are expected to yield an increase in classification performance. However, the curse of dimensionality and the interplay of normalisation and finite data lengths need to be taken into account as well.

Experiment[2]

The validation scheme, as summarised in figure 8.1, will be adhered to for window lengths 1, 5, 10, 20, 30 and 60 s and this for different feature extractor combinations and hyper-parameter settings (different CSP frequency bands and CCA number of filters). However, in order to keep the experiment tractable, the number of combinations is somewhat limited.

To this end, the CCA feature extractor is explored in isolation, for amount of filters ranging from 1 to 5. In this way, the influence of the number of CCA filters can be studied individually. In addition, the CSP feature extractor is studied in isolation, both for each individual frequency band (delta, theta, alpha and beta) separately and for the combination of all frequency bands. Moreover, in order to limit the curse of dimensionality, this experiment is limited to only one filter to maximise attention output energy and one filter to maximise inattention output energy. This study in isolation can be justified due to CSP's spectacular separation performance in the experiments of section 7.2.9. Indeed, combining CSP with other feature extractors might not be insightful due to the CSP's expected performance gain with respect to the other methods. Furthermore, regarding the KLD feature extractor, only the early fusion approach will be included since both fusion approaches seem to be based on the same information, as they both have large weights in the frontal(-right temporal) sensor region (see figure 7.8). Regarding the remaining left-in feature extractors (LS,

[2]Figures 8.2a-8.2c only provide a summary of the most important results for the sake of avoiding figure clutter. Nevertheless, the interested reader can find means and standard deviations of all considered feature combinations in appendix D.

LASSO, KLD early fusion and entropy), all 15 combinations are evaluated.

Regarding the hyper-parameter choice, we refer the reader to table 6.1. In addition, the beta-band entropy will be averaged over both the RT and RB groups (see figure 3.3), as both regions have proven differentiating in section 7.2.10. In addition, only the continuous story data are utilised, given that the results of chapter 7 suggest that task interference might have occurred in the matrix data.

To compare the classification performance of the features with a random classifier, the upper bound of a 95% one-sided confidence interval of a binomial distribution with success-rate 0.5 serves as chance level. Correspondingly, to compare methods with one another, the LME and ANOVA framework is applied to window lengths 10 s and 30 s, in combination with the Benjamini-Hochberg correction and a 5% significance level. A fixed effect on the accuracy and offset, and a random effect on the subject identifier are assumed.

Result

Figure 8.2a[3] displays the mean classification accuracies of the LS and CCA feature extractors in combination with the LDA classifier. Only the LS standard deviation is shown for the sake of avoiding figure clutter and since this standard deviation is representable for the CCA methods as well. Comparing the methods per window length, the mean accuracy of the 1-filter CCA approach exceeds the mean accuracy of the other ones at a window length of 1 s. Similarly, the 2-filter CCA approach attains the highest performance at window lengths 5, 30 and 60 s, and the LS approach does so at window lengths 10 and 20 s. Nevertheless, pairwise hypothesis tests with respect to the LS feature at the 10 s window show no significant differences ($\min(p)=7.1 \cdot 10^{-1}$) between the LS method and any CCA design, as is the case at the 30 s window ($\min(p)=7.1 \cdot 10^{-1}$). All methods lie above chance level.

Correspondingly, figure 8.2b displays the mean accuracies and accompanying standard deviations of the CSP feature extractor. As expected from section 7.2.9, mean classification accuracies all outperform chance level. In fact, the minimal mean accuracy for any frequency band and for any window size still equals 0.86 (theta band, 1 s window). Furthermore, all frequency band CSP classifies all cases correctly from a window length of 10 s onwards. Similarly, alpha band CSP achieves this maximal performance starting from a window length of 30 s and beta band CSP achieves this maximal performance at window lengths 20 and 30 s. At a window length of 10 s, all frequency band CSP significantly outperforms the individual CSP methods ($\max(p)=4.8 \cdot 10^{-2}$). At a larger window length of 30 s, all frequency band CSP still achieves this significance in comparison to delta ($4.7 \cdot 10^{-2}$) and theta ($4.7 \cdot 10^{-2}$) band CSP. However, insignificance is attained in comparison with alpha ($p=1.5 \cdot 10^{-1}$) and beta band CSP ($p=1.5 \cdot 10^{-1}$). In fact, this can also be seen

[3]This and subsequent line plots have been converted into tikz format using an unpublished matlab2tikz function written by S. Geirnaert as available on 11/11/2021.

visually in figure 8.2b, where either all frequency band CSP alone or all frequency band CSP together with alpha and beta band CSP achieve the highest classification accuracy. Moreover, alpha, beta and all frequency band CSP achieve lower standard deviation values than delta and theta band CSP.

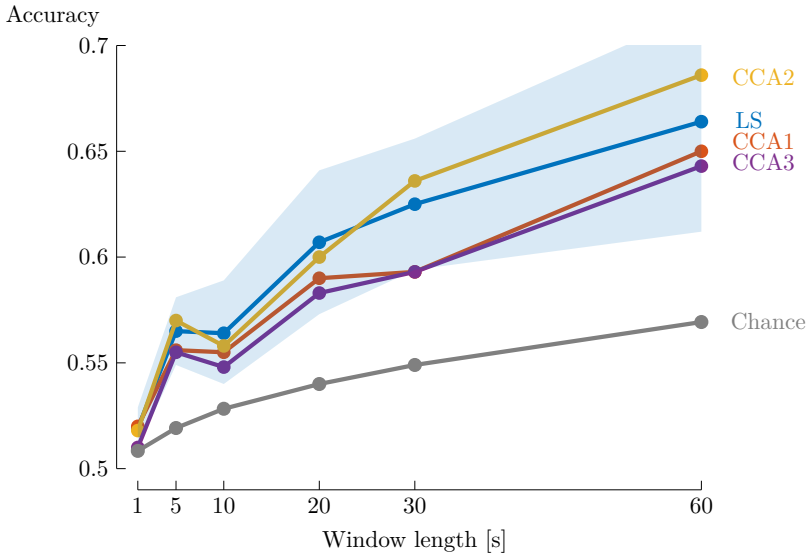
Next, figure 8.2c displays the mean accuracies and accompanying standard deviations of relevant LS, LASSO, KLD early fusion and entropy combinations. Whereas the KLD metric accuracy initially increases with increasing window length, the mean accuracy drops at the 20 and 30 s windows. The resulting mean accuracy nevertheless lies above chance level, and the KLD achieves higher mean accuracies at the 1, 5, 10 and 20 s windows than the LASSO, entropy and LS methods individually. On the minus side, the standard deviation of the KLD feature is about $0.5 \cdot 10^{-2}$ higher than for the other individual features. Next, the entropy feature outperforms chance level for every window length. Finally, in line with section 7.2.4, the LASSO feature does not prove to be discriminating between attention and inattention since the mean accuracy lies around chance level at every window length.

When combining the features, the LASSO-KLD-entropy combination yields the highest mean accuracy at window length 1 s, the LS-KLD-entropy combination yields the highest mean accuracy at window lengths 5, 10, 20 and 60 s and the KLD-entropy combination yields the highest mean accuracy at window length 30 s. However, in pairwise comparison with respect to the LS-KLD-entropy combination at the 10 and 30 s windows, the difference between these methods is insignificant ($\min(p)=9.4 \cdot 10^{-2}$). At the 10 s window, the LS-KLD-entropy combination furthermore proves insignificantly different from the KLD feature ($p=8.5 \cdot 10^{-2}$), whereas the combination proves differentiating with respect to the entropy feature ($p=4.3 \cdot 10^{-2}$). The reverse story is true at the 30 s window length, where the LS-KLD-entropy combination significantly outperforms the KLD feature ($p=8.5 \cdot 10^{-2}$), but does result in an insignificant p-value with respect to the entropy feature ($p=1.5 \cdot 10^{-1}$). Adding the LASSO feature to the mixture does not result in a significant gain in classification accuracy, both at window length 10 s ($\min(p)=2.0 \cdot 10^{-1}$) and at window length 30 s ($\min(p)=5.8 \cdot 10^{-1}$)

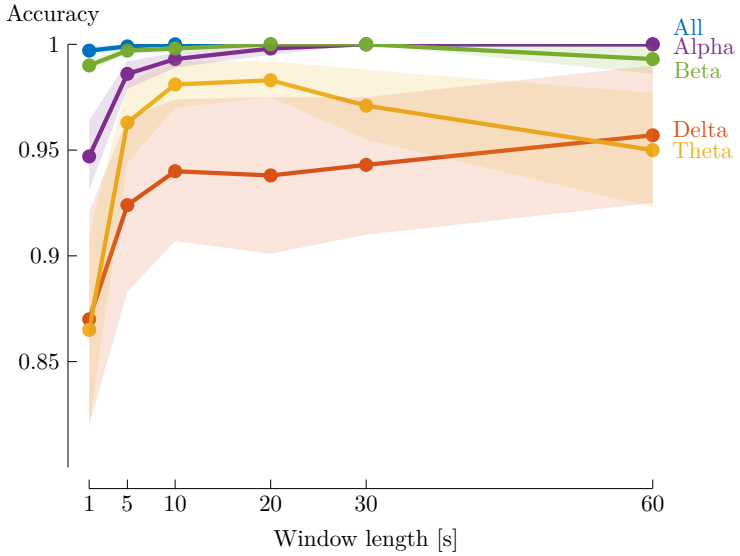
As can be seen in figures 8.2a, 8.2b and 8.2c, the increase in accuracy with increasing window length is fairly monotonic. Only regarding the KLD (based) features and the LASSO feature, this trend is absent. Moreover, in the figures 8.2a and 8.2c, generally an increase of $\approx 0.5 \cdot 10^{-2} - 10^{-1}$ in standard deviation is observed when increasing the window length from a 1 s window to a 60 s window. In figure 8.2b, however, the reversed trend is displayed, i.e., regarding alpha, beta and all frequency CSP, the standard deviation becomes smaller with increasing window length.

Discussion

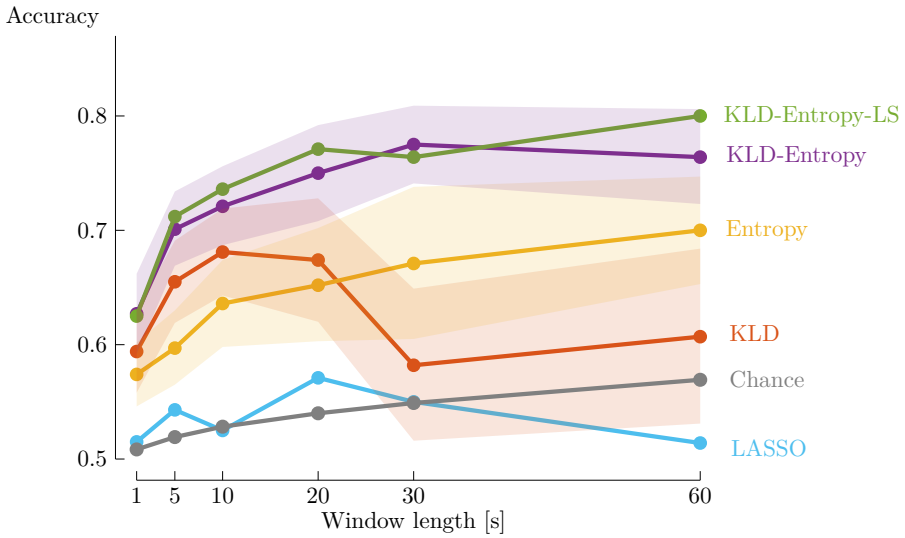
Regarding the first and second research questions, **no concrete optimal number of CCA filters can be found since neither CCA design does significantly differ with respect to the LS approach. Nevertheless, generally, simpler models**



(A) LS and CCA: CCAx ($x \in \mathbb{N}_0$) reflects subsequent filters generated by the CCA algorithm. LS and CCA2 result overall in the highest mean accuracies, although the methods do not significantly differ at window lengths 10 s and 30 s.



(B) CSP: All frequency band CSP versions attain performance well above chance level. Nevertheless, alpha, beta and all frequency band CSP visually attain larger mean accuracies than delta and theta band CSP. Chance level is not shown since it falls below the lower limit of the y-axis.



(c) Combinations of LS, LASSO, KLD and entropy: Accuracies of KLD-entropy combinations lie generally higher.

FIGURE 8.2: Classification accuracy of the feature extractors in function of the window length, given as mean accuracy (line) and standard deviation (shading) on the Vanthornhout dataset. 'Chance' denotes the upper bound of a 95% confidence interval of a binomial distribution with success-rate 0.5. Note that the accuracy scales (y-axes) differ between figures.

are preferred, such that the LS feature extractor might be favoured. Interestingly, this seems to denote that there is no additional information exploited when using the more flexible CCA model with respect to the LS methodology. Moreover, visibly, there does not seem to be a global trend of a decreasing favoured number of filters with an increasing window length. This can possibly be attributed to the regularisation mechanisms in the LDA classifier, which makes the regularised LDA classifier suitable for sparse feature spaces [79].

Regarding the third research question, **all CSP frequency bands significantly outperform chance level**. These results are indeed in line with our hypothesis and the results of section 7.2.9. Nevertheless, the classification results are complementary: Firstly, as hypothesised, **all frequency band CSP does not significantly differ from alpha and beta band CSP at the 30 s window, whereas significance is achieved with respect to delta and theta band CSP**. Secondly, **the high accuracies ($> 85\%$) are maintained at low window lengths.**

Referring to figure 8.2b, CSP seems to benefit from including all frequency bands, although this combination requires a feature dimension of 8 instead of 2. Possibly, the regularisation of the LDA classifier mitigates the effect of the increased feature dimensionality from 2 to 8. In addition, we note that this combination does not

significantly outperform the alpha and beta band CSP algorithms. Referring to section 7.2.9, **results nonetheless need to be interpreted with care. Due to the 'black-box' nature of CSP, it is not (entirely) clear what data properties CSP is effectively exploiting.** In fact, CSP can possibly be artefact driven, despite muscle and eye blink artefacts being (largely) removed using MWF filters (see section 2.2).

Regarding the fourth research question, **the LS-KLD-entropy combination yields the highest mean accuracy for 4/6 windows**. Nevertheless, this combination does not significantly outperform the KLD-entropy combination, both at the 10 s and at the 30 s window. This LS-KLD-entropy combination does however outperform the individual entropy feature at a window length of 10 s, and the individual KLD feature at a window length of 30 s. **Since simpler models are generally preferred, the KLD-entropy combination thus seems the favourable combination**. Given that the KLD feature did not result in any significance in section 7.2.6, it might be surprising to encounter the KLD feature in this favourable combination. This can be explained by the fact that **the KLD metric yields accuracies well above chance level for smaller window lengths, despite experiencing a large drop in mean accuracy at larger window lengths ($>$ 20 s)**. Possibly, this behaviour is caused by the curse of dimensionality, although given that this drop in accuracy does not occur for any other feature (combination), there are likely other (still unknown) factors in play. In addition, we note that the standard deviation of the KLD (based) features lies somewhat higher ($\approx 0.5 \cdot 10^{-2} - 10^{-1}$) than the standard deviation of the other feature combinations.

Regarding the fifth research question, **a tradeoff between decision time and accuracy is generally observed**, i.e., larger window lengths allow to incorporate more information and hence generally result in higher classification accuracies. Yet, this trend is not exactly monotonic, possibly due to the interplay of finite data length, normalisation procedures and the curse of dimensionality. No such trend is furthermore observed for the LASSO feature since this feature lies around chance level for every window length. In addition, as mentioned in the previous paragraph, the KLD feature also does not follow this monotonic behaviour since it experiences a drop in accuracy at larger window lengths ($>$ 20 s). Furthermore, except for the CSP metric, the standard deviation is generally seen to increase with increasing window length, possibly due to the curse of dimensionality.

8.2.2 Validation on the Brouckmans-Dewit-Vanhaelen dataset

To better assess architecture generalisation over different attention conditions, the proposed KLD-entropy feature combination and the CSP methods are validated on the Brouckmans-Dewit-Vanhaelen dataset as follows:

Research questions

1. What classification performance can be attained by the proposed **KLD-entropy** combination on the **Brouckmans-Dewit-Vanhaelen dataset**?
2. What classification performance can be attained by **CSP** on the **Brouckmans-Dewit-Vanhaelen dataset**?

Hypothesis

Regarding the first research question, similar performance to the KLD-entropy combination on the Vanthornhout dataset (section 8.2.1) might be expected. The mathematics and text conditions in the Brouckmans-Dewit-Vanhaelen dataset[4] do nevertheless differ from the movie condition in the Vanthornhout dataset, such that results might differ. Indeed, brain regions could possibly be less or more involved depending on the distractor condition.

Regarding the second research question, we hypothesise to find similar effects as in section 8.2.1, namely higher classification accuracies for CSP than for the KLD-entropy combination and CSP mainly exploiting alpha and beta band properties. However, CSP could exploit different patterns in the Brouckmans-Dewit-Vanhaelen dataset than in the Vanthornhout dataset since the distraction conditions differ, and CSP could possibly be artefact driven.

Experiment

CSP and the KLD-entropy combination are applied to the Brouckmans-Dewit-Vanhaelen dataset according to the validation procedure, as summarised in figure 8.1. Within this procedure, the mathematics and text conditions are concatenated and considered as one inattention condition. We justify this approach since ideally feature extractors should be able to discriminate the auditory attention condition from any other condition, although the relative influence of the mathematical distraction condition is of course limited in this setup.

As in section 8.2.1, this experiment is repeated for window lengths 1, 2, 5, 10, 20, 30 and 60 s. The entropy feature is averaged over both RT and RB groups and the other hyper-parameters are fixed according to table 6.1. Furthermore, hypothesis tests are performed using a framework of LME, ANOVA and Benjamini-Hochberg corrections, and a significance level of 5%. Chance level is computed as the upper bound of a 95% one-sided confidence interval of a binomial distribution with success-rate 0.5.

[4]Due to a strong baseline drift of the EEG data in this Brouckmans-Dewit-Vanhaelen dataset, the EEG is firstly linearly detrended, resampled to 256 Hz and highpass filtered (Chebychev type 2 where an attenuation of 80 dB attenuation is attained at the frequency 10% outside the passband) using a cut-off frequency of 0.5 Hz before being fed into the preprocessing framework of chapter 2.

Result

Figure 8.3a displays the classification accuracies of CSP. All frequency band CSP achieves the highest mean accuracy over all window lengths. At the 10 s window, this all frequency band CSP is slightly insignificant with respect to delta and beta band CSP ($\min(p)=8.8 \cdot 10^{-2}$), and also insignificant with respect to theta and alpha band CSP ($\min(p)=1.3 \cdot 10^{-1}$). Similarly, at the 30 s window, all frequency band CSP is only slightly insignificant with respect to beta band CSP ($\min(p)=8.8 \cdot 10^{-2}$) and more clearly insignificant with respect to the other individual CSP bands ($\min(p)=1.3 \cdot 10^{-1}$). Moreover, all frequency band and beta band CSP respectively achieve standard deviations $\approx 1 \cdot 10^{-1}$ and $\approx 5 \cdot 10^{-2}$ lower than delta, theta and alpha band CSP, of which the alpha band standard deviation in figure 8.3a is representative.

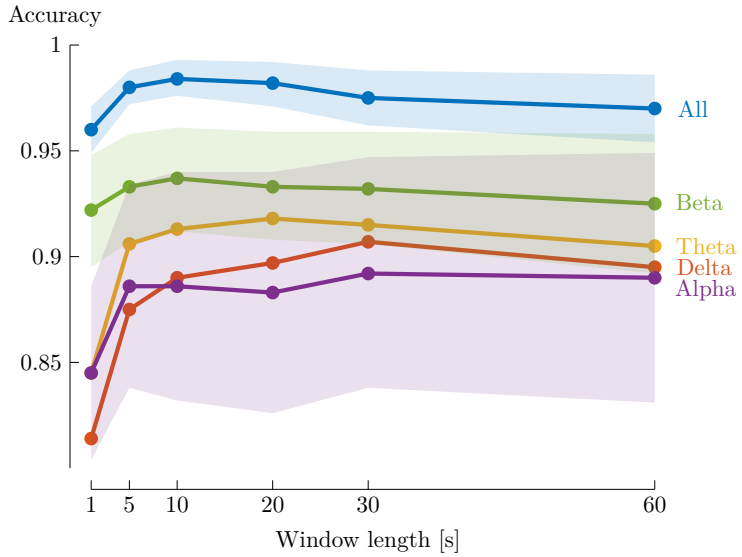
Similarly, figure 8.3b shows the performance of the KLD-entropy combination, as well as the performance of the individual features. As before, both features attain mean accuracies well above chance level. In addition, the KLD-entropy combination significantly outperforms the individual features both at the 10 s ($\max(p)=3.0 \cdot 10^{-2}$) and at the 30 s ($\max(p)=4.7 \cdot 10^{-2}$) window. Standard deviations of all methods yield comparable results of $O(10^{-2})$.

Appendix D contains tables with the numeric mean accuracies and accompanying standard deviations for the interested reader.

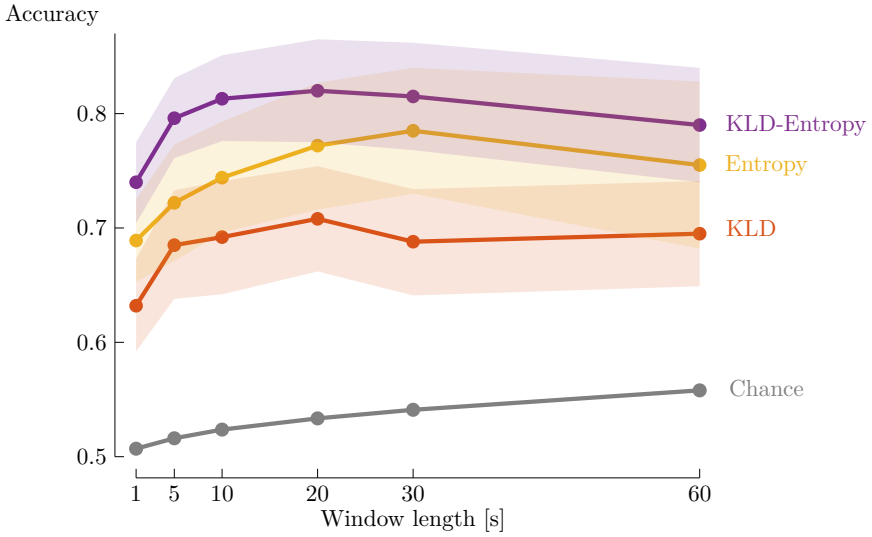
Discussion

Regarding the second research question, all frequency band CSP classification accuracies are in line with the results of section 8.2.1 since a mean accuracy of 0.98 is attained from a window length of 5 s onwards using all frequency band CSP. Furthermore, these accuracies are accompanied by a small standard deviation of $O(10^{-2})$. CSP seems again to mainly exploit beta band properties and, as before, the mean accuracy of delta band CSP generally lies below the other frequency band CSP curves. Yet, the alpha band does not seem as prominent as in the Vanthornhout dataset. In the hypothesis tests with respect to all frequency band CSP, beta band CSP nevertheless attains the lowest p-values ($p=8.8 \cdot 10^{-2}$), presumably due to lower standard deviations in beta band CSP than in the other individual CSP bands.

The current setup is interesting from the viewpoint of **architecture generalisation**: The CSP architecture with retraining per condition seems to generalise well over conditions. Indeed, at the 10 s window, all frequency band CSP attains a mean accuracy of 1 in the Vanthornhout dataset and a mean accuracy of 0.98 in the Brouckmans-Dewit-Vanhaelen dataset. Repeating the experiment with only the mathematical distraction condition as inattention data yields a mean accuracy of 0.97 at the 10 s window. Similarly, repeating the experiment when utilising only the text distraction condition as inattention data yields a mean accuracy of 0.99.



(A) CSP: All frequency band CSP attains the highest mean accuracy for all frequency bands, although the accuracies at the 10 and 30 s windows are not significantly different ($\min(p)=8.8 \cdot 10^{-2}$, $\max(p)=2.4 \cdot 10^{-1}$) with respect to the individual frequency band CSP methods. Chance level is not shown since it falls below the lower limit of the y-axis.



(B) Combinations of KLD and entropy: All methods readily outperform chance level. The KLD-entropy combination significantly outperforms the individual features at window lengths 10 and 30 s ($\max(p)=4.7 \cdot 10^{-2}$).

FIGURE 8.3: Classification accuracy of the feature extractors in function of the window length, given as mean accuracy (line) and standard deviation (shading) on the Brouckmans-Dewit-Vanhaelen dataset. 'Chance' denotes the upper bound of a 95% confidence interval of a binomial distribution with success-rate 0.5. Note that the accuracy scales (y-axes) differ between figures.

While an in-depth **generalisation of the model parameters** requires a study on its own, the current setup permits for some intuition: Inspecting the individual mean accuracies per fold shows that chance level is surpassed for each fold in each frequency band. This also encompasses the case where CSP is almost exclusively trained on attention versus text and solely evaluated on attention versus mathematics. The accuracy in that validation fold nevertheless seems to drop with respect to the other folds since the mean accuracy in that fold lies a factor 1.07 lower than the mean accuracy over all folds.

Finding correspondences between the topographic pattern plots of CSP on the Vanthornhout dataset and the Brouckmans-Dewit-Vanhaelen dataset nevertheless proves difficult. The interested reader can find these figures in appendix C.

The KLD-entropy combination attains similar performance as in the Vanthornhout dataset. In fact, the individual entropy and KLD features achieve mean accuracies well above chance level. Contrary to the Vanthornhout dataset, the KLD feature achieves similar standard deviations as the entropy feature and there is no large dip in mean accuracy.

As with CSP, the current setup allows to study **architecture generalisation**:

- The KLD-entropy combination and individual features surpass chance level in both datasets, indicating generalisability of the model architecture.
- Regarding the entropy, a regular 10 fold cross validation shows that in all separate distraction conditions, i.e., movie, text and mathematics, the mean inattention entropy in the RB and RT groups surpasses the corresponding auditory attention entropy. This further indicates generalisability of the model architecture.

Again, although a study of the **generalisation of model parameters** falls outside the scope of this study, some intuition can be distilled from the current experimental setup:

- When solely classifying based on the entropy, the mathematical validation fold attains an accuracy a factor 1.18 below the mean accuracy over all folds at the 10 s window. Nevertheless, except at the 60 s window, the accuracies of that mathematical validation fold still surpass chance level.
- Classifying solely using the KLD metric shows classification accuracies not surpassing chance level from the 20 s window onwards on the mathematical validation fold. Windows below 20 s nevertheless still outperform chance level.

Moreover, both CSP and the KLD-entropy combination achieve high performance at low window lengths: All frequency band CSP and the KLD-entropy combination respectively attain a mean accuracy of 0.96 and 0.74 at the 1 s window length.

In summary, both the CSP and KLD-entropy combination achieve significant performance on both the Vanthornhout dataset and the Brouckmans-Dewit-Vanhaelen dataset, which is indicative of well generalising model architectures across attention conditions. In order to draw general conclusions, these architectures should however be applied to more types of attention conditions as well. The CSP model parameters themselves seem in addition to perform less well across different attention conditions, although CSP still seems to significantly outperform chance level in that case. The same does not seem to be valid regarding the KLD-entropy feature combination.

8.3 Conclusion

In line with chapter 7, the CSP feature extractor seems to be the most promising. Both in the Vanthornhout dataset and in the Brouckmans-Dewit-Vanhaelen dataset, all frequency band CSP attains mean accuracies exceeding 96% over all windows from 1 s to 60 s. Training CSP mainly on the text inattention condition and validating on the mathematics inattention condition seems to indicate that CSP generalises over different conditions since classification accuracies surpass chance level. These results need, however, to be interpreted with caution due to the CSP's 'black-box' nature and the limited scope of this experiment.

Furthermore, no concrete number of CCA filters is seen to be optimal, in the sense that no number of CCA filters yields a significantly different performance with respect to the LS feature.

Regarding the LS, LASSO, KLD early fusion and entropy combinations, the KLD-entropy combination seems the favourable choice since this KLD-entropy model does not perform significantly worse than any other feature extractor combination on the Vanthornhout dataset, and it is the simplest model to do so. Validation on the Brouckmans-Dewit-Vanhaelen dataset seems to confirm that this KLD-entropy combination performs well above chance level. Model parameters mainly trained on the text condition do nonetheless not seem to generalise well to the mathematical condition.

Moreover, a general trend is visible of an increasing accuracy with increasing window length, although the standard deviation generally also increases. This trend is, nevertheless, absent regarding the KLD feature extractor on the Vanthornhout dataset. Moreover, the LDA assumptions do not seem to be entirely satisfied (see appendix B). Finally, note that the limited number of attention conditions in the dataset hampers the drawing of general conclusions.

Future work could focus on further investigating the generalisability of the feature parameters across conditions by, e.g., leveraging the text versus mathematics condition in the Brouckmans-Dewit-Vanhaelen dataset.

Chapter 9

Performance unsupervised algorithms

9.1 Introduction

This chapter presents the experiments, linked to the third high-level objective: *How can the classic machine learning framework be* ***converted*** *into an* ***unsupervised*** *one?* Subject-specific feature extractors generally attain higher performance than subject-independent ones [8]. However, the performance of these feature extractors might degrade over time, e.g., due to displacements of the EEG cap, malfunctioning sensors and changing background EEG activity [13]. In addition, collecting subject-specific data requires resources, posing practical constraints. In order to mitigate these problems, we study unsupervised domain adaptation methodologies (CCA-DA, PCA-DA, SA-DA, TPCA-DA and D-DA) as described in chapter 5.

Section 9.2 details the experiments and section 9.3 the associated conclusions.

9.2 Experiments

Ideally, we would have access to subject-specific source and target EEG data, recorded at a different time. However, no such dataset is available at present time and could no longer be recorded within the timeframe of this thesis. Therefore, we will focus on the adaption of feature extractors, trained on one subject (the source subject) in which we are currently not interested in but for which we have access to labelled data, to another subject (the target subject) in which we are currently interested in but for which we have only access to unlabelled data.

Operation-wise, the CCA-DA methodology requires an equal amount of source and target domain data due to the application of the QR-SVD based CCA algorithm as described in [87, 88]. Therefore, the SI-CV approach cannot readily be applied. To nevertheless cope with this additional constraint, the validation methodology is altered according to figure 9.1. The availability of labelled EEG data of one subject

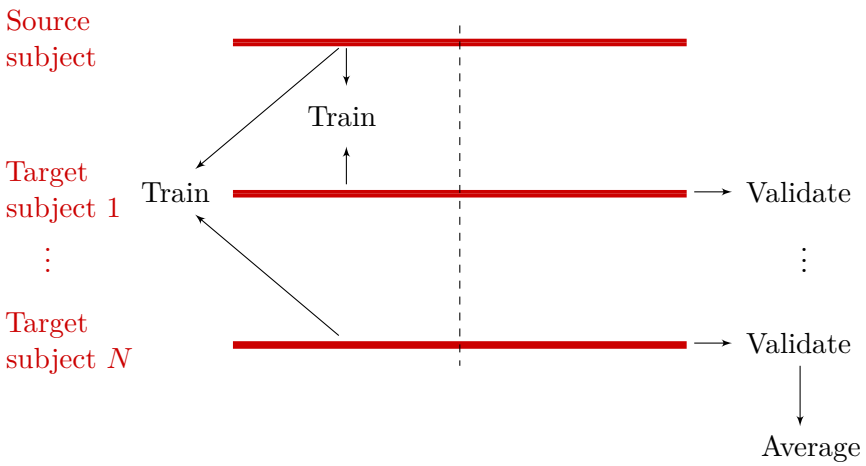


FIGURE 9.1: Schematic overview of the unsupervised feature extraction evaluation strategy using $N \in \mathbb{N}_0$ target subjects.

(source domain) and unlabelled EEG data of another subject (target domain) is assumed. Using a 2 fold cross validation approach, half of the data are used as a training set and the other half are used as a validation set. This process is repeated for different target subjects, such that each source-target subject combination is endeavoured.

Using this validation methodology, the following experiment is conducted:

Research questions

1. How does the domain adapted feature extractor compare to a supervised feature extractor, trained on **source** data?
2. How does the domain adapted feature extractor compare to a supervised feature extractor, trained on **target** data?
3. How does the domain adapted feature extractor compare to the **state-of-the-art unsupervised, iterative** feature extractor design?

Hypothesis

Since the domain adaptation methodologies essentially boil down to adding constraints to the feature extractor design, we hypothesise domain adaption to perform less well than supervised target design, i.e., subject-specific designs. For the same reason, we expect the unsupervised iterative design to outperform the domain adapted designs [75]. However, due to the joint embedding enforcement of domain adaptation, we hypothesise these domain adapted feature extractors to outperform the source trained feature extractors.

Experiment

Although the DA methodologies are inherently not restricted to the LS feature extractor, we will limit ourselves to it for the sake of clarity and because the LS feature can be readily transferred to other fields, such as the AAD domain [8, 15]. To this end, the procedure as illustrated in figure 9.1 is repeated for the different DA methodologies (CCA-DA, PCA-DA, SA-DA, TPCA-DA and D-DA). As a baseline, LS decoders are trained, in a supervised way, on both labelled source and labelled target data. As before, these source and target trained decoders are regularised using PCA as designed on the attention source, respectively, attention target EEG data. Note however that supervised target training is not feasible in practise since it requires labelled target data. In addition, the ILS procedure serves as a benchmark because it is the state-of-the-art unsupervised LS method in the AAD domain [75]. Regarding the hyper-parameters, the reader is referred to table 6.1. Hypothesis tests are performed using a framework of LME, ANOVA and Benjamini-Hochberg corrections. A fixed effect on the correlation value and offset, and a random effect on the subject identifier are assumed. Significance level equals 5%.

Result

Figure 9.2 displays the correlation levels, attained by the baseline and domain adaptation procedures. The CCA-DA approach significantly outperforms the source trained decoders for the attention case ($p=1.6 \cdot 10^{-2}$), and does almost so for the movie case ($p=8.2 \cdot 10^{-2}$). On the contrary, the PCA-DA and TPCA-DA approaches do not significantly differ from the source trained designs ($\min(p)=1.2 \cdot 10^{-1}$), and the SA-DA approach even performs significantly worse ($\max(p)=3.4 \cdot 10^{-2}$) from these source trained designs. CCA-DA correlation levels additionally significantly exceed the PCA-DA ones ($\max(p)=3.7 \cdot 10^{-2}$).

The D-DA approach also performs significantly worse than the source trained decoders ($\max(p)=1.6 \cdot 10^{-2}$). Importantly, these D-DA correlations cannot be interpreted absolutely, yet only relative to the source correlations. Indeed, referring to section 6.3, the discriminator weight ν has been tuned to only yield a significant direction of correlation shift. In this light, the source domain correlations exceed the D-DA correlations for 7/7 subjects in both the attention and the movie state.

In addition, apart from the CCA-DA movie correlations ($p=8.4 \cdot 10^{-2}$), all domain adapted correlations prove significantly inferior to their target trained counterparts ($\max(p)=2.6 \cdot 10^{-2}$).

The ILS baseline procedure significantly outperforms source training ($\max(p)=2.8 \cdot 10^{-6}$) and does not significantly differ from target training ($\min(p)=2.8 \cdot 10^{-1}$). Correspondingly, the ILS correlation levels significantly exceed the levels of the CCA-DA procedure ($\max(p)=2.1 \cdot 10^{-4}$).

Nevertheless, for practical purposes, the difference in attention-movie correlation is mainly important. As a baseline, the target trained decoders attain a mean correlation difference (attention-movie) 5.37 times larger than the source trained decoders. This correlation difference is nonetheless insignificant ($p=1.7 \cdot 10^{-1}$). On the contrary, the ILS procedure does achieve a significant difference in correlation with respect to source training ($p=3.3 \cdot 10^{-2}$), despite only attaining a mean correlation difference, with respect to this source training, of 2.75. The CCA-DA procedure attains a similar mean correlation difference of 2.40 with respect to source training, even though this difference is not significantly different from source training ($p=3.1 \cdot 10^{-1}$). Finally, the PCA-DA procedure achieves a worse mean correlation difference by a factor 1.52 in comparison with source training, although this correlation difference does not significantly lie below source training ($p=3.1 \cdot 10^{-1}$).

Lastly, we note that the domain adaptation procedures can also be used in conjunction with the iterative procedure. To this end, a joint framework of CCA-DA and the iterative procedure further increases the mean correlation difference with respect to the source domain to a factor 4.40 (compared to a factor 5.37 of target versus source training). This increase in correlation is furthermore significant with respect to the difference in correlation for both the ILS method ($p=3.3 \cdot 10^{-2}$) and the CCA method ($p=3.3 \cdot 10^{-2}$).

Discussion

On one hand, the CCA-DA procedure attains significantly lower attention correlations in comparison to target training, and the CCA-DA mean correlation difference is a factor 2.24 lower than its target trained counterpart. On the other hand, the CCA-DA procedure (almost) significantly improves source trained correlations (attention: $p=1.6 \cdot 10^{-2}$, movie: $p=8.2 \cdot 10^{-2}$), and achieves a mean difference in correlation only a factor 1.15 worse than the state-of-the-art ILS procedure. On the contrary, **PCA-based approaches do not seem to improve performance over source training**: the PCA-DA and TPCA-DA correlation levels do not significantly differ from source training, and SA-DA[1] even performs significantly worse than source training. The mean difference in correlation for the PCA-DA approach also is a factor 1.52 worse than source training. **Therefore, CCA seems to be the more appropriate dimensionality reduction mechanism for domain adaptation and shows the potential to replace PCA as the conventional regularisation technique.**

Nevertheless, given the fact that target training significantly outperforms the CCA and PCA based domain adaptation methodologies, there still seems to be a loss of neural-tracking information in the dimensionality reductions. Indeed, CCA tries to maximise the correlations between source and target domain and PCA tries to maximise the variance, but these conditions do not guarantee the preservation of

[1] In [82], it is advised to z-score the data. However, this does not yield any difference in trend in the current setup such that this z-scoring is neglected.

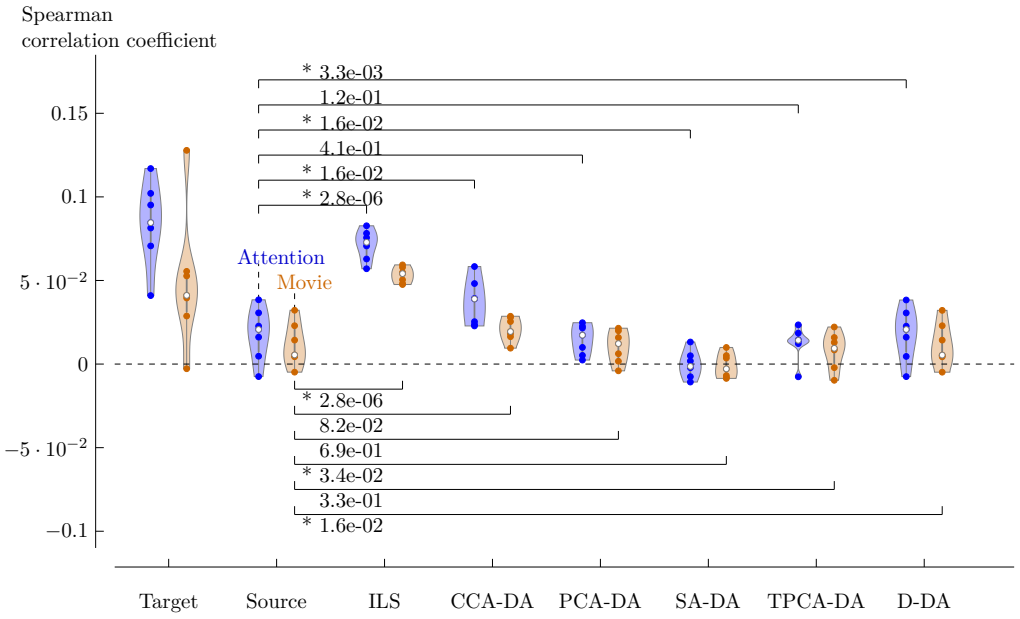


FIGURE 9.2: The spearman correlation of LS decoders, trained on labelled source data, labelled target data, trained in an iterative manner (ILS) and trained on labelled source data with domain adaptation (CCA-DA, PCA-DA, SA-DA, TPCA-DA and D-DA). The CCA-DA methodology significantly improves the attention source correlations and almost significantly the movie source correlations. PCA-DA and TPCA-DA do not significantly differ from source training. The SA-DA and D-DA methodologies perform significantly worse than source domain training.

neural tracking information. Since CCA-DA nonetheless outperforms the source domain training, relevant neural tracking information seems to be preserved in the CCA transform. The same does not seem to hold with respect to PCA.

Regarding the discriminator approach, **D-DA performs significantly worse than both source and target decoders.** Possibly, the LS feature extractor is not flexible enough to incorporate the discriminator term, such that both attention and inattention correlations are forced to zero.

The ILS procedure still significantly outperforms any tested domain adaptation methodology and does not significantly differ from target training, for both the attention and the inattention states ($\min(p)=2.8 \cdot 10^{-1}$). **Nevertheless, the difference in mean correlation between the CCA-DA and ILS approach is only a factor 1.15**. In addition, domain adaptation theoretically provides a general framework applicable to any feature. On the contrary, the ILS procedure cannot readily be extended to all other feature extraction methodologies as is. Combining CCA-DA with the ILS procedure also seems to further boost the mean difference in

correlation to a factor 4.40.

Finally, we note that the theoretical mitigation for unsupervised classifiers does not seem to entirely hold regarding the CCA-DA procedure: Validation on the left-out source fold reveals mean attention and movie correlation levels respectively a factor 2.21 and 2.03 larger than the mean correlations on the left-out target folds. As a result, classifiers calculated on the source validation set cannot be readily transferred to the target validation set.

9.3 Conclusion

Although domain adaptation performs worse than target training in terms of mean correlation difference, the CCA-DA approach outperforms source trained decoders, both in terms of mean correlation difference and in terms of individual correlation levels. In addition, the PCA based domain adaptation approaches do either perform significantly worse than regular source training (SA-DA) or do not significantly differ from it (TPCA-DA, PCA-DA). This paves the way for CCA based domain adaptation and for CCA to replace PCA as a regularisation technique. Nevertheless, there still seems to be a loss in neural tracking information when performing both CCA and PCA based dimensionality reductions, albeit to a lesser extent regarding CCA. Concerning the D-DA method, the LS feature extractor possibly is not flexibly enough to cope with the discriminator term.

Whereas the ILS procedure achieves the highest correlation levels for both the attention and inattention states, the mean difference in correlation between the CCA-DA and ILS procedure is only a factor 1.15. In addition, the CCA-DA and ILS procedures can also be combined, further increasing the mean correlation difference with respect to source training to a factor 4.40, in comparison with a factor 5.37 in target versus source training, a factor 2.40 in CCA versus source training and a factor 2.75 in ILS versus source training.

Finally, note that the drawing of general conclusions is hampered by the limited number of attention conditions in the dataset.

Chapter 10

Boosting auditory attention decoding

10.1 Introduction

This chapter presents the experiments, linked to the fourth high-level objective: *Does the **selection** of **high attention** time frames **improve** the **auditory attention decoding** (AAD) framework?* Consequently, we will investigate whether a joint auditory attention selection and AAD framework can improve AAD performance.

In an AAD setup, the goal is to select the attended speaker in a multispeaker environment. To this end, generally, subject-specific LS decoders are trained on the attended stream according to the procedure of section 3.3.2 [8, 15]. At inference, this decoder is applied to the EEG data and compared to the ground truth envelope of all auditory streams. Referring to section 3.3, attended streams result in higher levels of envelope tracking and hence result in higher correlation levels. Therefore, attended and unattended streams are detected based on the magnitude of the correlation levels.

Thus, a subject is instructed to focus on a specified auditory stream. This subject might, however, not always be effectively attentive to any auditory stream, interfering with the setup: A subject could be distracted by, e.g., a visual stream. Therefore, it might be useful to filter out these inattentive periods before AAD decoder training.

Section 10.2 presents the experiments and section 10.3 concludes this chapter.

10.2 Experiments

Ideally, we would have a dataset at our disposal wherein the same subjects attend both AAD experiments and experiments concerning auditory attention with a variety of distractor conditions. Such a dataset does, unfortunately, not exist at present time and could not be recorded within the timeframe of this thesis, such that we utilise the AAD dataset of [15]. This dataset consists of the EEG data of 16 normal-hearing

subjects, who were exposed to a two-speaker acoustic environment and instructed to attend one of the auditory streams while ignoring the other one. In total, approximately 72 minutes of data is available per subject. The interested reader can find the details of this setup in [15]. For conformity reasons with the attention data, this AAD dataset is preprocessed similarly to the scheme detailed in chapter 2[1].

The closest related experiment is performed in [6]. There, high attention segments were selected using an entropy based measure. They report increased AAD training correlations on the high attention training data with respect to the low attention training data. However, no such effect is reported on the validation set. Our experiment is nevertheless both more extensive and complementary to [6]: Firstly, we have investigated a broader range of feature extractors in chapters 7 and 8, although not all features can be included in the experiment due to the lack of subject-specific joint AAD and auditory attention versus inattention datasets, as will be detailed infra. Secondly, we will look explicitly at the correlations between the attended and unattended track on the validation set to compare decoders solely trained on high attention data with decoders trained on the entire training set. Therefore, our setup more closely resembles the true operation of an AAD enhanced framework.

Research questions

Does the AAD performance, i.e., the difference in correlation between attended and unattended streams, increase when only using high auditory attention segments at training time?

Hypothesis

We hypothesise to find an increased difference in envelope tracking (i.e. an increased difference in correlation) between the attended and unattended stream when only leveraging the high attention segments as a training set. By filtering out the inattention segments, we expect the AAD decoder to be better tailored towards the attended stream, hence increasing the correlation difference between the attended and unattended streams.

Experiment

Subject-specific LS AAD decoders are trained in a 10 fold SS-CV manner. Herein, an inner 10 fold cross validation is leveraged to select the high attention segments. In other words, attention features are trained on the left-in inner folds and applied on windows of 60 s to the left-out inner fold. After completing this inner cross validation, each 60 s segment has been leveraged once as validation set and has received an accompanying attention feature value. These feature values can subsequently be

[1]The EEG data in the AAD dataset are only available after prior downsampling to 128 Hz and highpass filtering with cutoff frequency 0.5 Hz. To this end, the downsampling blocks in the preprocessing framework of figure 2.1 were removed. This slightly alters the preprocessing framework since downsampling blocks and filtering blocks cannot be switched as such [105].

rank-ordered from high auditory attention (low rank value) to low auditory attention (high rank value). The 75% segments for which the sum of the associated rank labels over all features is the lowest form the attention-enhanced dataset. After the inner fold cross validation, subject-specific AAD decoders are trained on this attention-enhanced dataset, in a 10 fold SS-CV manner. In addition, training on the entire left-in outer folds is performed to serve as a practical baseline measure. Another baseline is created by 3 times randomly selecting 75% of the training data and averaging the resulting correlations. This allows to compare the attention-enhanced trained decoder to a decoder trained on the same amount of data.

Nonetheless, due to the unavailability of an accompanying subject-specific attention dataset, this experiment setup comes with the following assumptions/limitations:

- The attended stream is assumed to contain sufficient attention segments to reliably train attention feature extractors that need an attention training set (e.g. LS, CCA and KLD), although we acknowledge that an iterative procedure could provide a way out (e.g. [75, 79]): In each iteration a feature extractor could self-label the training data, on which a new feature extractor could subsequently be trained. However, since this iterative procedure would require convergence studies and experiments on its own, this approach is not pursued.
- Validation data are not partitioned in attention versus inattention segments. Therefore, it is assumed that LS AAD correlations are higher for enhanced AAD decoders, regarding both attentive and inattentive segments. Referring to section 3.3 and chapter 7, this assumption seems to hold true.
- The attention metrics all have the same weight in this attention selection mechanism, which might be suboptimal.
- In the AAD experiment subjects were instructed to attend an auditory stream. Therefore, some level of auditory attention might still be expected at any moment in time, whereas in practise subjects might completely ignore any auditory stream. In addition, in the Vanthornhout dataset, subjects were instructed to completely ignore the auditory stream. This difference in the level of auditory attention might interfere with the setup.
- Possibly the features could also be used to discriminate between attended and unattended segments in an AAD setting (e.g. [75]). Therefore, it might be the case that segments wherein subjects briefly attend the other auditory stream are filtered out as well. Whereas removing these segments is not the goal of our study, the end result of having a feature extractor better tailored to one subject remains the same and seems subsequently of less practical importance.

Furthermore, due to the unavailability of an accompanying subject-specific attention dataset, the subset of available features is rather limited. Indeed, such an attention dataset is required for CSP to be trained accurately (see section 3.4.3). As noted supra, an iterative procedure could possibly form a way out, although this would require convergence proofs and experiments on its own, such that CSP is not included in this experiment. CCA is also not included since this method does not significantly outperform LS, despite being a generalisation of the methodology (see section 8.2).

Moreover, LASSO is excluded, since, referring to section 8.2, the LASSO feature does not add significant information to the feature combinations. Finally, the entropy feature cannot readily be included: Sections 7.2.10 and 8.2 suggest that the relevant entropy values in a visual, mathematical and text reading attention state exceed the relevant entropy values in an auditory attention state when using the RT and RB groups in the beta band. On the contrary, in [6] and [63], the relevant entropy values are found to be higher in an auditory attention state than in a passive state when leveraging respectively alpha and beta bands [6], and delta, theta, alpha and beta bands [63]. As a result, entropy thresholds seem required to discriminate between the attention states and these thresholds are currently not available. Therefore, only KLD early fusion and LS features are included. We refer the reader to table 6.1 for the hyper-parameters of the respective methods. Hypothesis tests are performed using a framework of LME, ANOVA and Benjamini-Hochberg corrections, wherein a fixed effect on the correlation and offset, and a random effect on the subject identifier are assumed. Significance level equals 5%.

Result

Figure 10.1 shows the Spearman correlations of AAD decoders, trained on the attention-enhanced subset, the averaged random subset and the entire training set. Although the attention-enhanced trained decoders result in a higher correlation with respect to decoders trained on the entire training set in 9/16 attended stream cases and in 10/16 unattended stream cases, no significant differences in correlation are observed. Indeed, both for the attended ($p=3.0 \cdot 10^{-1}$) and unattended ($p=3.3 \cdot 10^{-1}$) stream, there is no evidence for the attention-enhanced trained decoders to significantly outperform the decoders trained on the entire training dataset. Similarly, the attention-enhanced trained decoders do not significantly outperform the decoders trained on a random subset of data for the unattended stream ($p=3.0 \cdot 10^{-1}$). However, regarding the attention state, the attention-enhanced trained decoders are on the verge of significantly ($p=5.5 \cdot 10^{-2}$) outperforming the randomly trained decoders.

Regarding the practically important difference in correlation (attended-unattended), the attention-enhanced trained decoders do not significantly outperform the decoders trained on the entire training data ($p=2.8 \cdot 10^{-1}$). However, the attention-enhanced trained decoders do so with respect to the randomly trained decoders ($p=1.7 \cdot 10^{-2}$).

Discussion

Since the difference between attended stream and unattended stream correlations using the attention-enhanced data subset significantly differs from random selection ($p=1.7 \cdot 10^{-2}$), leveraging high attention segments seems beneficial. As hypothesised, auditory attention selection seems to result in AAD decoders better tailored to the attended stream. However, while this comparison is fair in terms of training data size, it is not a practical one: In practise, one would use either the attention-enhanced trained decoders or decoders trained

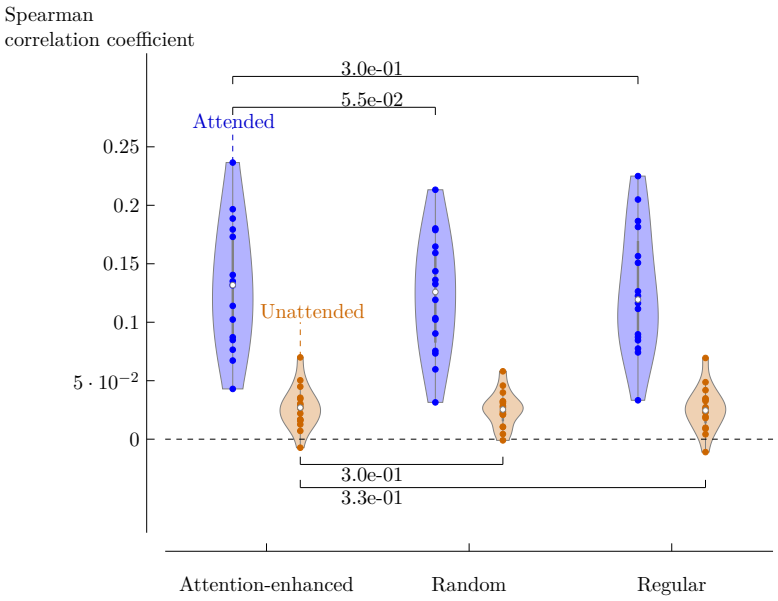


FIGURE 10.1: Spearman correlation for AAD decoders, trained on the attention-enhanced dataset (Attention-enhanced), trained on a random 75% partitioning of the training data (Random) and trained on the entire training dataset (Regular). This attention-enhanced training does not result in a significant improvement in attended stream or unattended stream correlations with respect to either a random partitioning or entire dataset training, although it is on the verge of achieving so regarding the random attention state.

on the entire dataset. In this light, **solely training on the attention-enhanced subset does not result in a significant increase in correlations with respect to decoders trained on the entire training set. Nonetheless, training on the attention-enhanced dataset could still be beneficial since it seems to reduce the amount of training data and accompanying computational cost of AAD training, without significant loss of AAD performance.** The attention mechanism of course still has a computational cost itself. Finally, the higher defined assumptions (see supra) should also be taken into account, such that the attention selection does likely not achieve its full potential in the current setup.

10.3 Conclusion

Attention selection results in a significant increase in the correlation difference between attended and unattended streams with respect to random selection. This suggests that auditory attention selection results in AAD decoders better tailored to the attended stream. However, auditory attention selection does not significantly improve correlation differences in comparison with training on the entire dataset. This auditory attention selection could nevertheless be beneficial in that case since it seems to allow for a training set size reduction without significant loss in performance.

Chapter 11

Conclusion and future work

11.1 Conclusion

The general goal of this thesis dissertation was to ***decode*** *whether a subject is in a* ***state of auditory attention***. To this end, different features have been investigated. Furthermore, the conversion of these features into unsupervised ones and the application of the features in an auditory attention decoding (AAD) framework have been studied. For this purpose, a dataset concerning auditory versus visual attention has been leveraged, as well as a dataset concerning attention towards audio, mathematical exercises and texts for validating final performance.

As a first objective, the performance of different feature extractor methodologies has been investigated. To this end, two feature extractor combinations are ultimately proposed: the combination of the novel Kullback-Leibler divergence (KLD) based feature with the entropy, and common spatial pattern (CSP).

The novel KLD based feature exploits neural envelope tracking properties and seems complementary to the entropy with respect to a linear discriminant analysis (LDA) classifier. Adding, the established least squares (LS) feature to this combination, however, does not result in a significant increase in performance, although LS and its generalisation canonical correlation analysis (CCA) individually are able to discriminate between attention and inattention to an auditory stream.

On the other hand, common spatial pattern (CSP) seems to be the most promising feature, yielding the highest classification accuracies with respect to other feature extractors. Results need, nevertheless, to be interpreted with caution. Although artefacts were (largely) removed, CSP could still be artefact driven, or CSP could exploit other dataset-specific properties.

Importantly, both the KLD-entropy combination and CSP architectures seem to generalise well across the datasets. Other feature explorations, involving band-power (BP) and least squares decoder sparsity using a least absolute shrinkage and selection

operator (LASSO) penalty, did not result in a proof for discriminating capabilities between attention and inattention to an auditory stream.

As a second objective, the conversion of the feature extractors to unsupervised ones has been investigated. To this end, the least squares feature extractor was focused upon due to its applicability in other domains. It was proposed to leverage domain adaptation methods, with the goal to adapt a supervised trained LS decoder (source subject) to another subject (target subject), in order to mitigate the need for subject-specific data collection while alleviating the performance gap between source trained and target trained feature extractors. A combination of CCA, as an instance of this domain adaptation method, with the state-of-the-art iterative least squares (ILS) design is suggested to this end.

Individually, the CCA method does achieve significantly worse correlation levels than the ILS method. Nevertheless, the practically important mean difference in attention-inattention correlation reaches similar levels, with only a difference of a factor 1.15 between both methods. The addition of CCA to the ILS method still seems beneficial. Indeed, the proposed combination significantly outperforms the mean attention-inattention difference in correlation with respect to both individual methods, indicating their complementarity.

Apart from CCA, also PCA-based approaches were explored to create a joint EEG subspace. These PCA based methods did not prove to significantly outperform the unadapted decoders, whereas CCA did. This result paves the way for CCA based domain adaptation and shows potential for CCA to replace PCA as the default regularisation technique.

Another flavour of domain adaptation was explored as well, namely a procedure consisting of a discriminator term, i.e., adding a source-target classifier to the feature extractor design in order to create a subspace at the feature value level. This has resulted in a novel expression for a discriminator LS-LDA based method. However, the adapted LS decoder did perform significantly worse than the unadapted decoder. The least squares feature extractor is possibly not flexible enough to cope with this discriminator adaptation method. Nevertheless, this methodology could still prove its worth combined with other features.

As a third objective, the application of the attention selection in an AAD framework has been investigated. Training AAD LS decoders solely on the highest attention segments did seem to significantly outperform decoders trained on a random subpart of the data in terms of differences in correlation between attended and unattended streams. Furthermore, no proof was found for a significantly different attended-unattended stream correlation difference with respect to regular training on the entire dataset. This seems to pave the way for data reduction possibilities, whilst not compromising on performance. However, this setup did not allow to readily apply subject-specific CSP and entropy features, such that this experiment possibly

did not attain its full potential.

Finally, note that results need to be interpreted with caution since the utilised datasets do not contain all possible attention states, such that results might be difficult to generalise. Nevertheless, overall, this thesis dissertation has taken a step towards the ultimate goal of combining an unsupervised auditory attention selection and an unsupervised AAD framework with applications in neurosteered hearing devices.

11.2 Future work

Foremost, future research could be focused on collecting larger datasets with more attention states in order to better generalise conclusions, although the Brouckmans-Dewit-Vanhaelen dataset could be further explored as well. In addition, an attention dataset could be recorded, wherein subjects are exposed to different environments to allow for the validation of unsupervised subject-tailored methodologies. Accompanying subject-specific AAD datasets could be acquired as well.

In addition, future work could further investigate the generalisability of the feature extraction methods over different inattention conditions.

Regarding the features, we identify three directions of possible future research: Firstly, lexical and linguistic features, such as phoneme onset, cohort entropy and word onset, show great potential within the attention context [106, 107, 108]. Indeed, in [106], it has been demonstrated that only the phoneme onset and cohort entropy of the attended stream are tracked in an AAD setup. Secondly, features from chaos theory [109, 110] (e.g. correlation dimension [111] and Lyapunov exponents [110]) and features from fractal theory [110, 112] (e.g. detrended fluctuation analysis [113] and boxcounting dimension [114]) could be interesting to explore. These metrics are popular in, e.g., sleep applications (e.g. [115, 116]) and could be utilised to assess whether certain brain regions exhibit fractal or chaotic properties in an auditory attention state. Finally, deep learning based approaches form another direction to explore. To this end, existing networks, relating EEG and auditory streams, could serve as a starting point [117, 118, 119].

Regarding the classification stage, the classifiers could also be seen as a hyperparameter to tune. The LDA assumptions indeed do not seem to be entirely satisfied (see appendix B). To this end, the support vector machine (SVM) classifier could be explored since this classifier transforms the features before linearly separating the feature space [52, 120]. In addition, neural network based classifiers exhibit great flexibility to separate feature spaces [64].

Regarding the conversion to unsupervised algorithms, the proposed unsupervised methodologies of chapter 5 could be applied to other features (e.g. see 3). Further-

more, other domain adaptation methodologies could be explored [76]. In this light, e.g., unsupervised neural network approaches could be leveraged [76, 78].

Regarding the combination of auditory attention selection and AAD, ultimately, unsupervised auditory attention selection and unsupervised AAD could be combined in a practical neurosteered hearing device.

Appendices

Appendix A

Non-convexity proofs

Convex optimisation problems are desired since each local minimum of such a problem corresponds to a global minimum [44]. It is thus useful to assess the convexity of the novel data-driven, early fusion KLD approach and the novel least squares-linear discriminant analysis discriminator-domain adaptation (D-DA) method. We will first review some convexity properties (section A.1) and thereafter leverage them to show that the KLD early fusion optimisation problem is in fact non-convex (section A.2). Similarly, we prove that the least squares-linear discriminant analysis D-DA approach is non-convex in section A.3.

A.1 Convexity properties

The following theorems define convexity and describe its properties [44]:

Theorem A.1.1 (Convexity optimisation problem) *An optimisation problem is convex if and only if the constraints form a convex set and the cost function is a convex function.*

Theorem A.1.2 (Convex set) *A set C is convex if it satisfies the following condition:*

$$\theta x + (1-\theta)y \in C,\ \forall x, y \in C,\ \theta \in [0,1].$$

Theorem A.1.3 (Convex function) *A function f: $C \to \mathbb{R}$ is convex if it satisfies the following condition:*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y),\ \forall x, y \in C,\ \theta \in [0,1].$$

Theorem A.1.4 (Positive semidefinite Hessian and convexity) *If the Hessian $\nabla^2 f(x)$ of a function $f(x)$ is positive semidefinite $\forall x \in C$, then $f(x)$ is convex over C.*

A.2 Early fusion Kullback-Leibler divergence

The cost function of the early fusion KLD feature extractor can be found in equation 3.21 (section 3.3), but is repeated here for the sake of clarity:

$$\begin{aligned} \min_{\mathbf{a}} \quad & KLD\left(\frac{\sum_{c=1}^{C} a(c)\bar{\mathcal{P}}_{X_c,n}(f)}{\sum_{f=f_1}^{f_2}\sum_{c=1}^{C} a(c)\bar{\mathcal{P}}_{X_c,n}}(f)||\bar{\mathcal{P}}_{Y,n}(f)\right) \\ \text{s.t.} \quad & \mathbf{a} \geq \mathbf{0}_{C\times 1} \\ & ||\mathbf{a}||_2 = 1. \end{aligned} \tag{A.1}$$

We leverage theorem A.1.4 and derive the Hessian of the KLD early fusion cost function to evaluate its positive semidefiniteness. In fact, we will show that equation A.1 is non-convex by means of a counterexample. To ease the derivation, the following notation is introduced:

$$\begin{aligned} C &= \sum_{f=f_1}^{f_2} h(f) log_2\left(\frac{h(f)}{\bar{\mathcal{P}}_{Y,n}(f)}\right); \\ h(f) &= \frac{\sum_{c=1}^{C} a(c)\bar{\mathcal{P}}_{X_c,n}(f)}{\sum_{f=f_1}^{f_2}\sum_{c=1}^{C} a(c)\bar{\mathcal{P}}_{X_c,n}(f)}; \\ \alpha(f) &= \sum_{c=1}^{C} a(c)\bar{\mathcal{P}}_{X_c,n}(f); \\ \beta &= \sum_{f=f_1}^{f_2} \alpha(f). \end{aligned} \tag{A.2}$$

By applying the chain rule to the cost function of equation A.1, the first and second order derivatives of the cost function are retrieved:

$$\begin{aligned} \frac{\partial C}{\partial a(i)} &= \sum_{f=f_1}^{f_2} \frac{\partial h(f)}{\partial a(i)} log_2\left(\frac{h(f)}{\bar{\mathcal{P}}_{Y,n}(f)}\right) + \frac{\partial h(f)}{\partial a(i)}\frac{1}{log(2)}, \ \forall i = 1..C; \\ \frac{\partial^2 C}{\partial a(i)\partial a(j)} &= \sum_{f=f_1}^{f_2} \frac{\partial^2 h(f)}{\partial a(i)\partial a(j)} log_2\left(\frac{h(f)}{\bar{\mathcal{P}}_{Y,n}(f)}\right) + \frac{\partial h(f)}{\partial a(i)}\frac{\partial h(f)}{\partial a(j)}\frac{1}{h(f)log(2)} + \\ & \quad \frac{\partial^2 h(f)}{\partial a(i)\partial a(j)}\frac{1}{log(2)}, \ \forall i,j = 1..C. \end{aligned} \tag{A.3}$$

Herein, the first and second order derivatives of the intermediate variable $h(f) \in \mathbb{R}$ can be calculated as follows:

$$\begin{aligned} \frac{\partial h(f)}{\partial a(i)} &= \frac{\bar{\mathcal{P}}_{X_i,n}(f)\beta - \alpha(f)}{\beta^2}; \\ \frac{\partial^2 h(f)}{\partial a(i)\partial a(j)} &= \frac{\bar{\mathcal{P}}_{X_i,n}(f) - \bar{\mathcal{P}}_{X_j,n}(f)}{\beta^2} - \frac{2\left(\bar{\mathcal{P}}_{X_i,n}(f)\beta - \alpha(f)\right)}{\beta^3}. \end{aligned} \tag{A.4}$$

Using these expressions, the non-convexity of the KLD early fusion cost function can be proven by means of a counterexample. Assume the following values:

$$\begin{aligned} C &= 2; \\ \bar{\boldsymbol{\mathcal{P}}}_{X_1,n} &= \begin{bmatrix} \frac{2}{6} & \frac{4}{6} \end{bmatrix}^\top; \\ \bar{\boldsymbol{\mathcal{P}}}_{X_2,n} &= \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \end{bmatrix}^\top; \\ \bar{\boldsymbol{\mathcal{P}}}_{Y,n} &= \begin{bmatrix} \frac{4}{5} & \frac{1}{5} \end{bmatrix}^\top; \\ \mathbf{a} &= \begin{bmatrix} cos\left(\frac{\pi}{6}\right) & sin\left(\frac{\pi}{6}\right) \end{bmatrix}^\top. \end{aligned} \tag{A.5}$$

Using these values, the Hessian of the cost function possesses eigenvalues $8.65 \cdot 10^{-2}$ and $-4.86 \cdot 10^{-2}$, resulting in an indefinite (i.e. not positive semidefinite) matrix. Referring to theorems A.1.1 and A.1.4, the early fusion KLD method is consequently non-convex.
□

A.3 Least squares-linear discriminant analysis discriminator-domain adaptation (D-DA)

This appendix studies the convexity of equation 5.5, which is repeated here for the sake of completeness:

$$\begin{aligned} \min_{\mathbf{d}} \quad & -\mathbf{d}^\top \mathbf{r}_{\underline{x}_{\mathcal{S},A} y} + \frac{\nu}{N_\mathcal{S} + N_\mathcal{T}} \sum_{l=\{\mathcal{S},\mathcal{T}\}} \sum_{n=1}^{N_l} z_l(n)\sigma(\beta_0 + \beta c_l(n)) + \\ & (1 - z_l(n))(1 - \sigma(\beta_0 + \beta c_l(n))) \\ s.t. \quad & \mathbf{d}^\top R_{\underline{x}_{\mathcal{S},A}\underline{x}_{\mathcal{S},A}} \mathbf{d} = 1 \\ & c_l(n) = \rho(X_l(n)\mathbf{d}, \mathbf{y}_l(n)), \quad n = 1..N_l,\ l = \{\mathcal{S},\mathcal{T}\} \\ & \bar{\Sigma} = \frac{1}{N_\mathcal{S}} \sum_{n=1}^{N_\mathcal{S}} (c_\mathcal{S}(n) - \bar{\mu}_\mathcal{S})^2 + \frac{1}{N_\mathcal{T}} \sum_{n=1}^{N_\mathcal{T}} (c_\mathcal{T}(n) - \bar{\mu}_\mathcal{T})^2 \\ & \bar{\mu}_\mathcal{S} = \frac{1}{N_\mathcal{S}} \sum_{n=1}^{N_\mathcal{S}} c_\mathcal{S}(n) \\ & \bar{\mu}_\mathcal{T} = \frac{1}{N_\mathcal{T}} \sum_{n=1}^{N_\mathcal{T}} c_\mathcal{T}(n) \\ & \beta = \bar{\Sigma}^{-1}(\bar{\mu}_\mathcal{T} - \bar{\mu}_\mathcal{S}) \\ & \beta_0 = -\frac{1}{2}\beta(\bar{\mu}_\mathcal{T} + \bar{\mu}_\mathcal{S}). \end{aligned} \tag{A.6}$$

Referring to theorems A.1.1-A.1.4, the non-convexity of equation A.6 can be proven by means of a counterexample: Finding a feasible point, wherein the Hessian is

not positive semidefinite. To this end, firstly, an expression of the Hessian of the cost function needs to be derived. To simplify calculations, the $-\mathbf{d}^\top \mathbf{r}_{\underline{x}_{\mathcal{S},A} y}$ term is further on neglected since its second order derivative is zero. Furthermore, the feature dimension is assumed to equal one and the decoder $\mathbf{d} \in \mathbb{R}^{C\times 1}$ is assumed to only leverage spatial information. The remaining cost term is symbolised by the notation *BCEm*. Its first order derivative can be written as follows:

$$
\begin{aligned}
\frac{\partial BCEm}{\partial d(i)} &= \frac{\nu}{N_{\mathcal{S}}+N_{\mathcal{T}}} \sum_{l=\{\mathcal{S},\mathcal{T}\}} \sum_{n=1}^{N_l} (-1+2z_l(n))(\sigma(\beta_0+\beta c_l(n))- \\
&\quad \sigma(\beta_0+\beta c_l(n))^2)\left(\frac{\partial \beta_0}{\partial d(i)} + \beta \frac{\partial c_l(n)}{\partial d(i)} + c_l(n)\frac{\partial \beta}{\partial d(i)}\right); \\
\frac{\partial \beta}{\partial d(i)} &= \frac{1}{\bar{\Sigma}}\left(\frac{\partial \bar{\mu}_{\mathcal{T}}}{\partial d(i)} - \frac{\partial \bar{\mu}_{\mathcal{S}}}{\partial d(i)}\right) - \frac{1}{\bar{\Sigma}^2}\left(\bar{\mu}_{\mathcal{T}} - \bar{\mu}_{\mathcal{S}}\right)\frac{\partial \bar{\Sigma}}{\partial d(i)}; \\
\frac{\partial \beta_0}{\partial d(i)} &= \frac{-1}{2}\left[\frac{\partial \beta}{\partial d(i)}\left(\bar{\mu}_{\mathcal{S}}+\bar{\mu}_{\mathcal{T}}\right) + \beta\left(\frac{\bar{\mu}_{\mathcal{S}}}{\partial d(i)} + \frac{\partial \bar{\mu}_{\mathcal{T}}}{\partial d(i)}\right)\right]; \\
\frac{\partial \bar{\Sigma}}{\partial d(i)} &= \sum_{l=\{\mathcal{S},\mathcal{T}\}} \frac{1}{N_l} \sum_{n=1}^{N_l} 2(c_l(n)-\bar{\mu}_l)\left(\frac{\partial c_l(n)}{\partial d(i)} - \frac{\partial \bar{\mu}_l}{\partial d(i)}\right); \\
\frac{\partial \mu_l}{\partial d(i)} &= \frac{1}{N_l}\sum_{n=1}^{N_l} \frac{\partial c_l(n)}{\partial d(i)}; \\
\frac{\partial c_l(n)}{\partial d(i)} &= \frac{\sum_{t=0}^{T-1} x_{l,n}(t,i) y_{l,n}(t)}{\left(\sum_{t=0}^{T-1} y_{l,n}(t)^2\right)^{1/2}} \frac{1}{\left(\sum_{t=0}^{T-1}\left(\sum_{c=1}^{C} d(c) x_{l,n}(t,c)\right)^2\right)^{1/2}} + \\
&\quad \frac{-1}{2\left(\sum_{t=0}^{T-1} y_{l,n}(t)^2\right)^{1/2}} \frac{\left(\sum_{t=0}^{T-1} 2\left(\sum_{c=1}^{C} d(c) x_{l,n}(t,c)\right) x_{l,n}(t,i)\right)}{\left(\sum_{t=0}^{T-1}\left(\sum_{c=1}^{C} d(c) x_{l,n}(t,c)\right)^2\right)^{3/2}} \cdot \\
&\quad \left(\sum_{t=0}^{T-1} y_{l,n}(t) \sum_{c=1}^{C} d(c) x_{l,n}(t,c)\right); \\
&\quad \forall l \in \{\mathcal{S},\mathcal{T}\},\ \forall i = 1..C.
\end{aligned}
\tag{A.7}
$$

Herein, $x_{l,n}(t,c) \in \mathbb{R}$ is shorthand notation for the tth ($t = 0..T-1$) EEG sample in the cth channel ($c = 1..C$) of the nth ($n = 1..N_l,\ l = \{\mathcal{S},\mathcal{T}\}$) window. Similarly, $y_{l,n}(t) \in \mathbb{R}$ is shorthand notation for the tth ($t = 0..T-1$) audio sample in the nth ($n = 1..N_l,\ l = \{\mathcal{S},\mathcal{T}\}$) window. In other words, these values form the elements of the EEG data $X_l(n) \in \mathbb{R}^{T\times C}$, in window n, and the accompanying audio data

$\mathbf{y}_l(n) \in \mathbb{R}^{T\times 1}$:

$$X_l(n) = \begin{bmatrix} x_{l,n}(0,1) & x_{l,n}(0,2) & \dots & x_{l,n}(0,C) \\ x_{l,n}(1,1) & x_{l,n}(1,2) & \dots & x_{l,n}(1,C) \\ \vdots & \vdots & \dots & \vdots \\ x_{l,n}(T-1,1) & x_{l,n}(T-1,2) & \dots & x_{l,n}(T-1,C) \end{bmatrix}; \tag{A.8}$$
$$\mathbf{y}_l(n) = \begin{bmatrix} y_{l,n}(0) & y_{l,n}(2) & \dots & y_{l,n}(T-1) \end{bmatrix}^\top.$$

The second order derivative takes the following form:

$$\begin{aligned}
\frac{\partial^2 BCEm}{\partial d(i)\partial d(j)} = \frac{\nu}{N_{\mathcal{S}}+N_{\mathcal{T}}} \sum_{l=\{\mathcal{S},\mathcal{T}\}} \sum_{n=1}^{N_l} (-1+2z_l(n)) \Bigg[\Big(\sigma(\beta_0+\beta c_l(n)) - \\
\sigma(\beta_0+\beta c_l(n))^2\Big) \frac{\partial^2\beta_0}{\partial d(i)\partial d(j)} + \\
\left(\frac{\partial\sigma(\beta_0+\beta c_l(n))}{\partial d(j)} - 2\sigma(\beta_0+\beta c_l(n))\frac{\partial\sigma(\beta_0+\beta c_l(n))}{\partial d(j)}\right)\frac{\partial\beta_0}{\partial d(i)} + \\
\Big(\sigma(\beta_0+\beta c_l(n)) - \sigma(\beta_0+\beta c_l(n))^2\Big)\cdot \\
\left(\beta\frac{\partial^2 c_l(n)}{\partial d(i)\partial d(j)} + \frac{\partial\beta}{\partial d(j)}\frac{\partial c_l(n)}{\partial d(i)}\right) + \\
\left(\frac{\partial\sigma(\beta_0+\beta c_l(n))}{\partial d(j)} - 2\sigma(\beta_0+\beta c_l(n))\frac{\partial\sigma(\beta_0+\beta c_l(n))}{\partial d(j)}\right)\beta\frac{\partial c_l(n)}{\partial d(i)} + \\
\Big(\sigma(\beta_0+\beta c_l(n)) - \sigma(\beta_0+\beta c_l(n))^2\Big)\cdot \\
\left(c_l(n)\frac{\partial^2\beta}{\partial d(i)\partial d(j)} + \frac{\partial c_l(n)}{\partial d(j)}\frac{\partial\beta}{\partial d(i)}\right) + \\
\left(\frac{\partial\sigma(\beta_0+\beta c_l(n))}{\partial d(j)} - 2\sigma(\beta_0+\beta c_l(n))\frac{\partial\sigma(\beta_0+\beta c_l(n))}{\partial d(j)}\right)\cdot \\
c_l(n)\frac{\partial\beta}{\partial d(i)}\Bigg];
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\sigma(\beta_0+\beta c_l(n))}{\partial d(j)} &= \sigma(\beta_0+\beta c_l(n))(1-\sigma(\beta_0+\beta c_l(n)))\cdot \\
&\left(\frac{\partial\beta_0}{\partial d(j)} + \beta\frac{\partial c_l(n)}{\partial d(j)} + c_l(n)\frac{\partial\beta}{\partial d(j)}\right); \\
\frac{\partial^2\beta}{\partial d(i)\partial d(j)} &= \frac{1}{\bar{\Sigma}^2}\left[\left(\frac{\partial^2\bar{\mu}_{\mathcal{T}}}{\partial d(i)\partial d(j)} - \frac{\partial^2\bar{\mu}_{\mathcal{S}}}{\partial d(i)\partial d(j)}\right)\bar{\Sigma} - \frac{\partial\bar{\Sigma}}{\partial d(j)}\left(\frac{\partial\bar{\mu}_{\mathcal{T}}}{\partial d(i)} - \frac{\partial\bar{\mu}_{\mathcal{S}}}{\partial d(i)}\right)\right] \\
&- \frac{1}{\bar{\Sigma}^4}\left[\frac{\partial^2\bar{\Sigma}}{\partial d(i)\partial d(j)}(\bar{\mu}_{\mathcal{T}}-\bar{\mu}_{\mathcal{S}})\bar{\Sigma}^2 + \frac{\partial\bar{\Sigma}}{\partial d(i)}\left(\frac{\partial\bar{\mu}_{\mathcal{T}}}{\partial d(j)} - \frac{\partial\bar{\mu}_{\mathcal{S}}}{\partial d(j)}\right)\bar{\Sigma}^2 -\right.
\end{aligned}$$

$$2\bar{\Sigma}\frac{\partial\bar{\Sigma}}{\partial d(j)}(\bar{\mu}_{\mathcal{T}}-\bar{\mu}_{\mathcal{S}})\frac{\partial\bar{\Sigma}}{\partial d(i)}\Bigg];$$

$$\frac{\partial^2\beta_0}{\partial d(i)\partial d(j)}=\frac{-1}{2}\Bigg[\frac{\partial^2\beta}{\partial d(i)\partial d(j)}(\bar{\mu}_{\mathcal{S}}+\bar{\mu}_{\mathcal{T}})+\frac{\partial\beta}{\partial d(i)}\left(\frac{\partial\bar{\mu}_{\mathcal{S}}}{\partial d(j)}+\frac{\partial\bar{\mu}_{\mathcal{T}}}{\partial d(j)}\right)+$$
$$\frac{\partial\beta}{\partial d(j)}\left(\frac{\partial\bar{\mu}_{\mathcal{S}}}{\partial d(i)}+\frac{\partial\bar{\mu}_{\mathcal{T}}}{\partial d(i)}\right)+\beta\left(\frac{\partial^2\bar{\mu}_{\mathcal{S}}}{\partial d(i)\partial d(j)}+\frac{\partial^2\bar{\mu}_{\mathcal{T}}}{\partial d(i)\partial d(j)}\right)\Bigg];$$

$$\frac{\partial^2\bar{\Sigma}}{\partial d(i)\partial d(j)}=\sum_{l=\mathcal{S},\mathcal{T}}\frac{1}{N_l}\sum_{n=1}^{N_l}2c_l(n)\frac{\partial^2c_l(n)}{\partial d(i)\partial d(j)}+2\frac{\partial c_l(n)}{\partial d(j)}\frac{\partial c_l(n)}{\partial d(i)}-$$
$$2\frac{\partial c_l(n)}{\partial d(j)}\frac{\partial\bar{\mu}_l}{\partial d(i)}-2c_l(n)\frac{\partial^2\bar{\mu}_l}{\partial d(i)\partial d(j)}-2\frac{\partial\bar{\mu}_l}{\partial d(j)}\frac{\partial c_l(n)}{\partial d(i)}-$$
$$2\bar{\mu}_l\frac{\partial^2c_l(n)}{\partial d(i)\partial d(j)}+2\bar{\mu}_l\frac{\partial^2\bar{\mu}_l}{\partial d(i)\partial d(j)}+2\frac{\partial\bar{\mu}_l}{\partial d(j)}\frac{\partial\bar{\mu}_l}{\partial d(i)};$$

$$\frac{\partial^2\bar{\mu}_l}{\partial d(i)\partial d(j)}=\frac{1}{N_l}\sum_{n=1}^{N_l}\frac{\partial^2c_l(n)}{\partial d(i)\partial d(j)};$$

$$\frac{\partial^2c_l(n)}{\partial d(i)\partial d(j)}=\frac{\sum_{t=0}^{T-1}x_{l,n}(t,i)y_{l,n}(t)}{\left(\sum_{t=0}^{T-1}y_{l,n}(t)^2\right)^{1/2}}\left(\frac{-1}{2}\right)\frac{1}{\left(\sum_{t=0}^{T-1}\left(\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)^2\right)^{3/2}}\cdot$$
$$\left(\sum_{t=0}^{T-1}2x_{l,n}(t,j)\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)+$$
$$\frac{-1}{2\left(\sum_{t=0}^{T-1}y_{l,n}(t)^2\right)^{1/2}\left(\sum_{t=0}^{T-1}\left(\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)^2\right)^{3}}\cdot$$
$$\Bigg\{\Bigg[\left(\sum_{t=0}^{T-1}2x_{l,n}(t,j)x_{l,n}(t,i)\right)\left(\sum_{t=0}^{T-1}\left(\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)y_{l,n}(t)\right)+$$
$$\left(\sum_{t=0}^{T-1}2x_{l,n}(t,i)\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)\left(\sum_{t=0}^{T-1}x_{l,n}(t,j)y_{l,n}(t)\right)\Bigg]\cdot$$
$$\left(\sum_{t=0}^{T-1}\left(\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)^2\right)^{3/2}$$
$$-\frac{3}{2}\left(\sum_{t=0}^{T-1}\left(\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)^2\right)^{1/2}\left(\sum_{t=0}^{T-1}2x_{l,n}(t,j)\left(\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)\right)\cdot$$
$$\left(\sum_{t=0}^{T-1}2x_{l,n}(t,i)\left(\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)\right)\left(\sum_{t=0}^{T-1}y_{l,n}(t)\sum_{c=1}^{C}d(c)x_{l,n}(t,c)\right)\Bigg\};$$
$$\forall l\in\{\mathcal{S},\mathcal{T}\},\ \forall i=1..C.$$

Let $\nu = 1$, $T = 1$, $N_{\mathcal{S}} = N_{\mathcal{T}} = 2$ and $C = 2$, and assume the following values:

$$\begin{aligned}
X_{\mathcal{S}}(1) &= \begin{bmatrix} 1.0 & 1.1 \\ 2.0 & 1.0 \end{bmatrix}; \\
X_{\mathcal{S}}(2) &= \begin{bmatrix} 2.0 & 1.9 \\ 2.2 & 0.7 \end{bmatrix}; \\
X_{\mathcal{T}}(1) &= \begin{bmatrix} 1.5 & 1.7 \\ 2.3 & 0.9 \end{bmatrix}; \\
X_{\mathcal{T}}(2) &= \begin{bmatrix} 3.0 & 2.8 \\ 1.7 & 2.7 \end{bmatrix}; \\
\mathbf{y}_{\mathcal{S}}(1) &= \begin{bmatrix} 1.0 & 0.2 \end{bmatrix}^{\top}; \\
\mathbf{y}_{\mathcal{S}}(2) &= \begin{bmatrix} 2.0 & 1.3 \end{bmatrix}^{\top}; \\
\mathbf{y}_{\mathcal{T}}(1) &= \begin{bmatrix} 0.5 & 0.8 \end{bmatrix}^{\top}; \\
\mathbf{y}_{\mathcal{T}}(2) &= \begin{bmatrix} 3.0 & 3.2 \end{bmatrix}^{\top}; \\
\mathbf{d} &= \begin{bmatrix} 1.2 \cdot 10^{-1} & 6.2 \cdot 10^{-2} \end{bmatrix}^{\top}.
\end{aligned} \tag{A.9}$$

Using these values, the second order derivatives can be computed. These derivatives result in a Hessian with eigenvalues -1.8 and 2.8. Since both eigenvalues differ in sign, the Hessian is indefinite and the optimisation problem is subsequently non-convex. □

Appendix B

Assessment of the linear discriminant analysis assumptions

The linear discriminant analysis (LDA) classifier assumes normality of the feature space and equal covariance matrices for the class conditional distributions. Although the LDA method can still prove its worth if these assumptions are not satisfied [52], it is insightful to check them. To this end, section B.1 details Mardia's test for normality assessment and section B.2 details Box's M test for equal covariance assessment. Thereafter, both tests are applied to the Vanthornhout dataset features in section B.3.

B.1 Normality assessment using Mardia's test

To confirm multivariate normality, Mardia's test is used [90]. This test assumes the null-hypothesis that the $N \in \mathbb{N}_0$ observations of F-dimensional ($F \in \mathbb{N}_0$) feature vectors $\mathbf{f}_n \in \mathbb{R}^{F\times 1}$, $n = 1..N$ are normally distributed. Define the sample mean $\bar{\boldsymbol{\mu}} \in \mathbb{R}^{F\times 1}$ and sample covariance matrix $\bar{\Sigma} \in \mathbb{R}^{F\times F}$ as follows:

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{n=1}^{N} \mathbf{f}_n; \\ \bar{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N} (\mathbf{f}_n - \bar{\boldsymbol{\mu}})(\mathbf{f}_n - \bar{\boldsymbol{\mu}})^\top. \end{aligned} \tag{B.1}$$

Then, Mardia's test compares the sample kurtosis $\in \mathbb{R}$ [121]:

$$Kurtosis = \frac{1}{N}\sum_{n=1}^{N} \left[(\mathbf{f}_n - \bar{\boldsymbol{\mu}})^\top \bar{\Sigma}^{-1} (\mathbf{f}_n - \bar{\boldsymbol{\mu}})\right]^2, \tag{B.2}$$

to the expected kurtosis of a normal distribution, namely $F(F+2)$. Similarly, the sample skewness $\in \mathbb{R}$ [121]:

$$Skewness = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\left[(\mathbf{f}_n - \bar{\boldsymbol{\mu}})^\top \bar{\Sigma}^{-1}(\mathbf{f}_{n'} - \bar{\boldsymbol{\mu}})\right]^3, \tag{B.3}$$

is compared to the expected skewness of a normal distribution, namely 0.

This translates to the following statistics under the null-hypothesis that the data is normally distributed: Statistic $A \in \mathbb{R}$ tests the skewness and is (asymptomatically) χ^2 distributed with $\frac{1}{6}F(F+1)(F+2)$ degrees of freedom. Statistic $B \in \mathbb{R}$ tests the kurtosis and is (asymptomatically) standard normal distributed [90]:

$$\begin{aligned}
A &= \frac{1}{6N}\sum_{n=1}^{N}\sum_{n'=1}^{N}\left[(\mathbf{f}_n - \bar{\boldsymbol{\mu}})^\top \bar{\Sigma}^{-1}(\mathbf{f}_{n'} - \bar{\boldsymbol{\mu}})\right]^3;\\
B &= \left(\frac{N}{8F(F+2)}\right)^{1/2}\left[\frac{1}{N}\sum_{n=1}^{N}\left((\mathbf{f}_n - \bar{\boldsymbol{\mu}})^\top \bar{\Sigma}^{-1}(\mathbf{f}_n - \bar{\boldsymbol{\mu}})\right)^2 - F(F+2)\right];\\
\bar{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{f}_n;\\
\bar{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{f}_n - \bar{\boldsymbol{\mu}})(\mathbf{f}_n - \bar{\boldsymbol{\mu}})^\top.
\end{aligned} \tag{B.4}$$

B.2 Equal covariance assessment using Box's M test

To study the assumption of equal covariance matrices of both classes, Box's M test is used [91]. This test, solely applicable to normally distributed data, assumes the null-hypothesis that the covariance matrices of F-dimensional features $\bar{\Sigma}_1, \bar{\Sigma}_2 \in \mathbb{R}^{F\times F}$, corresponding to classes K_1 and K_2, are equal. Assume respectively $N_1 \in \mathbb{N}_0$ and $N_2 \in \mathbb{N}_0$ observations of features $\mathbf{f}_1 \in \mathbb{R}^{F\times 1}$ and $\mathbf{f}_2 \in \mathbb{R}^{F\times 1}$ corresponding to classes K_1 and K_2. The following F-statistics $F^+ \in \mathbb{R}$ and $F^- \in \mathbb{R}$ can subsequently be computed [91]:

$$\begin{aligned}
F^+ &= \frac{M\left(1 - c_1 - \frac{df_1}{df_2}\right)}{df_1};\\
F^- &= \frac{df_2 M}{df_1\left(\frac{df_2}{1-c_1+\frac{2}{df_2}} - M\right)};\\
df_1 &= \frac{F(F+1)}{2};\\
df_2 &= \frac{df_1 + 2}{|c_2 - c_1^2|};\\
c_1 &= \frac{2F^2+3F-1}{6(F+1)}\left[\sum_{i=1}^{2}\left(\frac{1}{N_i - 1}\right) - \frac{1}{N_1 + N_2 - 2}\right];
\end{aligned} \tag{B.5}$$

$$
\begin{aligned}
c_2 &= \frac{(F-1)(F+2)}{6}\left[\sum_{=1}^{2}\left(\frac{1}{(N_i-1)^2}\right)-\frac{1}{(N_1+N_2-2)^2}\right];\\
M &= (N_1+N_2-2)log(det(S))-\sum_{i=1}^{2}(N_i-1)log(det(\bar{\Sigma}_i));\\
S &= \frac{1}{N_1+N_2-2}\sum_{i=1}^{2}(N_i-1)\bar{\Sigma}_i;\\
\bar{\boldsymbol{\mu}}_1 &= \frac{1}{N}\sum_{n=1}^{N_1}\mathbf{f}_{1,n};\\
\bar{\boldsymbol{\mu}}_2 &= \frac{1}{N}\sum_{n=1}^{N_2}\mathbf{f}_{2,n};\\
\bar{\Sigma}_1 &= \frac{1}{N}\sum_{n=1}^{N_1}(\mathbf{f}_{1,n}-\bar{\boldsymbol{\mu}}_1)(\mathbf{f}_{1,n}-\bar{\boldsymbol{\mu}}_1)^\top;\\
\bar{\Sigma}_2 &= \frac{1}{N}\sum_{n=1}^{N_2}(\mathbf{f}_{2,n}-\bar{\boldsymbol{\mu}}_2)(\mathbf{f}_{2,n}-\bar{\boldsymbol{\mu}}_2)^\top.
\end{aligned}
\tag{B.6}
$$

If $c_2 > c_1^2$, F^+ is F-distributed with parameters df_1 and df_2, else this F-distribution holds for F^-.

B.3 Application of Mardia's and Box's M tests

Both tests are applied to all combinations of LS, LASSO, KLD and entropy, CCA with $1-5$, filters and delta, theta, alpha, beta and all frequency band CSP with one filter maximising the attention state and one filter maximising the inattention state. Indeed, these feature combinations are attained in the classification experiments of chapter 8 and should therefore be tested for LDA assumption violations. Features are extracted using a 10 fold cross validation, wherein the validation segments are partitioned into windows of 1, 5 and 10 s. Mardia's test and Box's M test are performed for each subject, window and attention-inattention state separately, and p-values are corrected for multiple comparisons using a Benjamini-Hochberg correction. Significance level equals 5%.

Tables B.1a-B.1c display the minimum p-values for each group (LS-LASSO-KLD-entropy combination, CCA and CSP) over the different subjects, windows and attention states. Regarding Mardia's skewness test also the maximum skewness, the mean skewness and the standard deviation of the skewness over the group results are shown. Similar measures are extracted from the ratio between the calculated kurtosis and the expected kurtosis for a normal distribution $F(F+2)$: the maximum and minimum ratio in absolute value, the mean ratio and its standard deviation. An identical approach is taken regarding the maximum absolute value of the elementwise ratio of the attention over inattention covariance matrices.

Mardia's test results in a rejection of the normality assumption, both regarding the skewness test (min(p)=$0.0 \cdot 10^0$) and the kurtosis test (min(p)=$0.0 \cdot 10^0$) in a worst case scenario. Nevertheless, the mean skewness of the groups lies around $O(10^{-2}) - O(10^{-1})$ and the maximum value lies around 1, which indicate moderate skewness. The ratio of the kurtosis shows more extreme values since the minimum ratio for the LS-LASSO-KLD-entropy combinations and CCA equals 0 and the maximum kurtosis ratio for the LS-LASSO-KLD-entropy and CSP combinations is $O(10^0)$-$O(10^1)$. The mean kurtosis ratios are more moderate since these ratios are $O(10^{-1})$-$O(10^0)$. Furthermore, the best case scenario attains a p-value $1.0 \cdot 10^0$, both regarding the skewness and the kurtosis test, such that not for all cases evidence is found against the normality assumption.

Box's M test results in a rejection of the equal covariance assumption (min(p)=$0.0 \cdot 10^0$) in a worst case scenario. This result needs to be interpreted with caution since Box's M test is also sensitive to deviations from normality, such that this test might not truly check the equal covariance assumption [122]. We can nonetheless justify our approach since deviations from normality also violate the LDA assumptions, and the ratios of covariance matrices can furthermore be studied in conjunction with performing statistical tests. To this end, the maximum elementwise ratio is $O(10^1)$ for KLD-LASSO-LS-entropy combinations and $O(10^2)$ for CSP. In fact, the mean covariance ratio of CSP is $O(10^1)$. The mean covariance ratio regarding the KLD-LASSO-LS-entropy combinations on the other hand is closer to the expected value of 1 since it is $O(10^0)$. In addition, the best case scenario achieves a p-value $1.0 \cdot 10^0$, such that not in all cases evidence is found against the equal covariance assumption.

In conclusion, there is found evidence in the data for departure from normality and for inequality of covariance matrices, such that the LDA assumptions are not satisfied in the worst case scenario. The mean cases nevertheless seem to satisfy the LDA assumptions to a higher degree since the order of magnitude of skewness, kurtosis and covariance matrix ratio corresponds to the expected value. In addition, in the best case scenario, no evidence is found against these assumptions.

	Minimum P-value	Maximum absolute skewness	Mean skewness	Standard deviation skewness
LS-LASSO-KLD-entropy combinations	$0.0 \cdot 10^{0}$	$1.0 \cdot 10^{0}$	$1.3 \cdot 10^{-1}$	$2.6 \cdot 10^{-1}$
CCA	$1.2 \cdot 10^{-4}$	$1.0 \cdot 10^{0}$	$4.2 \cdot 10^{-1}$	$3.3 \cdot 10^{-1}$
CSP	$0.0 \cdot 10^{0}$	$8.6 \cdot 10^{-1}$	$7.5 \cdot 10^{-2}$	$1.9 \cdot 10^{-1}$

(A) Skewness test.

	Minimum P-value	Maximum absolute kurtosis ratio	Minimum absolute kurtosis ratio	Mean kurtosis ratio	Standard deviation kurtosis ratio
LS-LASSO-KLD-entropy combinations	$0.0 \cdot 10^{0}$	$7.3 \cdot 10^{0}$	$0.0 \cdot 10^{0}$	$1.4 \cdot 10^{0}$	$1.4 \cdot 10^{0}$
CCA	$0.0 \cdot 10^{0}$	$9.5 \cdot 10^{-1}$	$0.0 \cdot 10^{0}$	$1.2 \cdot 10^{-1}$	$1.5 \cdot 10^{-1}$
CSP	$0.0 \cdot 10^{0}$	$1.5 \cdot 10^{1}$	$3.2 \cdot 10^{-2}$	$2.3 \cdot 10^{0}$	$2.9 \cdot 10^{0}$

(B) Kurtosis test.

	Minimum P-value	Maximum absolute covariance ratio	Minimum absolute covariance ratio	Mean absolute covariance ratio	Standard deviation covariance ratio
LS-LASSO-KLD-entropy combinations	$0.0 \cdot 10^{0}$	$8.4 \cdot 10^{1}$	$1.6 \cdot 10^{-1}$	$2.2 \cdot 10^{0}$	$5.3 \cdot 10^{0}$
CCA	$2.4 \cdot 10^{-4}$	$2.8 \cdot 10^{0}$	$6.4 \cdot 10^{-1}$	$1.1 \cdot 10^{0}$	$2.6 \cdot 10^{-1}$
CSP	$0.0 \cdot 10^{0}$	$2.5 \cdot 10^{2}$	$9.4 \cdot 10^{-1}$	$7.3 \cdot 10^{1}$	$2.5 \cdot 10^{1}$

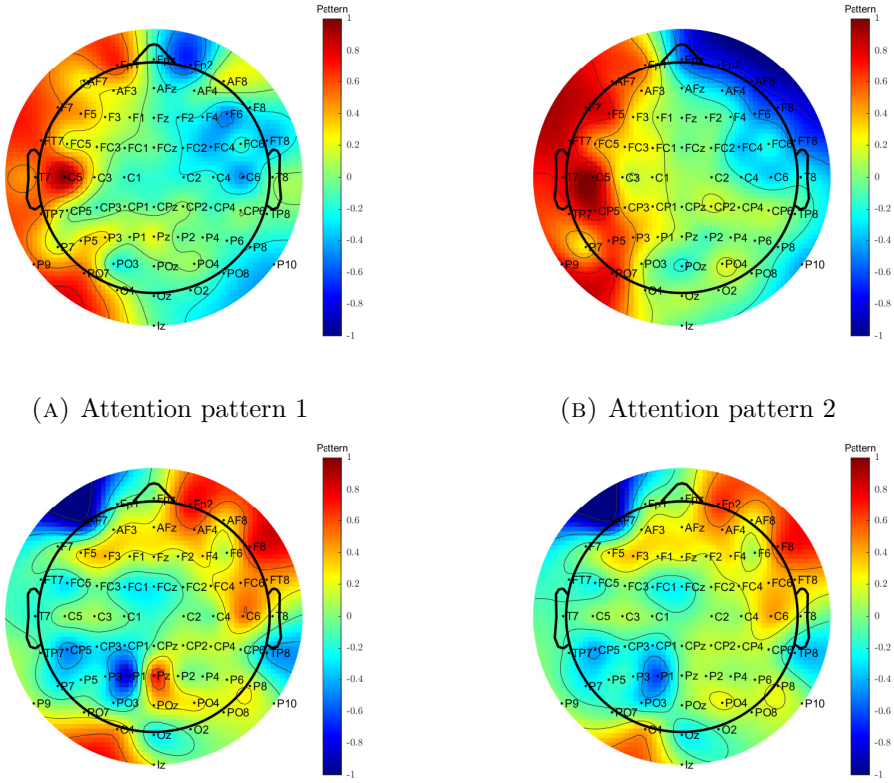
(C) Equal covariance test.

TABLE B.1: LDA assumption tests: normality and equal covariance using Mardia's and Box'M tests. Evidence is found for the rejection of the normality and equal covariance null-hypotheses. In addition, the maximum deviations are about an order of magnitude larger than the expected values for two normal distributions with equal covariance matrices. The mean cases nevertheless seem to have the desired order of magnitude.

Appendix C

Common spatial pattern topographic plots

Figures C.1-C.4 display the delta, theta, alpha and beta band CSP patterns using an SI-CV approach on the Vanthornhout dataset. Likewise, figures C.5-C.8 show the CSP patterns on the Brouckmans-Dewit-Vanhaelen dataset. For further explanation, regarding the neurological interpretation and experiment specifics, the reader is referred to chapters 3, 7 and 8.



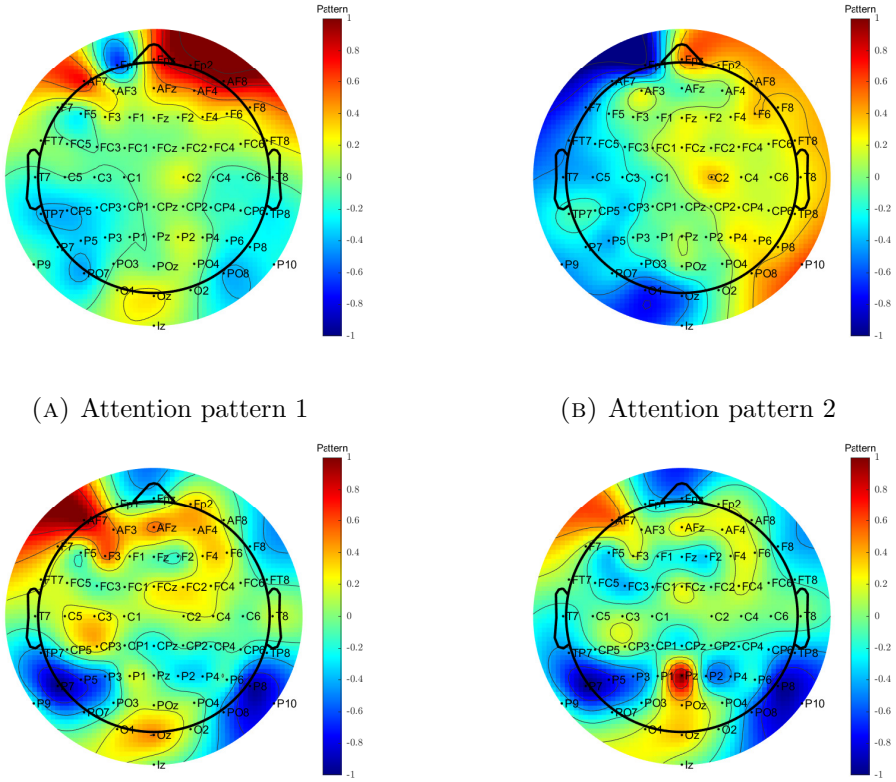
(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.1: SI-CV pattern topographic plot on the Vanthornhout dataset of delta band CSP. Structuring around the temporal lobes seems apparent.



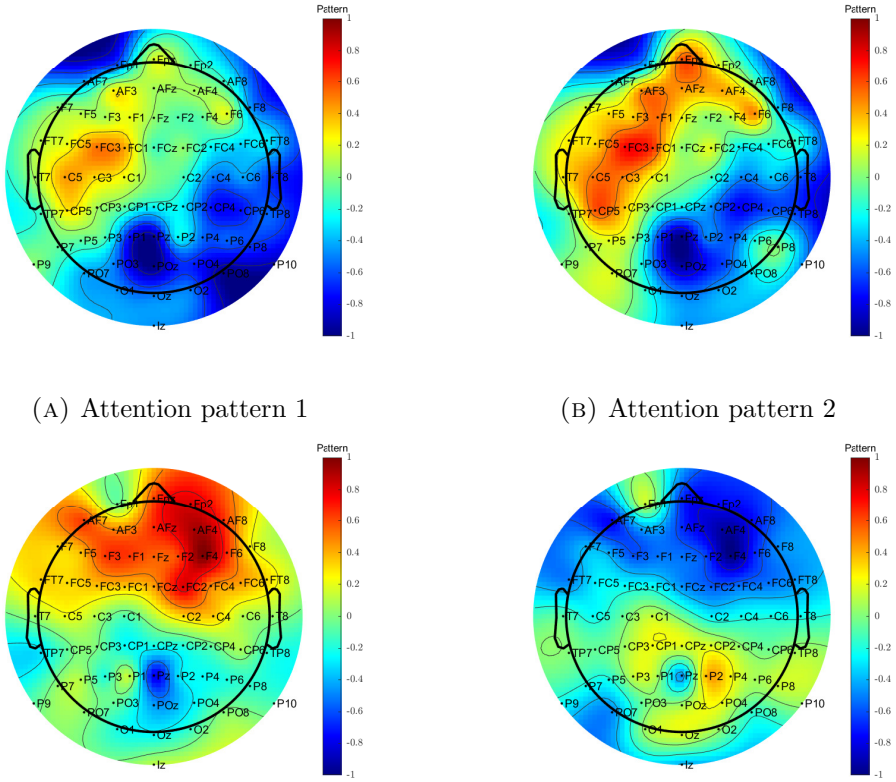
(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.2: SI-CV pattern topographic plot on the Vanthornhout dataset of theta band CSP. No clear, joint structuring seems apparent across the topographic plots.



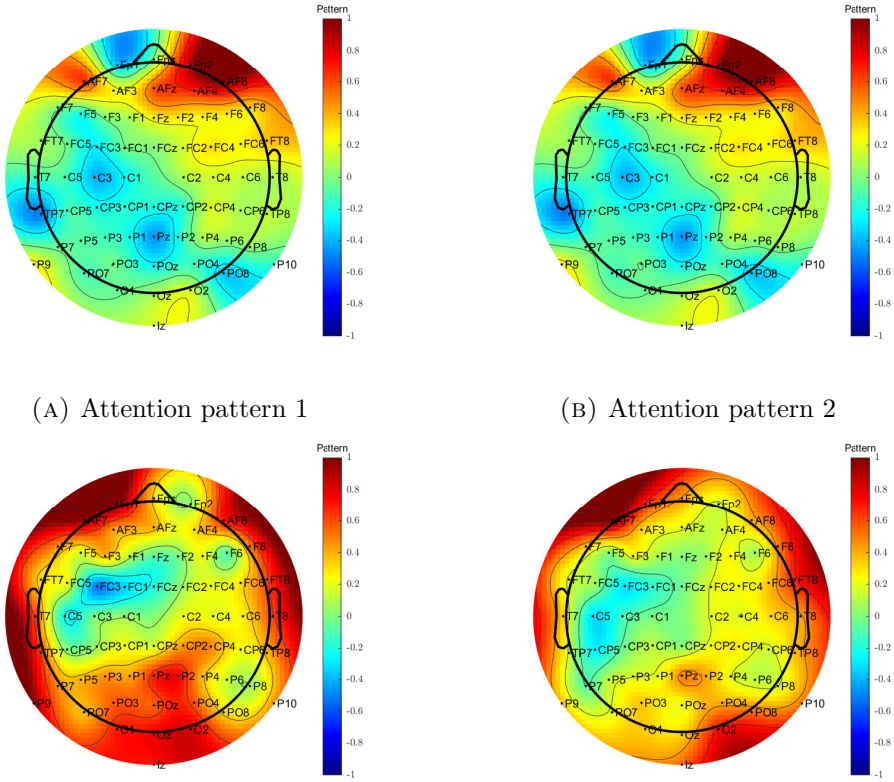
(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.3: SI-CV pattern topographic plot on the Vanthornhout dataset of alpha band CSP. No clear, joint structuring seems apparent across the topographic plots.



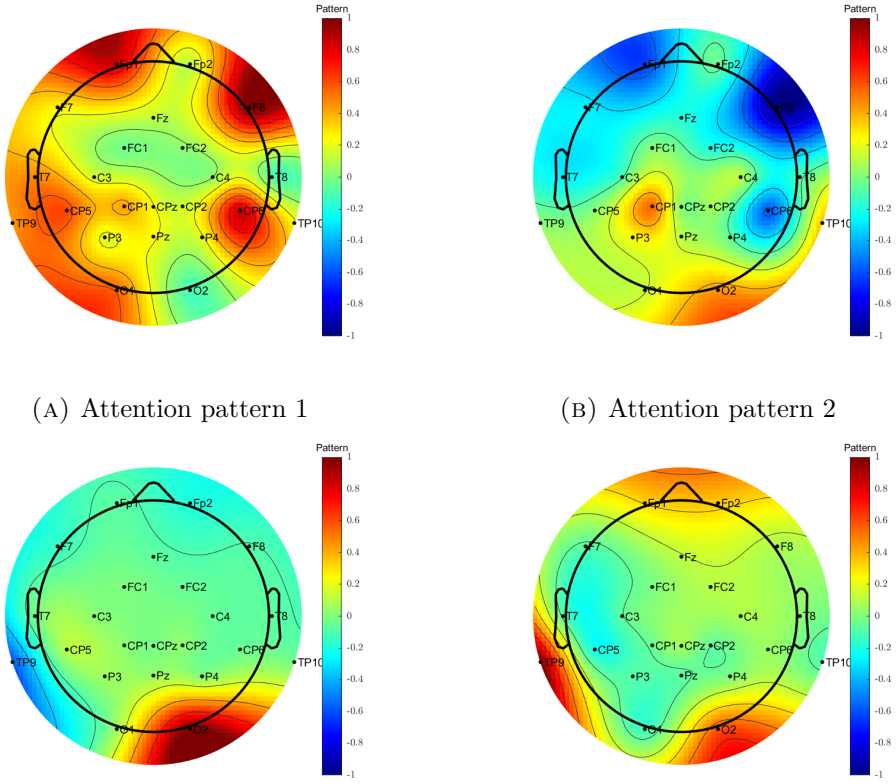
(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.4: SI-CV pattern topographic plot on the Vanthornhout dataset of beta band CSP. No clear, joint structuring seems apparent across the topographic plots.



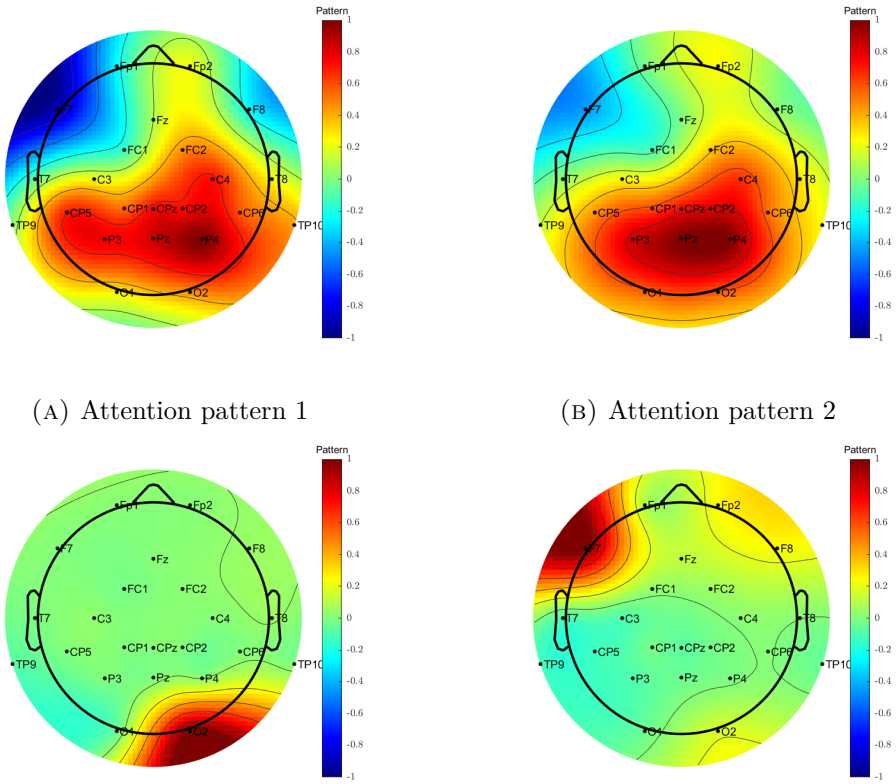
(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.5: SI-CV pattern topographic plot on the Brouckmans-Dewit-Vanhaelen dataset of delta band CSP.



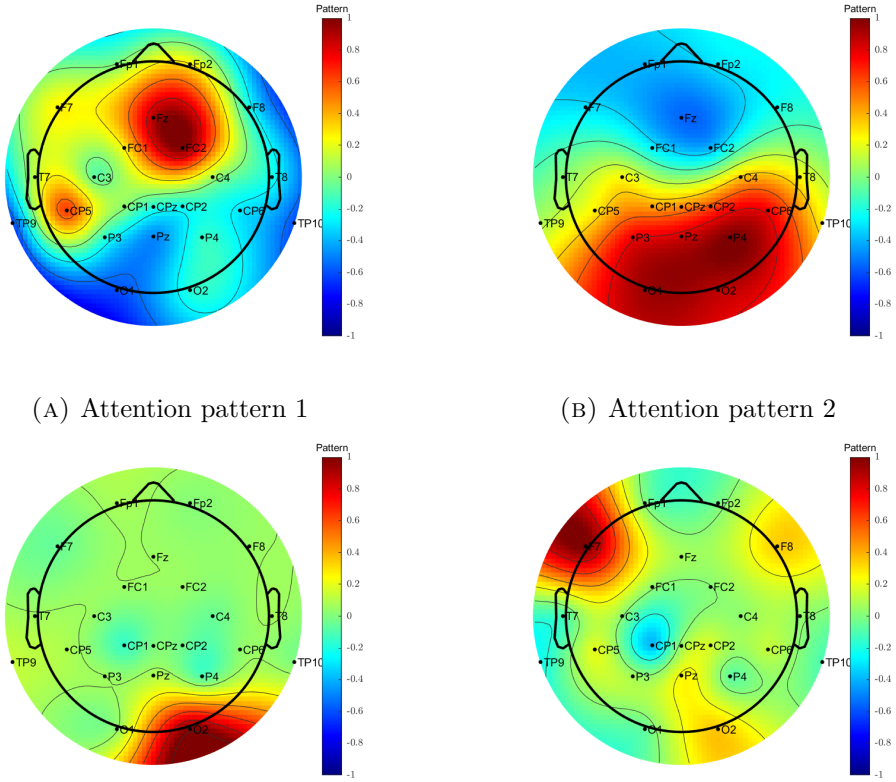
(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.6: SI-CV pattern topographic plot on the Brouckmans-Dewit-Vanhaelen dataset of theta band CSP.



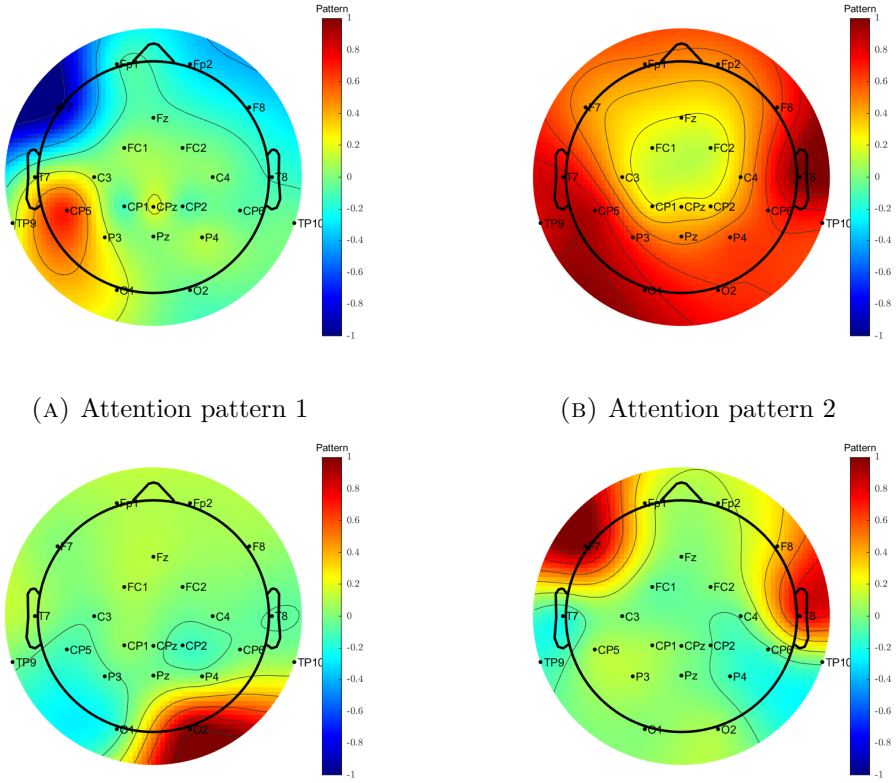
(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.7: SI-CV pattern topographic plot on the Brouckmans-Dewit-Vanhaelen dataset of alpha band CSP.



(A) Attention pattern 1

(B) Attention pattern 2

(C) Movie pattern 1

(D) Movie pattern 2

FIGURE C.8: SI-CV pattern topographic plot on the Brouckmans-Dewit-Vanhaelen dataset of beta band CSP.

Appendix D

Feature extractor-classifier performance

Tables D.1-D.3 display the mean classification accuracies and standard deviations of the CCA, CSP, LS, LASSO, KLD, entropy feature extractors and combinations thereof on the Vanthornhout dataset. Correspondingly, tables D.4 and D.5 show the mean classification accuracies and accompanying standard deviations of respectively CSP and the KLD-entropy combination on the Brouckmans-Dewit-Vanhaelen dataset. These results need to be seen in conjunction with the results of chapter 8. The highest mean accuracy per window length is marked in red.

Window [s]	1	5	10	20	30	60
LS	$0.52 \pm 2.4 \cdot 10^{-2}$	$0.56 \pm 4.2 \cdot 10^{-2}$	$0.56 \pm 6.5 \cdot 10^{-2}$	$0.61 \pm 9.0 \cdot 10^{-2}$	$0.63 \pm 8.1 \cdot 10^{-2}$	$0.66 \pm 1.4 \cdot 10^{-1}$
CCA1	$0.52 \pm 1.2 \cdot 10^{-2}$	$0.56 \pm 4.3 \cdot 10^{-2}$	$0.55 \pm 4.2 \cdot 10^{-2}$	$0.59 \pm 8.0 \cdot 10^{-2}$	$0.59 \pm 8.3 \cdot 10^{-2}$	$0.65 \pm 1.1 \cdot 10^{-1}$
CCA2	$0.52 \pm 1.2 \cdot 10^{-2}$	$0.57 \pm 3.9 \cdot 10^{-2}$	$0.56 \pm 4.5 \cdot 10^{-2}$	$0.60 \pm 4.5 \cdot 10^{-2}$	$0.64 \pm 4.6 \cdot 10^{-2}$	$0.69 \pm 9.5 \cdot 10^{-2}$
CCA3	$0.51 \pm 8.7 \cdot 10^{-3}$	$0.55 \pm 4.3 \cdot 10^{-2}$	$0.55 \pm 6.1 \cdot 10^{-2}$	$0.58 \pm 7.9 \cdot 10^{-2}$	$0.59 \pm 8.8 \cdot 10^{-2}$	$0.64 \pm 1.1 \cdot 10^{-1}$
CCA4	$0.51 \pm 1.4 \cdot 10^{-2}$	$0.54 \pm 4.7 \cdot 10^{-2}$	$0.53 \pm 7.2 \cdot 10^{-2}$	$0.60 \pm 6.8 \cdot 10^{-2}$	$0.61 \pm 1.1 \cdot 10^{-1}$	$0.65 \pm 1.0 \cdot 10^{-1}$
CCA5	$0.51 \pm 1.4 \cdot 10^{-2}$	$0.53 \pm 4.3 \cdot 10^{-2}$	$0.53 \pm 6.4 \cdot 10^{-2}$	$0.58 \pm 7.7 \cdot 10^{-2}$	$0.58 \pm 1.1 \cdot 10^{-1}$	$0.63 \pm 1.3 \cdot 10^{-1}$
CI-upper	0.51	0.52	0.53	0.54	0.55	0.57

TABLE D.1: Classification accuracy on the Vanthornhout dataset of LS and CCA feature extractors in function of the window length, given as mean accuracy ± standard deviation. CCAx ($x \in \mathbb{N}_0$) reflects subsequent filters generated by the CCA algorithm. The feature extractor with the highest classification accuracy per window length is marked in red. CI-upper denotes the upper bound of a 95% confidence interval of a binomial distribution with success-rate 0.5. LS and CCA2 result overall in the highest accuracies, although all mean accuracies lie above the upper bound of chance level and the methods do not significantly differ at window lengths 10 s and 30 s.

Window [s]	1	5	10	20	30	60
All	$1.00 \pm 4.3 \cdot 10^{-3}$	$1.00 \pm 2.7 \cdot 10^{-3}$	1.00 ± 0	1.00 ± 0	1.00 ± 0	1.00 ± 0
Delta	$0.87 \pm 1.3 \cdot 10^{-1}$	$0.92 \pm 1.1 \cdot 10^{-1}$	$0.94 \pm 8.9 \cdot 10^{-2}$	$0.94 \pm 9.9 \cdot 10^{-2}$	$0.94 \pm 8.6 \cdot 10^{-2}$	$0.96 \pm 8.6 \cdot 10^{-2}$
Theta	$0.86 \pm 1.2 \cdot 10^{-1}$	$0.96 \pm 5.0 \cdot 10^{-2}$	$0.98 \pm 2.9 \cdot 10^{-2}$	$0.98 \pm 2.2 \cdot 10^{-2}$	$0.97 \pm 4.3 \cdot 10^{-2}$	$0.95 \pm 7.1 \cdot 10^{-2}$
Alpha	$0.95 \pm 4.3 \cdot 10^{-2}$	$0.99 \pm 1.7 \cdot 10^{-2}$	$0.99 \pm 9.4 \cdot 10^{-3}$	$1.00 \pm 5.8 \cdot 10^{-3}$	1.00 ± 0	1.00 ± 0
Beta	$0.99 \pm 7.8 \cdot 10^{-3}$	$1.00 \pm 3.4 \cdot 10^{-3}$	$1.00 \pm 3.8 \cdot 10^{-3}$	1.00 ± 0	1.00 ± 0	$0.99 \pm 1.8 \cdot 10^{-2}$
CI-upper	0.51	0.52	0.53	0.54	0.55	0.57

TABLE D.2: Classification accuracy on the Vanthornhout dataset of the CSP feature extractor in function of the window length, given as mean ± standard deviation. The feature extractor with the highest classification accuracy per window length is marked in red. CI-upper denotes the upper bound of a 95% confidence interval of a binomial distribution with success-rate 0.5. Alpha and beta band CSP do not significantly differ from all frequency band CSP at a window length of 30 s, whereas delta and theta band CSP do. At a window length of 10 s, all frequency band CSP readily outperforms all individual CSP methods.

Window [s]	1	5	10	20	30	60
LS	$0.52 \pm 2.4 \cdot 10^{-2}$	$0.56 \pm 4.2 \cdot 10^{-2}$	$0.56 \pm 6.5 \cdot 10^{-2}$	$0.61 \pm 9.0 \cdot 10^{-2}$	$0.63 \pm 8.1 \cdot 10^{-2}$	$0.66 \pm 1.4 \cdot 10^{-1}$
LASSO	$0.52 \pm 2.2 \cdot 10^{-2}$	$0.54 \pm 2.7 \cdot 10^{-2}$	$0.53 \pm 8.8 \cdot 10^{-2}$	$0.57 \pm 4.3 \cdot 10^{-2}$	$0.55 \pm 1.3 \cdot 10^{-1}$	$0.51 \pm 1.6 \cdot 10^{-1}$
Entropy	$0.57 \pm 7.5 \cdot 10^{-2}$	$0.60 \pm 8.6 \cdot 10^{-2}$	$0.64 \pm 1.0 \cdot 10^{-1}$	$0.65 \pm 1.3 \cdot 10^{-1}$	$0.67 \pm 1.8 \cdot 10^{-1}$	$0.70 \pm 1.3 \cdot 10^{-1}$
KLD	$0.59 \pm 1.0 \cdot 10^{-1}$	$0.65 \pm 1.0 \cdot 10^{-1}$	$0.68 \pm 1.0 \cdot 10^{-1}$	$0.67 \pm 1.4 \cdot 10^{-1}$	$0.58 \pm 1.8 \cdot 10^{-1}$	$0.61 \pm 2.0 \cdot 10^{-1}$
LS-LASSO	$0.52 \pm 2.6 \cdot 10^{-2}$	$0.58 \pm 3.2 \cdot 10^{-2}$	$0.60 \pm 5.2 \cdot 10^{-2}$	$0.60 \pm 7.5 \cdot 10^{-2}$	$0.58 \pm 1.3 \cdot 10^{-1}$	$0.67 \pm 1.1 \cdot 10^{-1}$
LS-KLD	$0.59 \pm 1.0 \cdot 10^{-1}$	$0.66 \pm 9.1 \cdot 10^{-2}$	$0.69 \pm 9.6 \cdot 10^{-2}$	$0.68 \pm 1.4 \cdot 10^{-1}$	$0.65 \pm 1.5 \cdot 10^{-1}$	$0.68 \pm 1.6 \cdot 10^{-1}$
LS-entropy	$0.59 \pm 5.6 \cdot 10^{-2}$	$0.62 \pm 5.4 \cdot 10^{-2}$	$0.67 \pm 8.7 \cdot 10^{-2}$	$0.71 \pm 9.7 \cdot 10^{-2}$	$0.72 \pm 1.2 \cdot 10^{-1}$	$0.76 \pm 9.1 \cdot 10^{-2}$
LASSO-entropy	$0.58 \pm 7.0 \cdot 10^{-2}$	$0.62 \pm 6.9 \cdot 10^{-2}$	$0.65 \pm 1.1 \cdot 10^{-1}$	$0.67 \pm 1.1 \cdot 10^{-1}$	$0.66 \pm 1.8 \cdot 10^{-1}$	$0.68 \pm 8.4 \cdot 10^{-2}$
LASSO-KLD	$0.60 \pm 8.9 \cdot 10^{-2}$	$0.66 \pm 1.0 \cdot 10^{-1}$	$0.68 \pm 1.0 \cdot 10^{-1}$	$0.66 \pm 1.4 \cdot 10^{-1}$	$0.61 \pm 1.7 \cdot 10^{-1}$	$0.56 \pm 2.3 \cdot 10^{-1}$
KLD-entropy	$0.63 \pm 9.2 \cdot 10^{-2}$	$0.70 \pm 8.6 \cdot 10^{-2}$	$0.72 \pm 9.2 \cdot 10^{-2}$	$0.75 \pm 1.1 \cdot 10^{-1}$	$0.78 \pm 9.1 \cdot 10^{-2}$	$0.76 \pm 1.1 \cdot 10^{-1}$
LS-KLD-entropy	$0.63 \pm 9.4 \cdot 10^{-2}$	$0.71 \pm 9.0 \cdot 10^{-2}$	$0.74 \pm 8.5 \cdot 10^{-2}$	$0.77 \pm 1.0 \cdot 10^{-1}$	$0.77 \pm 1.2 \cdot 10^{-1}$	$0.80 \pm 9.3 \cdot 10^{-2}$
LS-LASSO-entropy	$0.59 \pm 5.0 \cdot 10^{-2}$	$0.63 \pm 5.7 \cdot 10^{-2}$	$0.68 \pm 9.5 \cdot 10^{-2}$	$0.71 \pm 1.0 \cdot 10^{-1}$	$0.71 \pm 1.3 \cdot 10^{-1}$	$0.72 \pm 1.0 \cdot 10^{-1}$
LS-LASSO-KLD	$0.60 \pm 9.0 \cdot 10^{-2}$	$0.67 \pm 8.6 \cdot 10^{-2}$	$0.68 \pm 9.1 \cdot 10^{-1}$	$0.69 \pm 1.3 \cdot 10^{-1}$	$0.64 \pm 1.7 \cdot 10^{-1}$	$0.63 \pm 2.0 \cdot 10^{-1}$
LASSO-KLD-entropy	$0.63 \pm 8.8 \cdot 10^{-2}$	$0.70 \pm 9.5 \cdot 10^{-2}$	$0.73 \pm 9.6 \cdot 10^{-2}$	$0.75 \pm 9.8 \cdot 10^{-2}$	$0.76 \pm 1.2 \cdot 10^{-1}$	$0.72 \pm 1.2 \cdot 10^{-1}$
LS-LASSO-KLD-entropy	$0.63 \pm 8.9 \cdot 10^{-2}$	$0.71 \pm 7.8 \cdot 10^{-2}$	$0.73 \pm 1.0 \cdot 10^{-1}$	$0.76 \pm 9.6 \cdot 10^{-2}$	$0.76 \pm 1.3 \cdot 10^{-1}$	$0.75 \pm 9.3 \cdot 10^{-2}$
CI-upper	0.51	0.52	0.53	0.54	0.55	0.57

TABLE D.3: Classification accuracy on the Vanthornhout dataset of combinations of LS, LASSO, KLD and entropy feature extractors in function of the window length, given as mean ± standard deviation. The feature extractor with the highest classification accuracy per window length is marked in red. CI-upper denotes the upper bound of a 95% confidence interval of a binomial distribution with success-rate 0.5. KLD refers to the early fusion KLD approach. Combinations including both KLD and entropy attain generally the highest classification accuracy.

Window [s]	1	5	10	20	30	60
All	$0.96 \pm 3.5 \cdot 10^{-2}$	$0.98 \pm 2.5 \cdot 10^{-2}$	$0.98 \pm 2.7 \cdot 10^{-2}$	$0.98 \pm 3.4 \cdot 10^{-2}$	$0.98 \pm 4.0 \cdot 10^{-2}$	$0.97 \pm 5.1 \cdot 10^{-2}$
Delta	$0.81 \pm 8.5 \cdot 10^{-2}$	$0.87 \pm 9.9 \cdot 10^{-2}$	$0.89 \pm 9.8 \cdot 10^{-2}$	$0.90 \pm 1.1 \cdot 10^{-1}$	$0.91 \pm 1.2 \cdot 10^{-1}$	$0.90 \pm 1.3 \cdot 10^{-1}$
Theta	$0.84 \pm 9.3 \cdot 10^{-2}$	$0.91 \pm 1.1 \cdot 10^{-1}$	$0.91 \pm 1.3 \cdot 10^{-1}$	$0.92 \pm 1.5 \cdot 10^{-1}$	$0.92 \pm 1.6 \cdot 10^{-1}$	$0.91 \pm 1.7 \cdot 10^{-1}$
Alpha	$0.84 \pm 1.3 \cdot 10^{-1}$	$0.89 \pm 1.5 \cdot 10^{-1}$	$0.89 \pm 1.7 \cdot 10^{-1}$	$0.88 \pm 1.8 \cdot 10^{-1}$	$0.89 \pm 1.7 \cdot 10^{-1}$	$0.89 \pm 1.9 \cdot 10^{-1}$
Beta	$0.92 \pm 8.3 \cdot 10^{-2}$	$0.93 \pm 8.1 \cdot 10^{-2}$	$0.94 \pm 7.7 \cdot 10^{-2}$	$0.93 \pm 8.1 \cdot 10^{-2}$	$0.93 \pm 8.4 \cdot 10^{-1}$	$0.93 \pm 1.1 \cdot 10^{-1}$
CI-upper	0.51	0.52	0.52	0.53	0.54	0.56

TABLE D.4: Classification accuracy on the Brouckmans-Dewit-Vanhaelen dataset dataset of the CSP feature extractor in function of the window length, given as mean ± standard deviation. The feature extractor with the highest classification accuracy per window length is marked in red. CI-upper denotes the upper bound of a 95% confidence interval of a binomial distribution with success-rate 0.5. All frequency band CSP numerically attains the highest mean accuracies for all windows. Nonetheless, these accuracies of all frequency band CSP, at window lengths of 10 and 30 s, are not significantly different ($\min(p)=8.8 \cdot 10^{-2}$, $\max(p)=2.4 \cdot 10^{-1}$) from individual CSP band accuracies.

Window [s]	1	5	10	20	30	60
Entropy	$0.69 \pm 1.1 \cdot 10^{-1}$	$0.72 \pm 1.6 \cdot 10^{-1}$	$0.74 \pm 1.5 \cdot 10^{-1}$	$0.77 \pm 1.8 \cdot 10^{-1}$	$0.79 \pm 1.8 \cdot 10^{-1}$	$0.76 \pm 2.3 \cdot 10^{-1}$
KLD	$0.63 \pm 1.3 \cdot 10^{-1}$	$0.69 \pm 1.5 \cdot 10^{-1}$	$0.69 \pm 1.6 \cdot 10^{-1}$	$0.71 \pm 1.5 \cdot 10^{-1}$	$0.69 \pm 1.5 \cdot 10^{-1}$	$0.70 \pm 1.5 \cdot 10^{-1}$
KLD-entropy	$0.74 \pm 1.1 \cdot 10^{-1}$	$0.80 \pm 1.1 \cdot 10^{-2}$	$0.81 \pm 1.2 \cdot 10^{-2}$	$0.82 \pm 1.4 \cdot 10^{-1}$	$0.82 \pm 1.5 \cdot 10^{-2}$	$0.79 \pm 1.6 \cdot 10^{-1}$
CI-upper	0.51	0.52	0.52	0.53	0.54	0.56

TABLE D.5: Classification accuracy on the Brouckmans-Dewit-Vanhaelen dataset dataset of combinations of KLD and entropy feature extractors in function of the window length, given as mean ± standard deviation. The feature extractor with the highest classification accuracy per window length is marked in red. CI-upper denotes the upper bound of a 95% confidence interval of a binomial distribution with success-rate 0.5. KLD refers to the early fusion KLD approach. All methods outperform chance level. The KLD-entropy combination performs significantly better, at 10 and 30 s decision windows, than the KLD and entropy methods separately.

Bibliography

[1] A. Biasiucci, B. Franceschiello, and M. M. Murray, "Electroencephalography," *Current Biology*, vol. 29, no. 3, pp. R80–R85, Feb. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960982218315513

[2] G. R. Müller-Putz, "Chapter 18 - Electroencephalography," in *Handbook of Clinical Neurology*, ser. Brain-Computer Interfaces, N. F. Ramsey and J. d. R. Millán, Eds. Elsevier, Jan. 2020, vol. 168, pp. 249–262. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444639349000184

[3] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal—state-of-the-art and guidelines," *Journal of Neural Engineering*, vol. 12, no. 3, p. 031001, Jun. 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/3/031001

[4] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," p. 7, 2007.

[5] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, Jul. 2012. [Online]. Available: https://www.pnas.org/content/109/29/11854

[6] D. Lesenfants and T. Francart, "The interplay of top-down focal attention and the cortical tracking of speech," *Scientific Reports*, vol. 10, no. 1, p. 6922, Dec. 2020. [Online]. Available: http://www.nature.com/articles/s41598-020-63587-3

[7] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, Sep. 1953. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1907229

[8] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015. [Online]. Available: https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bht355

[9] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, Oct. 1990. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.400247

[10] W. Kellermann, "Beamforming for Speech and Audio Signals," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, and M. Vorländer, Eds. New York, NY: Springer, 2008, pp. 691–702. [Online]. Available: https://doi.org/10.1007/978-0-387-30441-0_35

[11] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811918300338

[12] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-Based Decoding Of The Directional Focus Of Auditory Attention Using Common Spatial Patterns," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1557–1568, May 2021.

[13] W. Klonowski, "Everything you wanted to ask about EEG but were afraid to get the right answer," *Nonlinear Biomedical Physics*, vol. 3, p. 2, May 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698918/

[14] B. Somers, T. Francart, and A. Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *Journal of Neural Engineering*, vol. 15, no. 3, p. 036007, Jun. 2018.

[15] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, May 2017.

[16] J. Vanthornhout, L. Decruy, and T. Francart, "Effect of Task and Attention on Neural Tracking of Speech," *Frontiers in Neuroscience*, vol. 13, p. 977, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2019.00977

[17] P. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening, Modern Acoustics and Signal Processing*, Jan. 2013, pp. 33–56.

[18] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, Oct. 1995. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.414456

[19] Z.-N. Li, M. S. Drew, and J. Liu, "MPEG Audio Compression," in *Fundamentals of Multimedia*, ser. Texts in Computer Science, Z.-N. Li, M. S. Drew, and J. Liu, Eds. Cham: Springer International Publishing, 2014, pp. 457–482. [Online]. Available: https://doi.org/10.1007/978-3-319-05290-8_14

[20] S. S. Stevens, "The Measurement of Loudness," *The Journal of the Acoustical Society of America*, p. 15, Sep. 1995.

[21] B. Babadi and E. N. Brown, "A Review of Multitaper Spectral Analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1555–1564, May 2014.

[22] L. W. Couch, "Digital & Analog Communication Systems," in *Digital & Analog Communication Systems*, 8th ed. Edinburgh Gate: Pearson, 2013, pp. 436–513. [Online]. Available: https://www.pearson.com/store/p/digital-analog-communication-systems/P100001410611/9780133072716

[23] M. J. Prerau, R. E. Brown, M. T. Bianchi, J. M. Ellenbogen, and P. L. Purdon, "Sleep Neurophysiological Dynamics Through the Lens of Multitaper Spectral Analysis," *Physiology*, vol. 32, no. 1, pp. 60–92, Jan. 2017. [Online]. Available: https://www.physiology.org/doi/10.1152/physiol.00062.2015

[24] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, Sep. 1982.

[25] "Multitaper power spectral density estimate - MATLAB pmtm - MathWorks Benelux." [Online]. Available: https://nl.mathworks.com/help/signal/ref/pmtm.html#mw_243b8a58-f1e5-4df4-97f8-fece362dc840

[26] E. M. Zion Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder, "Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a 'Cocktail Party'," *Neuron*, vol. 77, no. 5, pp. 980–991, Mar. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3891478/

[27] K. C. Puvvada and J. Z. Simon, "Cortical Representations of Speech in a Multitalker Auditory Scene," *Journal of Neuroscience*, vol. 37, no. 38, pp. 9189–9196, Sep. 2017. [Online]. Available: https://www.jneurosci.org/content/37/38/9189

[28] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "EEG-based Auditory Attention Decoding: Towards Neuro-Steered Hearing Devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, Jul. 2021. [Online]. Available: http://arxiv.org/abs/2008.04569

[29] E. Edwards and E. F. Chang, "Syllabic (∼2–5 Hz) and fluctuation (∼1–10 Hz) ranges in speech and auditory processing," *Hearing Research*, vol. 305, pp. 113–134, Nov. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378595513002153

[30] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope," *Journal of the Association for Research in Otolaryngology: JARO*, vol. 19, no. 2, pp. 181–191, Apr. 2018.

[31] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[32] A. Bednar and E. C. Lalor, "Neural tracking of auditory motion is reflected by delta phase and alpha power of EEG," *NeuroImage*, vol. 181, pp. 683–691, Nov. 2018.

[33] O. Etard and T. Reichenbach, "Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise," *Journal of Neuroscience*, vol. 39, no. 29, pp. 5750–5759, Jul. 2019. [Online]. Available: https://www.jneurosci.org/content/39/29/5750

[34] J. C. F. de Winter, t. link will open in a new window Link to external site, S. D. Gosling, and J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," *Psychological Methods*, vol. 21, no. 3, pp. 273–290, Sep. 2016. [Online]. Available: https://www.proquest.com/docview/1790926596/abstract/AEF423FE88AE4979PQ/1

[35] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-Time Tracking of Selective Auditory Attention From M/EEG: A Bayesian Filtering Approach," *Frontiers in Neuroscience*, May 2018. [Online]. Available: https://www.proquest.com/docview/2306310020/abstract/8CFD258FFACA4EF0PQ/1

[36] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[37] J. Nocedal and S. J. Wright, in *Numerical Optimization*, 2nd ed., ser. Springer Series in Operations Research. New York: Springer, 2006, pp. 10–27.

[38] N. P. Hurley and S. T. Rickard, "Comparing Measures of Sparsity," *arXiv:0811.4706 [cs, math]*, Apr. 2009. [Online]. Available: http://arxiv.org/abs/0811.4706

[39] S. Rickard and M. Fallon, "The Gini index of speech," Jan. 2004.

[40] H. Dalton, "The Measurement of the Inequality of Incomes," *The Economic Journal*, vol. 30, no. 119, p. 348, Sep. 1920. [Online]. Available: https://www.jstor.org/stable/10.2307/2223525?origin=crossref

[41] M. O. Lorenz, "Methods of Measuring the Concentration of Wealth," *Publications of the American Statistical Association*, vol. 9, no. 70, p. 209, Jun. 1905. [Online]. Available: https://www.jstor.org/stable/2276207?origin=crossref

[42] A. de Cheveigné, M. Slaney, S. A. Fuglsang, and J. Hjortkjaer, "Auditory stimulus-response modeling with a match-mismatch task," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046040, Aug. 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/abf771

[43] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936. [Online]. Available: http://www.jstor.org/stable/2333955

[44] S. P. Boyd and L. Vandenberghe, in *Convex Optimization*. Cambridge, UK ; New York: Cambridge University Press, 2004, pp. 127–273.

[45] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda, and M. Hirano, "Quantification of EEG irregularity by use of the entropy of the power spectrum," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 3, pp. 204–210, Sep. 1991. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/001346949190138T

[46] H. Viertiö-Oja, V. Maja, M. Särkelä, P. Talja, N. Tenkanen, H. Tolvanen-Laakso, M. Paloheimo, A. Vakkuri, A. Yli-Hankala, and P. Meriläinen, "Description of the Entropy™ algorithm as applied in the Datex-Ohmeda S/5™ Entropy Module," *Acta Anaesthesiologica Scandinavica*, vol. 48, no. 2, pp. 154–161, 2004. [Online]. Available: http://onlinelibrary.wiley.com/doi/abs/10.1111/j.0001-5172.2004.00322.x

[47] T. T. Georgiou, "Distances between power spectral densities," *arXiv:math/0607026*, Jul. 2006. [Online]. Available: http://arxiv.org/abs/math/0607026

[48] T. Georgiou and A. Lindquist, "Kullback-Leibler approximation of spectral density functions," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 2910–2917, Nov. 2003.

[49] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. [Online]. Available: http://www.jstor.org/stable/2236703

[50] E. R. Dougherty, in *Dougherty, E: Random Processes for Image Signal Processing*, Bellingham, Wash, Nov. 1998, pp. 1–114.

[51] M. Aoyagi, T. Kiren, Y. Kim, Y. Suzuki, T. Fuse, and Y. Koike, "Optimal modulation frequency for amplitude-modulation following response in young children during sleep," *Hearing Research*, vol. 65, no. 1-2, pp. 253–261, Feb. 1993. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/037859559390218P

[52] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. Springer, 2017. [Online]. Available: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf

[53] J. T. Coull, C. D. Frith, R. S. Frackowiak, and P. M. Grasby, "A fronto-parietal network for rapid visual information processing: A PET study of sustained attention and working memory," *Neuropsychologia*, vol. 34, no. 11, pp. 1085–1095, Nov. 1996.

[54] J. J. Foxe, G. V. Simpson, and S. P. Ahlfors, "Parieto-occipital ∼1 0Hz activity reflects anticipatory state of visual attention mechanisms:," *NeuroReport*, vol. 9, no. 17, pp. 3929–3933, Dec. 1998. [Online]. Available: http://journals.lww.com/00001756-199812010-00030

[55] E. Fedorenko, J. Duncan, and N. Kanwisher, "Broad domain generality in focal regions of frontal and parietal cortex," *Proceedings of the National Academy of Sciences*, vol. 110, no. 41, pp. 16 616–16 621, Oct. 2013. [Online]. Available: https://www.pnas.org/content/110/41/16616

[56] S. Kweldju, "Neurobiology Research Findings: How the Brain Works during Reading," *PASAA: Journal of Language Teaching and Learning in Thailand*, vol. 50, pp. 125–142, 2015. [Online]. Available: https://eric.ed.gov/?id=EJ1088308

[57] N. Kaongoen, J. H. Choi, and S. Jo, "Speech-imagery-based BCI system using ear-EEG," *Journal of Neural Engineering*, Dec. 2020.

[58] A. Belyavin and N. A. Wright, "Changes in electrical activity of the brain with vigilance," *Electroencephalography and Clinical Neurophysiology*, vol. 66, no. 2, pp. 137–144, Feb. 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0013469487901830

[59] W. Klimesch, "Alpha-band oscillations, attention, and controlled access to stored information," *Trends in Cognitive Sciences*, vol. 16, no. 12, pp. 606–617, Dec. 2012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661312002434

[60] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-r. Muller, "Optimizing Spatial filters for Robust EEG Single-Trial Analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[61] Z. Koles, "The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 6, pp. 440–447, Dec. 1991. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/001346949190163X

[62] I. Xygonakis, A. Athanasiou, N. Pandria, D. Kugiumtzis, and P. D. Bamidis, "Decoding Motor Imagery through Common Spatial Pattern Filters at the EEG Source Space," *Computational Intelligence and Neuroscience*, vol. 2018, p. e7957408, Aug. 2018. [Online]. Available: https://www.hindawi.com/journals/cin/2018/7957408/

[63] D. Lesenfants, D. Habbal, C. Chatelle, A. Soddu, S. Laureys, and Q. Noirhomme, "Toward an Attention-Based Diagnostic Tool for Patients With Locked-in Syndrome," *Clinical EEG and Neuroscience*, vol. 49, no. 2, pp. 122–135, Mar. 2018. [Online]. Available: https://doi.org/10.1177/1550059416674842

[64] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. New York: Springer, 2006.

[65] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, Feb. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X03000964

[66] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901. [Online]. Available: https://doi.org/10.1080/14786440109462720

[67] F. Castells, P. Laguna, L. Sörnmo, A. Bollmann, and J. M. Roig, "Principal Component Analysis in ECG Signal Processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–21, Dec. 2007. [Online]. Available: https://asp-eurasipjournals.springeropen.com/articles/10.1155/2007/74580

[68] A. N. Tikhonov, "On the solution of ill-posed problems and the method of regularization," p. 5, 1963.

[69] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, "Canonical Correlation Analysis when the Data are Curves," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 55, no. 3, pp. 725–740, 1993. [Online]. Available: http://www.jstor.org/stable/2345883

[70] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE transactions on bio-medical engineering*, vol. 58, no. 2, pp. 355–362, Feb. 2011.

[71] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, Feb. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811913010914

[72] M. J. Crosse, N. J. Zuk, G. M. D. Liberto, A. Nidiffer, S. Molholm, and E. C. Lalor, "Linear Modeling of Neurophysiological Responses to Naturalistic Stimuli: Methodological Considerations for Applied Research," May 2021. [Online]. Available: https://psyarxiv.com/jbz2w/

[73] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Dec. 2015. [Online]. Available: http://www.degruyter.com/document/doi/10.1515/9781400874668/html

[74] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, Dec. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494619302947

[75] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding," *IEEE journal of biomedical and health informatics*, vol. PP, Apr. 2021.

[76] G. Csurka, Ed., *Domain Adaptation in Computer Vision Applications*, ser. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2017. [Online]. Available: http://link.springer.com/10.1007/978-3-319-58347-1

[77] ——, *Domain Adaptation in Computer Vision Applications*, ser. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2017. [Online]. Available: http://link.springer.com/10.1007/978-3-319-58347-1

[78] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," *arXiv:1409.7495 [cs, stat]*, Feb. 2015. [Online]. Available: http://arxiv.org/abs/1409.7495

[79] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, Jun. 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aab2f2

[80] P. Xiao, B. Du, and X. Li, "An Unsupervised Domain Adaptation Algorithm Based on Canonical Correlation Analysis," in *Computer Vision*, ser. Communications in Computer and Information Science, J. Yang, Q. Hu, M.-M. Cheng, L. Wang, Q. Liu, X. Bai, and D. Meng, Eds. Singapore: Springer, 2017, pp. 26–37.

[81] K. R. Anoop, R. Subramanian, V. Vonikakis, K. Ramakrishnan, and S. Winkler, "On the utility of canonical correlation analysis for domain adaptation in multi-view headpose estimation," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sep. 2015, pp. 4708–4712.

[82] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Subspace Alignment For Domain Adaptation," *arXiv:1409.5241 [cs]*, Oct. 2014. [Online]. Available: http://arxiv.org/abs/1409.5241

[83] B. Gong, K. Grauman, and F. Sha, "Geodesic Flow Kernel and Landmarks: Kernel Methods for Unsupervised Domain Adaptation," in *Domain Adaptation in Computer Vision Applications*, ser. Advances in Computer Vision and Pattern Recognition, G. Csurka, Ed. Cham: Springer International Publishing, 2017, pp. 59–79. [Online]. Available: https://doi.org/10.1007/978-3-319-58347-1_3

[84] Y. S. Aurelio, G. M. de Almeida, C. L. de Castro, and A. P. Braga, "Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function," *Neural Processing Letters*, vol. 50, no. 2, pp. 1937–1949, Oct. 2019. [Online]. Available: https://doi.org/10.1007/s11063-018-09977-1

[85] E. Brouckmans and L. Dewit-Vanhaelen, "Dataset (Unpublished raw data)," 2022.

[86] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974. [Online]. Available: http://www.jstor.org/stable/2984809

[87] "Canonical correlation - MATLAB canoncorr - MathWorks Benelux." [Online]. Available: https://nl.mathworks.com/help/stats/canoncorr.html

[88] N. Sun, "Canonical Correlation Analysis (CCA)," p. 4.

[89] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press, 2006.

[90] K. V. Mardia, "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, vol. 57, no. 3, pp. 519–530, 1970. [Online]. Available: http://www.jstor.org/stable/2334770

[91] G. E. P. Box, "A general distribution theory for a class of likilihood criteria," *Biometrika*, vol. 36, no. 3-4, pp. 317–346, Dec. 1949. [Online]. Available: https://doi.org/10.1093/biomet/36.3-4.317

[92] H.-M. Kaltenbach, "Hypothesis Testing," in *A Concise Guide to Statistics*, ser. SpringerBriefs in Statistics, H.-M. Kaltenbach, Ed. Berlin, Heidelberg: Springer, 2012, pp. 53–75. [Online]. Available: https://doi.org/10.1007/978-3-642-23502-3_3

[93] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: http://www.jstor.org/stable/2346101

[94] N. M. Laird and J. H. Ware, "Random-Effects Models for Longitudinal Data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982. [Online]. Available: http://www.jstor.org/stable/2529876

[95] S. R. Searle, G. Casella, and C. E. McCulloch, "Variance Components," *John Wiley & Sons*, p. 537, 1992.

[96] A. Gałecki and T. Burzykowski, *Linear Mixed-Effects Models Using R*, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2013. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-3900-4

[97] "Estimating Parameters in Linear Mixed-Effects Models - MATLAB & Simulink - MathWorks Benelux." [Online]. Available: https://nl.mathworks.com/help/stats/estimating-parameters-in-linear-mixed-effects-models.html

[98] "Linear Mixed-Effects Models - MATLAB & Simulink - MathWorks Benelux." [Online]. Available: https://nl.mathworks.com/help/stats/linear-mixed-effects-models.html

[99] R. A. Fisher, "Statistical Methods for Research Workers," in *Breakthroughs in Statistics: Methodology and Distribution*, ser. Springer Series in Statistics, S. Kotz and N. L. Johnson, Eds. New York, NY: Springer, 1992, pp. 66–70. [Online]. Available: https://doi.org/10.1007/978-1-4612-4380-9_6

[100] E. Ostertagova and O. Ostertag, "Methodology and Application of One-way ANOVA," *American Journal of Mechanical Engineering*, vol. 1, pp. 256–261, Nov. 2013.

[101] "Analysis of variance for linear mixed-effects model - MATLAB - MathWorks Benelux." [Online]. Available: https://nl.mathworks.com/help/stats/linearmixedmodel.anova.html

[102] "Development and normative data for the Flemish/Dutch Matrix test - KU Leuven." [Online]. Available: http://limo.libis.be/primo-explore/fulldisplay/LIRIAS1777721/Lirias

[103] K. Luyckx, H. Kloots, E. Coussé, and S. Gillis, *Klankfrequenties in Het Nederlands*, Jan. 2007.

[104] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003. [Online]. Available: https://www.jmlr.org/papers/v3/guyon03a

[105] A. V. Oppenheim, *Discrete-Time Signal Processing*. Pearson Education, 1999.

[106] C. Brodbeck, L. E. Hong, and J. Z. Simon, "Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech," *Current Biology*, vol. 28, no. 24, pp. 3976–3983.e5, Dec. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096098221831409X

[107] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech," *Current Biology*, vol. 28, no. 5, pp. 803–809.e3, Mar. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982218301465

[108] M. Gillis, J. Vanthornhout, J. Z. Simon, T. Francart, and C. Brodbeck, "Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics," *The Journal of Neuroscience*, vol. 41, no. 50, pp. 10 316–10 329, Dec. 2021. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0812-21.2021

[109] E. N. Lorenz, "Deterministic Nonperiodic Flow," *Journal of the Atmospheric Sciences*, vol. 20, no. 2, pp. 130–141, Mar. 1963. [Online]. Available: http://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml

[110] J. Gao, Y. Cao, W.-w. Tung, and J. Hu, *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond*. John Wiley & Sons, Dec. 2007.

[111] P. Grassberger and I. Procaccia, "Measuring the strageness of strange attractors," *North-Holland Publishing Company*, p. 20, May 1983.

[112] B. B. Mandelbrot, *The Fractal Geometry of Nature*. Henry Holt and Company, 1983.

[113] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E*, vol. 49, no. 2, pp. 1685–1689, Feb. 1994. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.49.1685

[114] A.-L. Barabási, A.-L. ászl ó Barabási, and H. E. Stanley, *Fractal Concepts in Surface Growth*. Cambridge University Press, Apr. 1995.

[115] M. Lavanga, O. De Wel, A. Caicedo, E. Heremans, K. Jansen, A. Dereymaeker, G. Naulaers, and S. Van Huffel, "Automatic quiet sleep detection based on multifractality in preterm neonates: Effects of maturation," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2017, pp. 2010–2013.

[116] C. Varon and S. Van Huffel, "Complexity and Nonlinearities in Cardiorespiratory Signals in Sleep and Sleep Apnea," in *Complexity and*

Nonlinearity in Cardiovascular Signals, R. Barbieri, E. P. Scilingo, and G. Valenza, Eds. Cham: Springer International Publishing, 2017, pp. 503–537. [Online]. Available: https://doi.org/10.1007/978-3-319-58709-7_19

[117] M. J. Monesi, B. Accou, J. Montoya-Martinez, T. Francart, and H. V. Hamme, "An LSTM Based Architecture to Relate Speech Stimulus to Eeg," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 941–945.

[118] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *eLife*, vol. 10, p. e56481, Apr. 2021. [Online]. Available: https://doi.org/10.7554/eLife.56481

[119] N. Das, J. Zegers, H. V. hamme, T. Francart, and A. Bertrand, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *Journal of Neural Engineering*, vol. 17, no. 4, p. 046039, Aug. 2020. [Online]. Available: https://doi.org/10.1088/1741-2552/aba6f8

[120] V. N. Vapnik, *The Vicinal Risk Minimization Principle and the SVMs*, ser. Statistics for Engineering and Information Science, V. N. Vapnik, Ed. New York, NY: Springer, 2000. [Online]. Available: https://doi.org/10.1007/978-1-4757-3264-1_9

[121] M. K. Cain, Z. Zhang, and K.-H. Yuan, "Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation," *Behavior Research Methods*, vol. 49, no. 5, pp. 1716–1735, Oct. 2017. [Online]. Available: https://doi.org/10.3758/s13428-016-0814-1

[122] M. Friendly and M. Sigal, "Visualizing Tests for Equality of Covariance Matrices," *The American Statistician*, vol. 74, no. 2, pp. 144–155, Apr. 2020. [Online]. Available: https://doi.org/10.1080/00031305.2018.1497537