NAIST-IS-DD1561019

Doctoral Dissertation

Noise-Removal From Electroencephalographies Toward Single-Trial Cognitive State Analysis

Hayato Maki

February 25, 2018

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Hayato Maki

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Associate Professor Sakriani Sakti	(Co-supervisor)
Assistant Professor Hiroki Tanaka	(Co-supervisor)
Professor Shin Ishii	(Kyoto University)

Noise-Removal From Electroencephalographies Toward Single-Trial Cognitive State Analysis[*]

Hayato Maki

Abstract

It is well known that the signal-to-ratio of electroencephalography (EEG) is low, which makes it challenging to perform estimation or information extraction using single-trial EEG. Single-trial analysis of EEG has gathered attention from researchers because it can bring new insights about cognitive or emotional states of human beings. In this doctoral dissertation, noise-removal from single-trial event-related potentials (ERP) and perceived quality prediction of synthesized speeches are investigated, respectively.

Noise-removal from ERP was tackled in two different approaches. First, considering the application of tensor factorization to ERP data, a novel regularization and initialization methods incorporating the geometrical information of EEG electrode location are respectively proposed, whose effectiveness was shown experimentally. Second, we proposed a multidimensional probabilistic generative model of EEG activities, whose maximum likelihood estimators of model parameters were utilized to construct a noise-removing filter. In addition, incorporating prior knowledge of EEG signals as prior distributions of model parameters was proposed that were calculated using previously recorded EEG signals.

Moreover, subjective quality evaluation scores of synthesized speeches were predicted using a single-trial EEG that was recorded during listening to the speeches. Consequently, it was shown that subjective rating scores can be predicted by a single-trial EEG, and effective features for the prediction was neurophysiologically plausible.

[*] Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1561019, February 25, 2018.

Keywords:

electroencephalography, tensor factorization, noise-removal, single-trial analysis

脳波による単一試行認知状態推定を目的としたノイズ除去法[*]

真木 勇人

内容梗概

脳波 (electroencephalography) の信号対ノイズ比は低いため，単一の脳波信号を対象とした推論や情報抽出は通常難しいことが知られている．単一試行の脳波解析技術が確立すれば，認知や感情の状態に関する分析の高精度化や，これまで不可能だった実験デザインが可能になることが期待できる．本論文では，単一試行の事象関連電位，及び合成音声聴取時の反応を解析する方法についてそれぞれ研究を行った．事象関連電位の抽出については，以下二つの異なるアプローチから研究を行った．(１) 脳波を時間や周波数など複数のモダリティをもった多次元配列データと捉え，テンソル分解の適用について検討した．テンソル分解の最適化において，頭皮上における脳波計の電極配置を事前知識とした正則化及び初期化の方法を提案し，これによって事象関連電位の抽出精度が向上することが確認された．(２) 複数のモダリティに対応する多変量の確率的生成モデルを用いて脳波の振幅生成を表現し，事後確率が最大となるパラメタを利用して信号分離を行う方法について提案した．また，既存の脳波データを用いてパラメタの事前分布を導入することで推定を安定化させることを提案した．また，合成音声聴取時の脳波と主観評価との関係を統計的に学習することによって，合成音声の品質予測を行う方法について検討した．結果，合成音声に対する主観評価をある程度予測可能なこと，神経生理学的に妥当な予測モデルを学習可能なことが示された．

キーワード

脳波, ノイズ除去, テンソル分解, 単一試行解析

[*] 奈良先端科学技術大学院大学 情報科学研究科 博士論文, NAIST-IS-DD1561019, 2018 年 2 月 25 日.

Acknowledgements

この博士論文は，奈良先端科学技術大学院大学 知能コミュニケーション研究室で行われた研究についてまとめたものである．五年間に渡る研究生活の間，多くの方々にたいへんお世話になった．それらの助けがなければこの論文は決して完成しなかったので，ここで感謝の意を表したい．

まず，主指導教員としてご指導くださった知能コミュニケーション研究室の中村哲教授に心より感謝申し上げる．ころころと気が変わりやすいわたしを厳しく，かつ寛容に導き続けてくださった．工学研究者としての姿勢について，仕事について，人生について，わたしなりに中村教授から学び，成長できたと感じている．また，インターンシップや留学など，わたしのために様々な機会を与えてくださった．極めて恵まれた，わたしの人生において貴重な五年間であったと思う．重ねて厚く御礼申し上げたい．

副指導教員として貴重なコメントをくださった奈良先端科学技術大学院大学自然言語処理研究室の松本裕治教授に感謝申し上げる．

学外から博士論文の審査員を務めてくださった京都大学 論理生命学研究室の石井信教授に感謝申し上げる．20013 年には株式会社国際電気通信基礎技術研究所 (ATR) 動的脳イメージング研究室室長として，当時修士課程一年次でまだこの研究分野についてほとんど無知であったわたしをインターンシップに迎えてくださった．

知能コミュニケーション研究室の Sakriani Sakti 特任准教授に感謝申し上げる．研究グループの指導教員として日頃から多くの助言をくださった．I I sincerely appreciate Associate Professor Sakriani Sakti for her continuous encouragement and advices at the meetings.

知能コミュニケーション研究室の田中宏季助教に感謝申し上げる．最初の三年間は同じ学生として，後の二年間は最も身近な教員として，唐突で飛躍の多いのわたしの話に辛抱強く耳を傾けてくださった．

名古屋大学の戸田智基教授に感謝申し上げる．研究の立ち上げからお世話になり，論文の書き方やプレゼンテーション法まで含め，非常に多くの助言をくださり，わたしの研究者としての基礎を作って頂いたと思う．そのおかげで，まがりなりにも博士課程を貫徹することができた．重ねて御礼申し上げる．

インターンシップとして迎えて頂いた Swartz Center for Computational Neuroscience を主宰する Scott Makeig 博士に感謝申し上げる．議論のために多くの

時間を割き，わたしの下手な英語に辛抱強く耳を傾けてくださった．I sincerely appreciate Dr. Scott Makeig's corporation to my research at Swartz Center for Computational Neuroscience. He kindly accepted me as a internship, spare his time for me, and even listened to my poor English.

Swartz Center for Computational Neuroscience の宮腰誠博士に感謝申し上げる．わたしの研究のために多くの時間を割いてくださり，公私ともにお世話になった．

Carnegie Mellon University の Graham Neubig 博士に感謝申し上げる．日頃のミーティングにおいて常に的確なコメントをくださった．また，わたしが英語論文を執筆した際には，何度も添削をしてくださった．

知能コミュニケーション研究室の吉野幸一郎助教に感謝申し上げる．日頃のミーティングにおいてアドバイスを頂いた．

知能コミュニケーション研究室秘書の松田真奈美氏に感謝申し上げる．書類の提出が遅れがちなわたしを我慢強く，寛容にサポートして頂いた．

人数が多いためここですべての方の名前を挙げることはしないが，知能コミュニケーション研究室の学生諸氏には，日頃から公私ともにお世話になった．

最後に，田舎の決して裕福ではない家庭に生まれたわたしにとって，大学院での研究や留学など本来は望むべくもないものである．にもかかわらず，わたしがこのような幸運に恵まれたのは，すべて父と母の献身的な支援のおかげにほかならない．わたしをここまで導いてくれた二人に，深く感謝したい．

Contents

List of Figures

List of Tables

Abbreviations

Table 1: Abbreviations

Abbreviation	Definition
BMI	Brain-machine interface
CPD	Canonical polyadic decomposition
EEG	Electroencephalography
EM	Expectation Maximization
ERP	Event-related potential
fMRI	Functional magnetic resonance imaging
GCPD	Graph regularized canonical polyadic decomposition
GFB	Graph Fourier Basis
HOPLS	Higher order partial least square
HOSVD	Higher order singular value decomposition
ICA	Independent component analysis
MAP	Maximum a posteriori
MEG	Magnetoencephalography
SVD	Singular value decomposition
PCA	Principal component analysis
PLS	Partial least square

Chapter 1

Introduction

1.1 Background

Nearly one hundred years ago, a German psychiatrist Hans Berger reported that he recorded the electrical activity of human brain by placing an electrode on the scalp [1]. Many scientists thought the Berger's finding was spurious initially but the same phenomenon was also reported by other researchers after a few years, which led to the acceptance of the findings as a real phenomenon. The electrical activity has been called *electroencephalography*, or shortly EEG. The word can be divided into subwords *electro*, *encephalo*, and *graphy*. *Encephalo* can be further divided into *en* and *cephalo*. *En* comes from ancient Greek and means "in," and *cephalo* also originates from an ancient Greek word *kephale*, which means "head." Therefore, EEG literally means *electrical recording of inside head*.

Over the decades, EEG have been widely used by researchers as a tool to study *inside head*. For example, in cognitive science, language processing [2–7], face recognition [8–10], emotions [11–13] have been studied using EEG. Clinical application of EEG includes the investigations of Alzheimer's disease [14–16], epilepsy [17–19], attention-deficit hyperactivity disorder (ADHD) [20–23]. Furthermore, endeavor to utilize it for engineering applications has also been continued. For example, EEG-based brain-machine interface (BMI) has attracted much attention from many researchers with different backgrounds for the purpose of assisting handicapped people and even augmenting our lifestyle [24–29]. Mental monitoring is an emerging

EEG application where mental or physiological condition including engagement, fatigue, or drowsiness of subjects are estimated from their EEG [30–32].

The representative applications of EEG are summarized in Figure 1.1. EEG potentials can be roughly divided into two categories. The first one is spontaneous potential, which is generated without specific stimuli presentation. On the other hand, the second one is evoked potential, which is generated as a response to specific stimuli (light, sound, sentence, image, etc). Especially, evoked potential is called event-related potential (ERP) if stimuli cause cognitive loads on a subject. Measuring ERP is one of the most widely used method to investigate human cognitive function [33, 34].

1.2 Research Objective

It is well known that the signal-to-ratio of electroencephalography (EEG) is low, which makes it difficult to get useful information from one EEG signal. Therefore, EEG is usually analyzed by aggregating multiple signals. For example, multiple EEG signals are recorded during repeated stimuli presentations to a subject. Then, the signals are averaged to cancel out noise, which is explained again in Section 2.1.1.

Single-trial EEG analysis means individually analyzing each EEG signal obtained by single stimuli presentation. It was proposed in [36] and has gathered attention from researchers because it can bring new insights about cognitive or emotional states of human beings. For example, it can allow us to investigate a subject's response variability among stimuli, which is impossible using conventional averaging [37–39].

Single-trial EEG analysis can be classified into two types, noise-removal and prediction (Figure 1.2). A single-trial noise-removal method inputs an observed single-trial EEG signal and outputs a cleaned signal. Definitions of *signal* and *noise* depend on research interests. For example, if you are interested in evoked potentials, spontaneous potentials are noise. In addition, artificial potentials are also noise that can be caused by body movements. On the other hand, spontaneous potentials are signal and evoked and artificial potentials are noise if you want to analyze the former.

In single-trial prediction problems, input is an observed single-trial EEG signal as

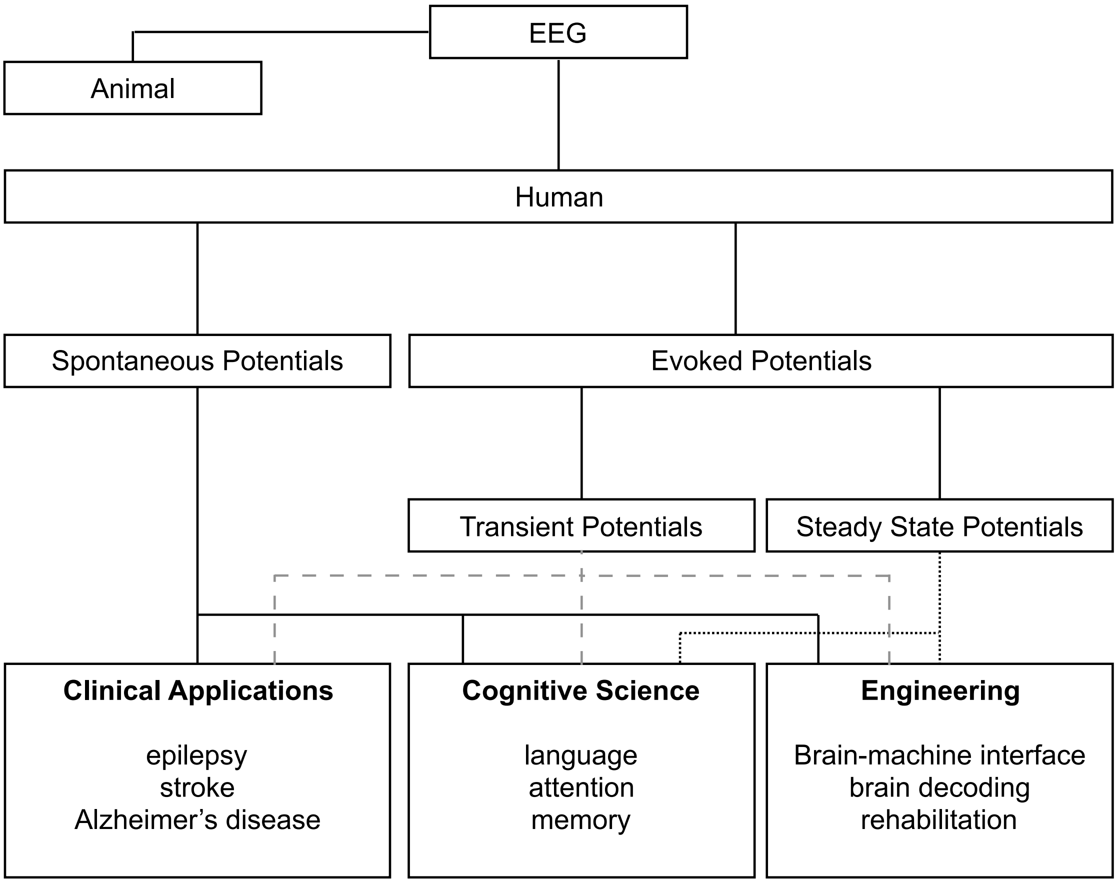


Figure 1.1: A summary of EEG potentials and their usage. Based on the Figure 1-3 in [35]. In this study, analysis methods for evoked transient potential are investigated.

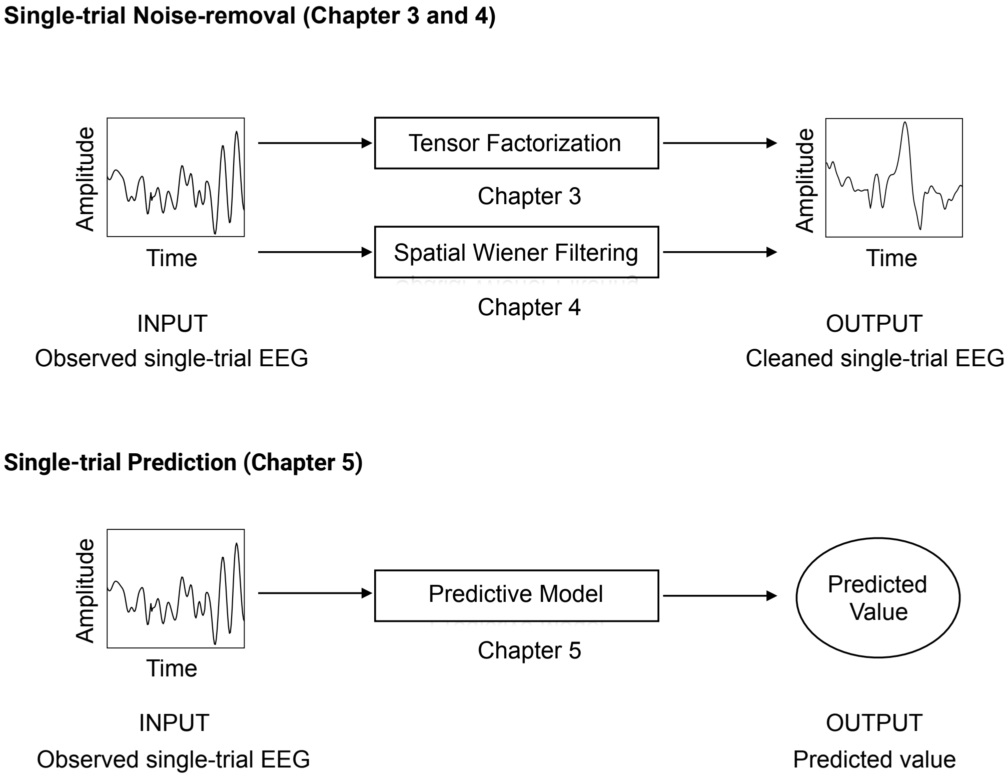


Figure 1.2: Two types of single-trial EEG analysis problems. In both of the problems, input is an observed single-trial EEG signal. (Upper) In single-trial noise-removal, output is a cleaned signal. (Lower) In single-trial prediction, output is a predicted value. The former is investigated in Chapter 3 and 4, and the latter is in Chapter 5.

same as noise-removal problems, output is a predicted value. The objective of this doctoral dissertation is to investigate both of the problems with a focus on analysis of evoked potentials.

1.3 Structure of Dissertation

The rest of this dissertation is structured as follows:

In Chapter 2, basic knowledges of EEG and its analysis methods are described.

Chapter 3 addresses a single-trial noise-removal problem investigating application of tensor factorization. We propose a novel tensor factorization algorithm that incorporates the geometric structure of the electrode location. Canonical polyadic decomposition (CPD), which is one of the tensor factorization methods is extended by adding a regularization term that considers the graphical information of EEG electrode location and controls the spatial smoothness of signals on a scalp surface. An initialization method considering the smoothness is also proposed. The geometric structure of EEG electrode location is expressed as an undirected graph where the similarities between electrodes are defined by their relative distances on a scalp.

Chapter 4 addresses a single-trial noise-removal problem, where we use a spatial Wiener filtering for noise-removal. A probabilistic generative model is created to express amplitude generation of EEG in a high dimensional space corresponding to multiple modes of EEG data, and its parameters are used to construct a noise-removing filter. Unifying view of the tensor factorization and spatial Wiener filtering is also explained.

In Chapter 5, we tackle to predict perceived qualities of synthesized speeches using EEG, which is an example of single-trial prediction problem. Subjective rating scores of speech quality are predicted by regression analysis with EEG features. Quality prediction of multi media based on EEG is challenging due to a small number of training trials, a low signal-to-noise ratio, and a high correlation among featrures extracted from EEG. We incorporate the structure of features to improve prediction performance.

Chapter 6 concludes the entire of this dissertation and gives possible directions for the future researches.

Chapter 2

Basics of EEG and its Analysis

2.1 EEG

EEG measures voltage fluctuations generated by the synchronized activities of numerous neurons using electrodes placed on scalp surface as shown in Figure 2.1.

Compared with other brain activity measurements, EEG has the following advantages. (1) *Non-invasiveness*: Recording EEG doesn't hurt a subject's body and have safety-related restriction while positron emission tomography (PET) exposes a subject to radiation. Micro electrode measurement inserts an electrode into the brain, which limits subjects to those who have medical reasons. (2) *High temporal resolution*: The temporal resolution of EEG is one millisecond (ms) or better, wheare those of functional magnetic resonance imaging (fMRI) and PET are several hundreds ms at best [33]. (3) *Portability*: EEG can be recorded almost everywhere and some EEG recording equipments are wireless [40] while conventional ones are wired to a data recording device. In contrast, magnetoencephalography (MEG) has to be used in a magnetically shield room and fMRI is also fixed to an experimental room, which makes it impossible to record brain activities in daily environment.

On the other hand, EEG has the following drawbacks. (1) *Low spatial resolution*: Field potentials that are generated by neuron activities come through cerebral spinal fluid and skull to scalp EEG electrodes, which causes volume conduction [35]. Therefore, a recorded EEG signal is a spatial average of neural activities within about 10–40 cm^2 of cortical sheets [24]. (2) *Low signal-to-noise ratio*: Various factors can

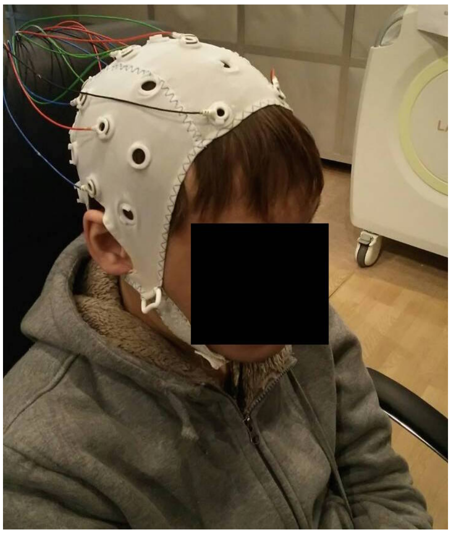

[redacted]

Figure 2.1: A man wearing an EEG cap.

affect noise level of EEG including an used EEG equipment, an electrical shield in an experimental room, electrode impedances, and the temperature and humidity in a recording place [41]. Signal power in delta (1–4 Hz), beta (12–30 Hz), or gamma (30–45 Hz) frequency bands are important to study brain functions but it is difficult to distinguish between them and muscle electricity [35]. Moreover, EEG get contaminated by artifacts seriously even by a small body movement, for example an eye blink. Artifacts often produce much higher amplitude than brain activities, which severely limits possible experimental paradigms. For example, when movie stimuli are presented to a subject with a wide monitor for a cognitive task, EEG data is likely to be contaminated by ocular and other body movement artifacts because seeing a wide range elicits movements of eye balls and the neck, which produces artificial potentials and conceals EEG activities of interest.

2.1.1 ERP

ERP is amplitude change of an EEG waveform caused by stimuli presented to a subject [33, 34]. Because it is always recorded mixed with spontaneous potentials (spontaneous potentials are also called background EEG in contrast to ERP) whose voltage are bigger than that of ERP, it is usually not visible. Therefore, synchronous averaging of multiple ERPs that are obtained by repeated stimuli presentation is commonly performed to attenuate spontaneous potentials.

A waveform of averaged ERP usually has a series of characteristic peaks that are named using their polarity and latency from the stimuli onset, for example N1 or N100 (the negative voltage shift that appears 100 ms after the stimuli onset) and P3 or P300 (the positive voltage shift that appears 300 ms after the stimuli onset). Each peak is called *an ERP component*.

There are a large body of studies that investigate relation between ERP components and cognitive functions. For example, the P3 component is related to stimuli categorization and cognitive resource allocation [7, 34, 42], and often utilized in brain-machine interfaces due to its stable appearance [27–29]. The N400 components is related to language processing in the brain [34, 43, 44]. The N170 component is related to perception of faces and expertise objects [34, 45, 46].

Examples of single-trial and averaged ERPs are shown in Figure 2.2, where a result of oddball paradigm is plotted [27, 47]. The oddball paradigm is the most widely used ERP experimental paradigm, where subjects are presented a series of stimuli that fall into two classes, target and non-target stimuli. Subjects make a response only to target stimuli (button press, count the number of their presences) ignoring the others. In Figure 2.2, finding each ERP component elicited by each stimulus is difficult, whereas it can be clearly seen the both of the averaged waveforms have the N1 component, and the averaged target trial waveform has the larger P3 component than non-target.

2.2 Noise-removal from EEG Data

2.2.1 Independent Component Analysis

Independent component analysis (ICA) has been widely studied to solve the blind source separation (BSS) problem [48, 49], and frequently used to remove noise from EEG [37, 50, 51]. The BSS problem is depicted in Figure 2.3. We observe signals that are generated from I sources using J sensors. The goal of the BSS problem

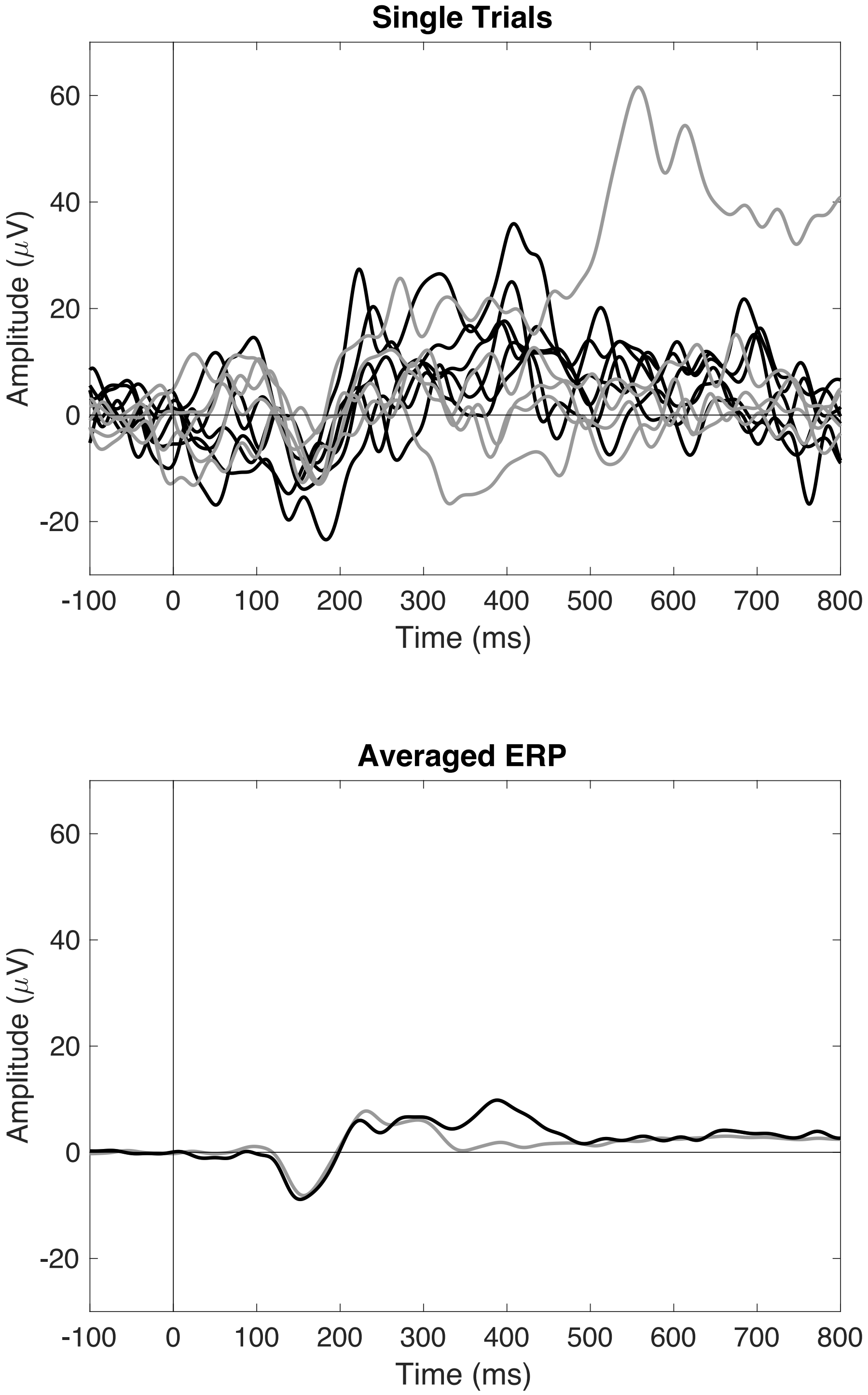


Figure 2.2: (upper) Example of five single-trial ERPs of target (black) and non-target (gray) trials. (lower) Averaged waveforms of the two stimuli classes.

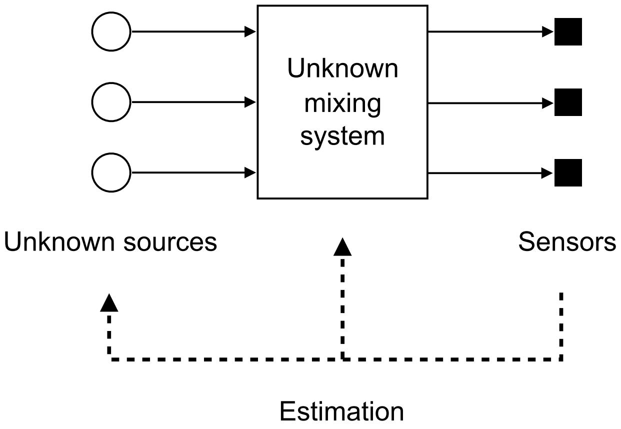


Figure 2.3: A schematic image of the BSS problem with three sources and three sensors. In an EEG case, sources are numerous synapses and sensors are electrodes on the scalp.

is to estimate sources without knowledge of sources and the mixing system. If the number of sources is equal to or less than that of sensors ($I \leq J$), the problem is called *the determined BSS*. On the other hand, if we have fewer number of sensors than sources ($I > J$), it is called *the under-determined BSS*.

If the mixing system is linear and instantaneous, BSS can be written as:

$$\begin{gathered} \boldsymbol{x}(t) = \mathrm{A}\boldsymbol{s}(t), \qquad (2.1) \\ \boldsymbol{x} \in \mathbb{R}^J, \quad \boldsymbol{s} \in \mathbb{R}^I, \quad \mathrm{A} \in \mathbb{R}^{J \times I}, \end{gathered}$$

where $\boldsymbol{x}(t)$ and $\boldsymbol{s}(t)$ are respectively the observed signal and the sources at the time t, and A represents the linear mixing system. The strategy of ICA to solve BSS is to find an estimator of inverse matrix of the mixing matrix $\hat{\mathrm{W}} = \mathrm{A}^{-1}$ and obtain source estimators $\hat{\boldsymbol{s}}$ as:

$$\hat{\boldsymbol{s}}(t) = \hat{\mathrm{W}}\boldsymbol{x}(t). \qquad (2.2)$$

W is estimated so that the source estimators are as independent as possible. ICA solves the BSS problem up to the scaling and permutation ambiguities between estimated sources but can not handle the under-determined BSS because it assumes the existence of the inverse of the mixing matrix.

2.2.2 Tensor Factorization

A tensor is a natural extension of a matrix so that its element is indexed by more than two integers. While we used the terms “row” and “column” to refer to elements of a matrix, we use the term “mode” to refer to those of a tensor (a vector and a matrix has one and two modes, respectively). A tensor that has N modes is called a N-th order tensor,

There are two important features of tensor factorization, that is, the linearity on the number of parameters and the symmetry among the mode. The linearity on the number of parameters means that the number of parameters to be estimated by a tensor factorization algorithm increases linearly with each mode dimension. The symmetry among the modes means that a result of tensor factorization doesn’t change even if a tensor is *rolled*, that is, a tensor factorization algorithm doesn’t depend on an order of modes. There are at least two tensor factorization algorithms that meet the two properties stated above. The one is canonical polydic decomposition (CPD) and the other is Tucker decomposition, which are reviewed in the following sections.

Notations

In this section notations and basic operations for tensor is introduced following [52].

A scalar, a vector, a matrix, and a tensor are denoted by a standard Italic letter, a boldface Italic lowercase letter, a standard Roman capital letter, and a Euler script letter, respectively, as a, $\boldsymbol{a}$, A, and $\mathcal{A}$.

The i-th entry of a vector $\boldsymbol{a}$ is denoted by $\boldsymbol{a}(i)$. An element of a matrix that is specified by the i-th row and the j-th column is denoted by $\mathrm{A}(i, j)$. A tensor element that is specified by index of each mode $n_1, n_2, \ldots, n_N$ is denoted as $\mathcal{A}(n_1, n_2, \ldots, n_N)$.These notations are summarized in Table 2.1.

2.2.3 Tensor Algebra

Vectorization

The vectorization of a I–by–J matrix A is denoted by vec(A) and creates a vector, which has the IJ elements by stacking all the columns of A.

Table 2.1: Notations of a scalar, a vector, a matrix, a tensor, and their elements.

	Font	Example	Element Example
Scalar	standard Italic	a	
Vector	boldface lowercase	$\boldsymbol{a}$	$\boldsymbol{a}(i)$
Matrix	standard Roman capital	A	$\mathrm{A}(i,j)$
Tensor	Euler's script	$\mathcal{A}$	$\mathcal{A}(i_1,i_2,\ldots,i_N)$

Mode-n fibers

The mode-n fibers are vectors that are obtained by varying the index of the n-th mode while fixing all the other mode indexes. For example, we have a third order tensor $\mathcal{A} \in \mathbb{R}^{4\times 3\times 2}$. Then, we can obtain the six mode-1 fibers with the four elements, the eight mode-2 fibers with the three elements, and the twelve mode-3 fibers with the two elements. A schematic image of the mode-n fibers is shown in Figure 2.4.

Mode-n matricization (unfolding)

The mode-n matricization (also called the mode-n unfolding) of a tensor $\mathcal{A} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ is an operation that flattening a tensor into a matrix, and denoted as:

$$\mathcal{A}_{[n]} \in \mathbb{R}^{I_n\times I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N}. \tag{2.3}$$

The mode-n matricization is obtained by arranging all the mode-n fibers of input tensor into the columns of output matrix. Employing the same example above, the mode-1, 2, and 3 matricizations are the 4–by–6, 3–by–8, and 2–by–12 matrices, respectively. In spite of the simple concept, the order of the fiber arrangement into the columns is not unique, which forces us to use the following complicated notation in accordance with [52] to eliminate the ambiguity, where a tensor element $(i_1, i_2, \ldots, i_N)$ is mapped to a matrix element (i_n, j) as follows:

$$\begin{aligned} j &= 1 + \sum_{\substack{k=1\\k\neq n}}^{N} (i_k - 1) J_k, \\ J_k &= \prod_{\substack{m=1\\m\neq n}}^{k-1} I_m. \end{aligned} \tag{2.4}$$

A schematic image of the mode-n matricization is also depicted in Figure 2.4.

Rank-1 tensor and outer product

Outer product of vectors is denoted by the symbol $\circ$, and yields a tensor by calculating scalar products of all possible element-pairs among the vectors as follows:

$$\boldsymbol{a}^{(n)} \in \mathbb{R}^{I_n \times 1}, \mathcal{A} = \boldsymbol{a}^{(1)} \circ \boldsymbol{a}^{(2)} \circ \cdots \circ \boldsymbol{a}^{(N)} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}, \tag{2.5}$$

$$\mathcal{A}(i_1, i_2, \ldots, i_N) = a^{(1)}(i_1) a^{(2)}(i_2) \cdots a^{(N)}(i_N), \tag{2.6}$$

$$n = 1, 2, \ldots, N.$$

We say a N-th order tensor $\mathcal{A}$ is *rank one* if and only if it can be written as the outer product of N vectors. A schematic image of outer product and a rank-1 tensor is depicted in Figure 2.5.

Mode-n product

The mode-n product of a tensor $\mathcal{A}$ and a matrix B is denoted by $\mathcal{A} \times_n \mathrm{B}$ and yields another tensor $\mathcal{X}$ as follows:

$$\mathcal{X} = \mathcal{A} \times_n \mathrm{B}, \tag{2.7}$$

where

$$\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times J \times \cdots \times I_N},$$

$$\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}, \quad \mathrm{B} \in \mathbb{R}^{J \times I_n}, \tag{2.8}$$

and elements of the resulting tensor are calculated as:

$$\mathcal{X}(i_1, i_2, \ldots, i_{n-1}, j, i_{n+1}, \ldots, i_N) = \sum_{i_n=1}^{I_n} \mathcal{A}(i_1, i_2, \ldots, i_{n-1}, j, i_{n+1}, \ldots, i_N) \mathrm{B}(j, i_n). \tag{2.9}$$

The mode-n product is defined when the tensor dimension of the n-th mode and the number of matrix column are identical.

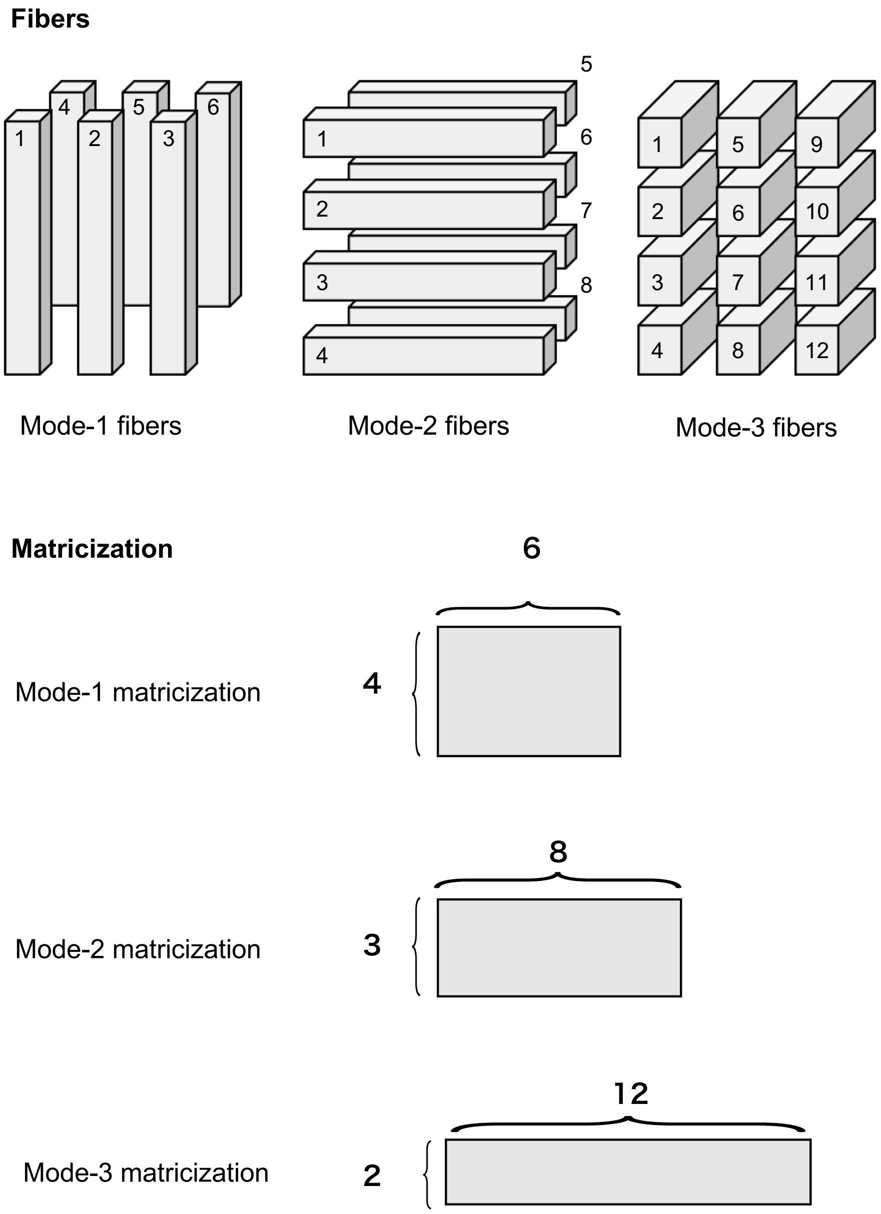


Figure 2.4: Examples of the fibers and matricizations for a third order tensor. The dimension of the first, second, and third are four, three, and two, respectively. The numbers attached to the fibers indicate the column order of the matricization.

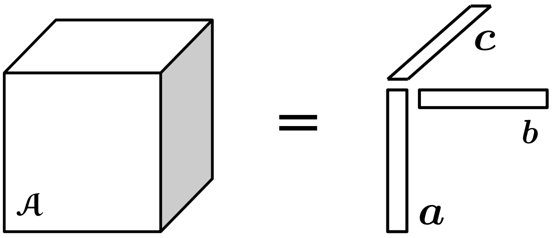


Figure 2.5: A schematic image of outer product.

Identity tensor

The identity tensor $\mathcal{I}$ is analogously defined to the identity matrix, whose all mode dimensions are identical and elements are one if their all indices are identical and otherwise zero, that is:

$$\mathcal{I} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}, \quad I_1 = I_2 = \cdots = I_N,$$
$$\mathcal{I}(i_1, i_2, \ldots, i_N) = \begin{cases} 1 & (i_1 = i_2 = \cdots = i_N) \\ 0 & (\text{otherwise}) \end{cases}. \tag{2.10}$$

Khatri-Rao product

The Khatri-Rao product [53] of two matrices with an identical number of columns is denoted by $\odot$ and yields another matrix. The n-th column of the resulting matrix are obtained by vectorizing the tensor calculated by outer product of the n-th column vectors of the matrices. It is formulated as:

$$\mathrm{A} = \left(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_K\right) \in \mathbb{R}^{I \times K}, \quad \mathrm{B} = \left(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_K\right) \in \mathbb{R}^{J \times K},$$
$$\mathrm{A} \odot \mathrm{B} = \left(\mathrm{vec}(\boldsymbol{a}_1 \circ \boldsymbol{b}_1), \mathrm{vec}(\boldsymbol{a}_2 \circ \boldsymbol{b}_2), \ldots, \mathrm{vec}(\boldsymbol{a}_K \circ \boldsymbol{b}_K)\right) \in \mathbb{R}^{IJ \times K}. \tag{2.11}$$

Frobenius norm

The Frobenius norm of a tensor $\mathcal{A}$ is denoted by $||\mathcal{A}||_{\mathrm{F}}$ and defined by the element-wise root sum square over all the indices:

$$||\mathcal{A}||_{\mathrm{F}} = \sqrt{\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_N=1}^{I_N}\mathcal{A}(i_1,i_2,\ldots,i_N)^2}, \tag{2.12}$$
$$\mathcal{A} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}.$$

2.2.4 Divergences

A divergence between two same-sized tensors is denoted by D and can be measured by the sum of the element-pair-wise divergences, that is:

$$D(\mathcal{A},\mathcal{B}) = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_N=1}^{I_N} d\left(\mathcal{A}(i_1,i_2,\ldots,i_N),\mathcal{B}(i_1,i_2,\ldots,i_N)\right), \tag{2.13}$$
$$\mathcal{A},\mathcal{B} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N},$$

where d is a function to measure divergence between two scalars and satisfies:

$$\forall a,b \in \mathbb{R} \quad d(a,b) \geq 0, \tag{2.14}$$

$$d(a,b) = 0 \quad \text{if and only if} \quad a = b. \tag{2.15}$$

Notice that a divergence function d is not necessarily symmetric:

$$d(a,b) = d(b,a), \tag{2.16}$$

and does not necessarily satisfy the triangular inequality:

$$d(a,b) \leq d(a,c) + d(c,b). \tag{2.17}$$

Here are examples of divergence functions.

Euclid distance (EUC distance)

$$d_{\mathrm{EUC}}(a,b)=\sqrt{(a-b)^2} \tag{2.18}$$

Generalized Kullback-Leibler divergence (gKL divergence)

$$d_{\mathrm{gKL}}(a,b)=\left(a\log\frac{a}{b}-a+b\right) \tag{2.19}$$

Itakura-Saito divergence (IS divergence)

$$d_{\mathrm{IS}}(a,b)=\log\frac{b}{a}+\frac{a}{b}-1 \tag{2.20}$$

Cauchy divergence

$$d_{\mathrm{Cauchy}}(a,b)=\log\frac{a}{b}+\frac{3}{2}\log\frac{2a^2+b^2}{3a^2} \tag{2.21}$$

It is worth noting that some divergences can be defined only when the both of the scalars are positive. All of the divergences above except Cauchy divergence can be derived from either of Alpha divergence, Beta divergence, and Bregman divergence [54].

When a divergence between two same-sized positive definite Hermitian matrices is measured, the generalized Itakura-Saito divergence (gIS divergence, also known as Log-determinant divergence or Stein's loss) [55–57], denoted by gIS, can be used. Let both of the matrices A and B be J–by–J positive definite Hermitian matrices. The gIS divergence is formulated as:

Generalized Itakura-Saito divergence (gIS divergence)

$$D_{gIS}\left(\mathrm{A},\mathrm{B}\right)=\mathrm{Tr}\left(\mathrm{AB}^{-1}\right)-\log\left(\det\left(\mathrm{AB}^{-1}\right)\right)-J, \tag{2.22}$$

where Tr and det are the trace and the determinant of matrix, respectively.

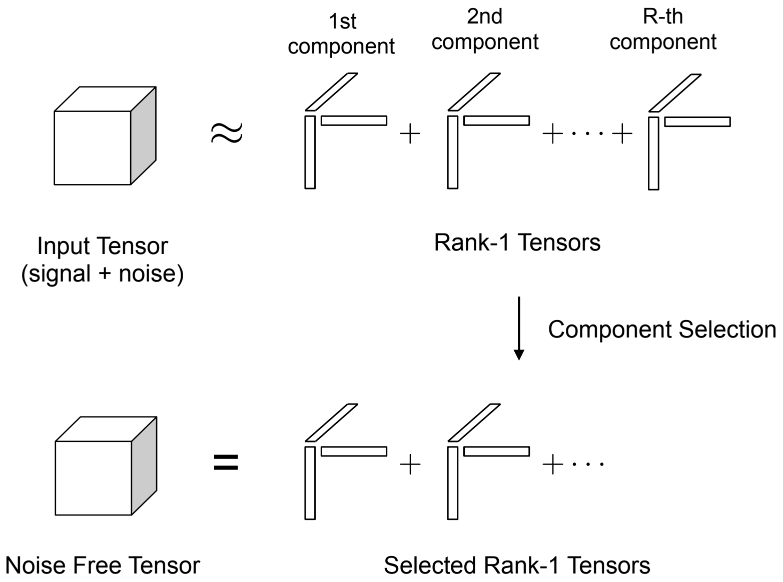


Figure 2.6: A schematic image of noise-removal based on rank-R CPD

2.2.5 Tensor Factorization Model

Canonical polyadic decomposition (CPD)

The CPD is also known as the canonical decomposition (CANDECOMP) or the parallel factor analysys (PARAFAC) [58–60]. It is formulated as a tensor decomposition into a sum of rank-1 tensors:

$$\hat{\mathcal{X}} = \sum_{r=1}^{R} \boldsymbol{a}_r^{(1)} \circ \boldsymbol{a}_r^{(2)} \circ \cdots \circ \boldsymbol{a}_r^{(I_N)} = \mathcal{I} \times_1 \mathrm{A}_1 \times_2 \mathrm{A}_2 \times_3 \cdots \times_N \mathrm{A}_N,$$
$$\hat{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N},$$
$$\mathrm{A}_n = \left(\boldsymbol{a}_1^{(n)}, \boldsymbol{a}_2^{(n)}, \ldots, \boldsymbol{a}_R^{(n)}\right) \in \mathbb{R}^{I_n \times R},$$

where $\mathcal{I}$ is the identity tensor. Each rank-1 tensor $\boldsymbol{a}_r^{(1)} \circ \boldsymbol{a}_r^{(2)} \circ \cdots \circ \boldsymbol{a}_r^{(I_N)}$ and R are called the r-th component of the decomposed tensor and the CPD rank, respectively. In addition, a vector $\boldsymbol{a}_r^{(n)}$ is called the mode-n basis of the r-th component. Because the above decomposition decomposes a tensor into R rank-1 tensors, it is called rank-R CPD. A schematic image of the noise-removal using CPD is depicted in Figure 2.6.

CPD finds factor matrices (or equivalently bases) by minimizing a divergence between its input and output tensor denoted by $\mathcal{X}$ and $\hat{\mathcal{X}}$, respectively. Using the Frobenius norm to measure the divergence, the objective function f_1 to minimize is formulated as follows:

$$f_1(\mathcal{X}, \hat{\mathcal{X}}) = \|\mathcal{X}, \hat{\mathcal{X}}\|_{\mathrm{F}}. \tag{2.23}$$

The function f_1 is not convex but the optimal solution is unique under mild conditions up to the ambiguities for scaling among modes and for permutation among components [61] and its local convergence property is shown in [62]. After the initialization of each factor matrix, an alternating gradient descent algorithm can be used for solving the minimization problem, which continues to update one factor matrix while fixing the others until convergence as follows:

$$\mathrm{A}_n \leftarrow \mathrm{A}_n - \eta \nabla_{\mathrm{A}_n} f_1, \tag{2.24}$$

where η is a learning rate, and

$$\nabla_{\mathrm{A}_n} f_1 = \frac{\partial f_1}{\partial \mathrm{A}_n} = (\hat{\mathcal{X}}_{[n]} - \mathcal{X}_{[n]})\mathrm{B}_n, \tag{2.25}$$

$$\mathrm{B}_n = \mathrm{A}_N \odot \cdots \odot \mathrm{A}_{n+1} \odot \mathrm{A}_{n-1} \odot \cdots \odot \mathrm{A}_1. \tag{2.26}$$

The n-mode matricization of the approximated tensor can be calculated using matrix product as follows:

$$\hat{\mathcal{X}}_{[n]} = \mathrm{A}_n \mathrm{B}_n^\top. \tag{2.27}$$

The overall procedure is shown in Algorithm 1.

Initialization of CPD

The initialization of factor matrices are usually done using random numbers or higher order singular value decomposition (HOSVD) [63]. When HOSVD is used for initialization, the matricization of each mode is caluculated at first. Then, standard SVD is applied for each unfolded matrix, and each factor matrix is initialized with the left singular vectors corresponding to the R largest singular values.

Algorithm 1 CPD based on the Frobenius norm

```
Require:
    $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$                ▷ Input tensor
    $R > 1$ is an integer                                     ▷ Number of components
    $\eta > 0$                                                ▷ Learning rate
    $\epsilon > 0$                                            ▷ Convergence tolerance
Ensure:
    $\hat{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$
function CPD($\mathcal{X}$, $\eta$)
    for all $n$ do
        initialize $\mathrm{A}^{(n)} \in \mathbb{R}^{I_n \times R}$
        $\mathcal{X}_{[n]} \leftarrow \mathrm{unfold}(\mathcal{X}, n)$       ▷ Unfold input tensor along with each mode
    end for
    while $D_{\mathrm{EUC}}(\mathcal{X}, \hat{\mathcal{X}}) > \epsilon$ do
        for all $n$ do
            $\mathrm{B}_n \leftarrow \mathrm{A}_N \odot \cdots \odot \mathrm{A}_{n+1} \odot \mathrm{A}_{n-1} \odot \cdots \odot \mathrm{A}_1$
            $\hat{\mathcal{X}}_{[n]} \leftarrow \mathrm{A}_n \mathrm{B}_n^\top$
            $\mathrm{A}_n \leftarrow (\hat{\mathcal{X}}_{[n]} - \mathcal{X}_{[n]}) \mathrm{B}_n$
        end for
    end while                                                 ▷ End iterative update
    return $\hat{\mathcal{X}}$
end function
```

Tucker Decomposition

Tucker decomposition is a generalization of CPD and formulated as follows:

$$\hat{\mathcal{X}} = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \cdots \sum_{k_N=1}^{K_N} \mathcal{G}(n_1, n_2, \ldots, n_N) \boldsymbol{a}_{k_1}^{(1)} \circ \boldsymbol{a}_{k_2}^{(2)} \circ \cdots \circ \boldsymbol{a}_{k_N}^{(I_N)}, \tag{2.28}$$

$$\hat{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}, \quad \mathcal{G} \in \mathbb{R}^{K_1 \times K_2 \times \cdots \times K_N},$$

$$\mathrm{A}_n = \left(\boldsymbol{a}_{k_1}^{(n)} \boldsymbol{a}_{k_2}^{(n)} \ldots \boldsymbol{a}_{K_N}^{(n)} \right) \in \mathbb{R}^{I_n \times K_N},$$

where $\mathcal{G}$ is called the core tensor. While CPD restricts the ranks of all components to R, Tucker decomposition allows them to differ. If the core tensor is diagonal, it reduces to CPD.

Chapter 3

Graph Regularized Tensor Factorization

3.1 Introduction

ERPs are usually analyzed using trial averaging to weaken the background potentials, that is always generated and irrelevant to stimuli presented to a subject. Noise-removal problems to realize single-trial ERP analysis have gathered much attention from researchers because such an averaging procedure conceals variabilities among trials.

Matrix factorization algorithms including independent component analysis (ICA) [37, 64] have been studied to tackle this problem. However, such algorithms can only be applied to a two-way array, i.e., a matrix, while ERP data usually have more than two modes (dimensions), for example time, electrodes, frequencies, trials, subjects, and experimental conditions, which naturally expresses such data as a multi-dimensional array, or a tensor. Therefore, the application of tensor factorization to ERP data has been studied recently [54, 65]. Especially, CPD has been commonly used for its high interpretability and simplicity [66–69].

However, it is well known that EEG data are noisy and sensitive to the outliers caused by body movements such as eye blinks. The amount of data is also usually small, which causes the above statical noise-removal methods to slip into over-fitting and fail to find meaningful components.

Regularization can be used to avoid over-fitting by adding the norm of parameters or incorporating prior knowledge of the data [70]. For example, researchers in the field of brain-machine interfaces have found that a common spatial pattern filter, which is widely used in two-class classification problems with EEG data, can be effectively constructed with regularization based on the geometric structure of the EEG electrode location [71, 72]. An EEG is a summation of an electrical recording of numerous neuron activities from which we can reasonably assume that multiple electrodes will observe signals that resemble each other if they are closely located on a scalp. However, existing CPD applications to EEG data have failed to utilize the spatial smoothness of EEGs and often identify spatially bumpy components [66, 69]. On the other hand, the nonnegative matrix factorization (NMF) algorithm that incorporates a graphical structure (GNMF) was proposed and applied to the facial image analysis [73].

In this chapter, a new CPD-based noise-removal method is proposed, which incorporates geometrical information of the EEG electrodes. It is a natural extension of GNMF to tensor factorization.

3.2 Graph Regularized Tensor Factorization

In this section, CPD is extended to GCPD, which uses the geometrical information of the electrode location for both regularization and initialization. First, we explain a method that models the geometrical information of the electrode location using a graph. Second, we introduce a regularization method for CPD that constrains the spatial smoothness of the components. Third, we explain how to use the graph structure for initialization.

3.2.1 Obtaining Adjacency Matrix

EEG electrodes are usually placed on a scalp according to the International 10-20, 10-10, or 10-5 systems [74–76] depicted in Figure 3.1. These systems assume that the electrodes are placed on a sphere and define the j-th electrode position using a three-dimensional vector as $\boldsymbol{z}_j = (x_j, y_j, z_j)$ by setting the sphere's center as the origin.

The local structure among J nodes can be expressed by an undirected graph which has J nodes and its corresponding J–by–J adjacency matrix [77]. In our study, we regards EEG electrodes as nodes. We denote an adjacency matrix by W, and its element $\mathrm{W}(i,j)$ express the similarity between the i-th and j-th nodes.

The commonly used methods to measure the similarity between the i-th and j-th nodes are inner product:

$$\mathrm{W}(i,j) = \boldsymbol{z}_i^\top \boldsymbol{z}_j, \tag{3.1}$$

and heat kernel weighting [78]:

$$\mathrm{W}(i,j) = \exp \frac{-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}{\sigma}. \tag{3.2}$$

With these measures employed, the adjacency matrix is symmetric and its element $\mathrm{W}(i,j)$ increases when the i-th and j-th nodes are similar and vice versa.

We have mentioned only the spatial structure of EEG but the temporal smoothness can be also considered in the same way by using each time index as a node to define a graph structure.

3.2.2 Graph Regularization

Given an adjacency matrix, the smoothness of the CPD bases of the n-th mode on its corresponding graph can be measured using the following term:

$$S_n\left(\mathrm{A}_n, \mathrm{W}^{(n)}\right) = \sum_{k=1}^{K} \sum_{i,j=1}^{I_N} ||\mathrm{A}_n(i,k) - \mathrm{A}_n(j,k)||^2 \mathrm{W}^{(n)}(i,j) = \mathrm{Tr}(\mathrm{A}_n^\top \mathrm{L}_n \mathrm{A}_n), \tag{3.3}$$

$$\mathrm{L}_n = \mathrm{D}^{(n)} - \mathrm{W}^{(n)}, \tag{3.4}$$

the matrix $D^{(n)}$ is a diagonal matrix, whose diagonal elements are a column (or equivalently a row) sum of $\mathrm{W}^{(n)}$, and L_n is called a graph Laplacian matrix [77]. In the spectral graph theory literature, the term $\mathrm{Tr}(\mathrm{A}_n^\top \mathrm{L}_n \mathrm{A}_n)$ is called the graph Laplacian quadratic form. It increases when the difference between the column entries of the factor matrix $\mathrm{A}_n(i,k)$ and $\mathrm{A}_n(j,k)$ is big although $\mathrm{W}^{(n)}(i,j)$ is big, in other words

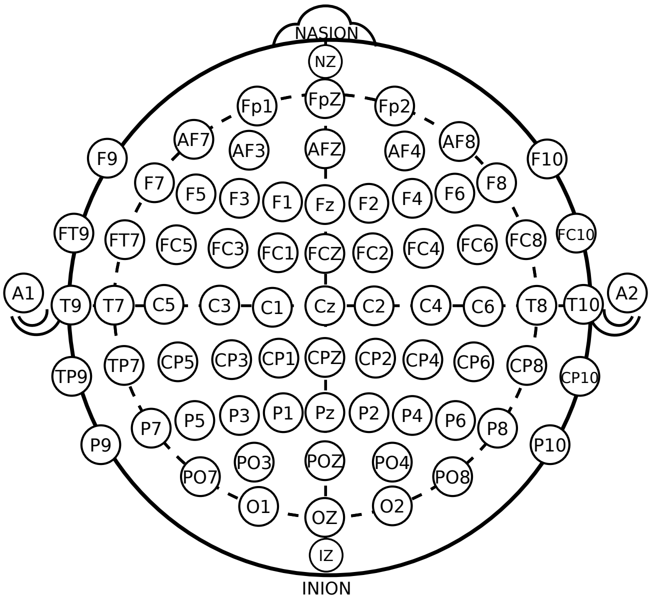


Figure 3.1: Modified international 10-20 system. This figure is distributed under the Creative Commons CC0 1.0 Universal Public Domain Dedication [79].

the i-th node and the j-th node are close to each other on the graph [78, 80].

By adding this term to the original objective function of CPD shown in Equation (2.23), the objective function of GCPD is formulated as follows:

$$f_2(\mathcal{X},\hat{\mathcal{X}}) = \|\mathcal{X}-\hat{\mathcal{X}}\|_{\mathrm{F}} + \sum_{n=1}^{N} \left(\lambda_n \|\mathrm{A}_n\|_{\mathrm{F}} + \psi_n S_n\left(\mathrm{A}_n, \mathrm{W}^{(n)}\right)\right), \tag{3.5}$$

where ψ_n is a graph regularization parameter. By minimizing this function we expect that bases can be found that respect the intrinsic graphical structure. A gradient of f_2 is given by the following terms:

$$\nabla_{\mathrm{A}_n} f_2 = (\hat{\mathrm{X}}_n - \mathrm{X}_n)\mathrm{B}_n + (\lambda_n I + \psi_n \mathrm{L}_n)\mathrm{A}_n, \tag{3.6}$$

$$\mathrm{B}_n = \mathrm{A}_N \odot \cdots \odot \mathrm{A}_{n+1} \odot \mathrm{A}_{n-1} \odot \cdots \odot \mathrm{A}_1, \tag{3.7}$$

where I is the identity matrix. Iterative updating is performed as in Equation (2.24).

3.3 Initialization Considering Graph Structure

While GNMF uses the graphical smoothness only for regularization, we propose to incorporate it not only for regularization but also initialization of CPD. The bases of each mode found by GCPD should respect its geometric structure: in other words, such bases make smoothness defined by Equation (3.3) small. Therefore, the proposed initialization method is given by solving the following minimization problem [78]:

$$\mathrm{A}_n^* = \arg\min_{\mathrm{A}_n} S(\mathrm{A}_n, \mathrm{W}_n). \tag{3.8}$$

The solution of this problem is given by the following eigenvalue problem:

$$\mathrm{L}_n \boldsymbol{b}^{(n)} = p\boldsymbol{b}^{(n)}. \tag{3.9}$$

Let the dimension of the n-th mode J_n. Since the graph Laplacian is a J_n–by–J_n real symmetric matrix, it has a complete set of the J_n eigenvectors $\{\boldsymbol{b}_k^{(n)}\}_{k=0,\dots,J_n-1}$ associated with eigenvalues $\{p_k^{(n)}\}_{k=0,\dots,J_n-1}$ that satisfy

$$0 = p_0 < p_1 \leq \cdots \leq p_{J_n-1}. \tag{3.10}$$

The eigenvectors of the graph Laplacian is called *the graph Fourier bases* (GFB). GFB corresponding to a smaller eigenvalue reduce the following term more greatly:

$$S'(\boldsymbol{b}_k^{(n)}, \mathrm{W}_n) = \sum_{i,j=1}^{J_N} ||\boldsymbol{b}_k^{(n)}(i) - \boldsymbol{b}_k^{(n)}(j)||^2 \mathrm{W}_n(i,j), \tag{3.11}$$

where $\boldsymbol{b}_k^{(n)}(j)$ is the j-th element of the vector $\boldsymbol{b}_k^{(n)}$. In other words, the following relation holds [80]:

$$0 = S'(\boldsymbol{b}_0^{(n)}, \mathrm{W}_n) < \cdots \leq S'(\boldsymbol{b}_{J_{N-1}}^{(n)}, \mathrm{W}_n). \tag{3.12}$$

Each factor matrix is initialized with the K GFBs corresponding to the K smallest eigenvalues.

While HOSVD initializes the factor matrices based on the orthogonality and the variance explained by the direction of the singular vectors and fails to consider the graph structures, the eigenvectors of the graph Laplacian are also orthogonal as well as they consider the smoothness on a given graph. Moreover, they are dependent only on a graph Laplacian matrix. Therefore, they always produce the same initialization result with the same Laplacian matrix (the same EEG electrode location) even if a subject or an experimental condition is different: HOSVD is dependent on input data, which means the quality of initialization will be largely affected by the data amount and the artifacts in the data. This can be an advantage for the proposed method because EEGs are easily affected by such artifacts as body movements and the amount of data is often severely limited.

3.4 Experimental Evaluation

An noise removal experiment is described in this section. Since it is difficult to discriminate which decomposed components really correspond to a *signal* (an EEG signal caused by brain activities of interest) or a *noise* (a signal caused by other brain activities or artifacts), a pseudo ERP data were created using real EEG signals to make it possible to evaluate the proposed method and the standard CPD objectively.

3.4.1 Data Acquisition

All EEG signals were recorded in a sound proof room with 25 scalp electrodes placed according to International 10-20 system at the sampling rate of 1000 Hz. Three healthy subjects aged from 23 to 26 years old without any neurological disorders participated in the experiment. All experimental procedures were approved by the Ethical Review Board of Nara Institute of Science and Technology. The recorded EEG signals were down-sampled to 200 Hz, and a band-pass filter was applied between 0.01 Hz and 30 Hz.

An auditory oddball paradigm was used [33] to elicit the ERP component of P300 from the subjects. A random sequence of 2000 Hz and 1000 Hz sound stimuli was presented to each subject by earphones. All the sound stimuli had a duration of 200 milliseconds and the intervals between them were 1.4 seconds. 2000 and

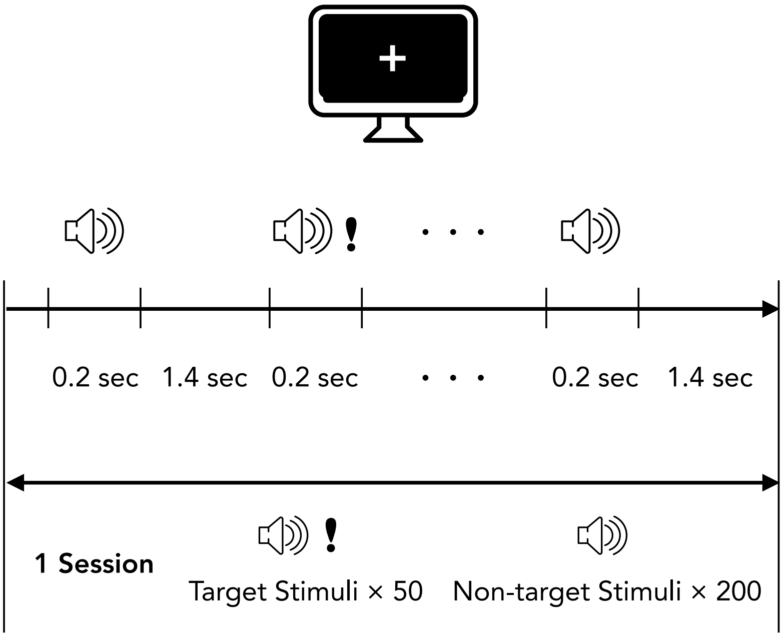


Figure 3.2: Timeline of the EEG recording experiment. All stimuli have the duration of 0.2 seconds and the intervals between stimuli were 1.4 seconds. Target and Non-target stimuli were presented to subjects 50 and 200 times in one session, respectively. Subjects looked at the fixation mark on the monitor and counted the number of target stimuli presences.

1000 Hz sounds were respectively presented 50 and 200 times. Subjects counted the number of 2000 Hz sound stimuli (target trials) while ignoring 1000 Hz sound (non-target trials). It has been well documented that the ERP of P300 appears more conspicuously when a subject is presented a target stimuli than a non-target one. After repeating this procedure twice (100 target and 400 non-target trials in total), the subjects were told to relax without a task or a stimulus to record their resting state EEG for two minutes.

3.4.2 Validation Using Pseudo ERP-Data

From the resting state EEG signal from the each subject, 100 epochs of 850 milli seconds duration (170 time samples) was extracted randomly. All the target-trials obtained from the oddball paradigm were averaged across trials. Then, the resulting averaged ERP that stands for *signal* was added to the each of 100 resting state epochs that stands for *noise* to make the 100 trials of pseudo ERP data. The obtained ERPs are shown in Figure 3.3. The phase of the added ERP was randomly shifted from -50 ms to 50 ms and amplitude also randomly changed by multiplying a random number sampled from the normal distribution with mean 1 and standard deviation 0.4. The averaged non-target trial ERP was also added to another 100 resting state epochs as same as the target-trial ERP. Concatenating the two data, a third order tensor $\mathcal{X} \in \mathbb{R}^{170\times25\times200}$ with modes of $time \times space(electrodes) \times trials$ was made for each subject.

The effectiveness of compared methods were measured by how well they removed the noise and extracted the added ERP components from the tensor.

The rank of CPD is set to 20 for both of the methods. Similarities between electrodes and time samples were defined by the heat kernel weighting with the variance parameter $\sigma = 1$. The hyper parameters were chosen following the previous researches. Spatial Wiener filtering introduced in Chapter 4 was also compared.

The noise removal performances were measured by root mean square error (RMSE):

3.4.3 Component Selection

As with other source separation techniques, CPD decomposes an input signal into multiple components without identifying the components of interest. Therefore, they must be selected based on prior knowledge [81].

Target-trials elicit larger ERP than non-target trials. If a component corresponds to an ERP, the magnitude of trial mode bases differ among the set of elements corresponding to target trials and the set of the non-target trials. On the other hand, if a component doesn't correspond to an ERP, the magnitude of the trial mode bases will take similar values among elements. Based on this idea, the inter-condition variance (ICV) of the r-th component is defined as follows:

$$\mathrm{ICV}(r) = \sum_{i\in P_T} \left|\mathrm{A}^{trial}(i,r)\right| - \sum_{i\notin P_T} \left|\mathrm{A}^{trial}(i,r)\right|, \tag{3.13}$$

where D_T is a set of component indexes that corresponds to the target trials. Components with a high ICV are chosen as ERP components.

A previous work adopted a similar approach [66] and selected components based

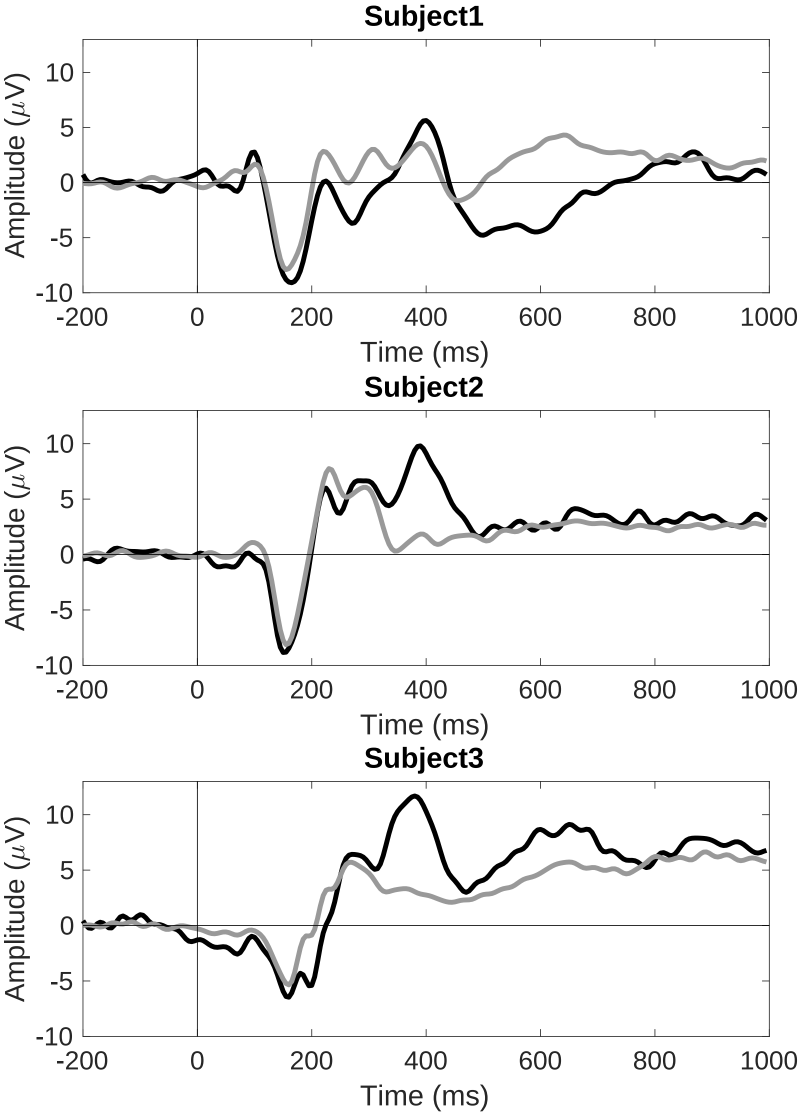


Figure 3.3: Plots of the averaged ERPs of all subjects. Black and gray lines are time courses of the averaged ERPs of target and non-target trials, respectively.

on the p-value of a statical test to the magnitude of the subject mode to investigate the difference between subjects with reading and with attention disabilities. ICV is simpler and more instinctive than the p-value of a statical significant test, which needs a sufficient amount of data to find significant differences and assumptions the about data, e.g., distribution or variance. We selected the three components that had the highest ICVs.

3.4.4 Result

The noise-removal performance of GCPD was superior to CPD for all subjects as summarized in Table 3.1, where the averaged RMSE values with the standard deviation (s.d.). The average was calculated across all the 200 trials and the 25 electrodes (5000 RMSEs in toal). The performance of the spatial Wiener filter was worse than CPD and GCPD, which is probably caused its high dimensionality of parameters to be estimated (a detailed discussion is in Section 4.8).

The bases of the component that was obtained from the subject 3 by GCPD and had the highest ICV is shown in Figure 3.3 for each mode. The temporal mode basis clearly shows the appearance of P300. The spatial mode basis shows the amplitude distribution on the scalp. It is spatially smooth and has the highest activity at near the central of the scalp, which is physiologically plausible. The absolute value of the trial mode basis is shown for both of trial and non-target trials. The mode is important for single-trial analysis because it is possible to know which trials affect to a cognitive state of a subject more strongly than others. From the first to one hundredth elements correspond to target trials and the rest to non-target trials. It can be seen that magnitude are larger at the target-trials than the non-target ones, which is why this component is selected by ICV.

Convergence results of GCPD initialized by GFB and HOSVD are shown in Figure 3.5. Although GCPD initialized by GFB achieved the better noise removal performance in all subjects, GCPD initialized by HOSVD converged faster than that by GFB in the subject 2.

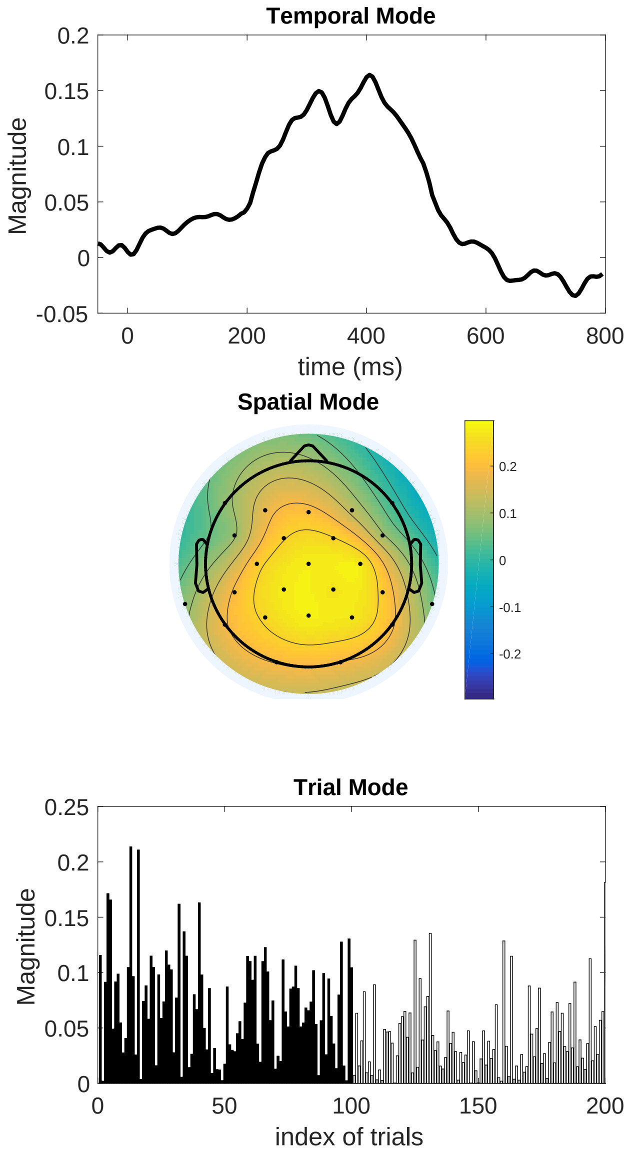


Figure 3.4: Each mode of the component that had the highest ICV. (Left) the temporal mode, (middle) the spatial mode, and (right) the absolute value of the trial mode with the highest ICV. The first one hundred bars and the rest indicate the target and non-target trials, respectively.

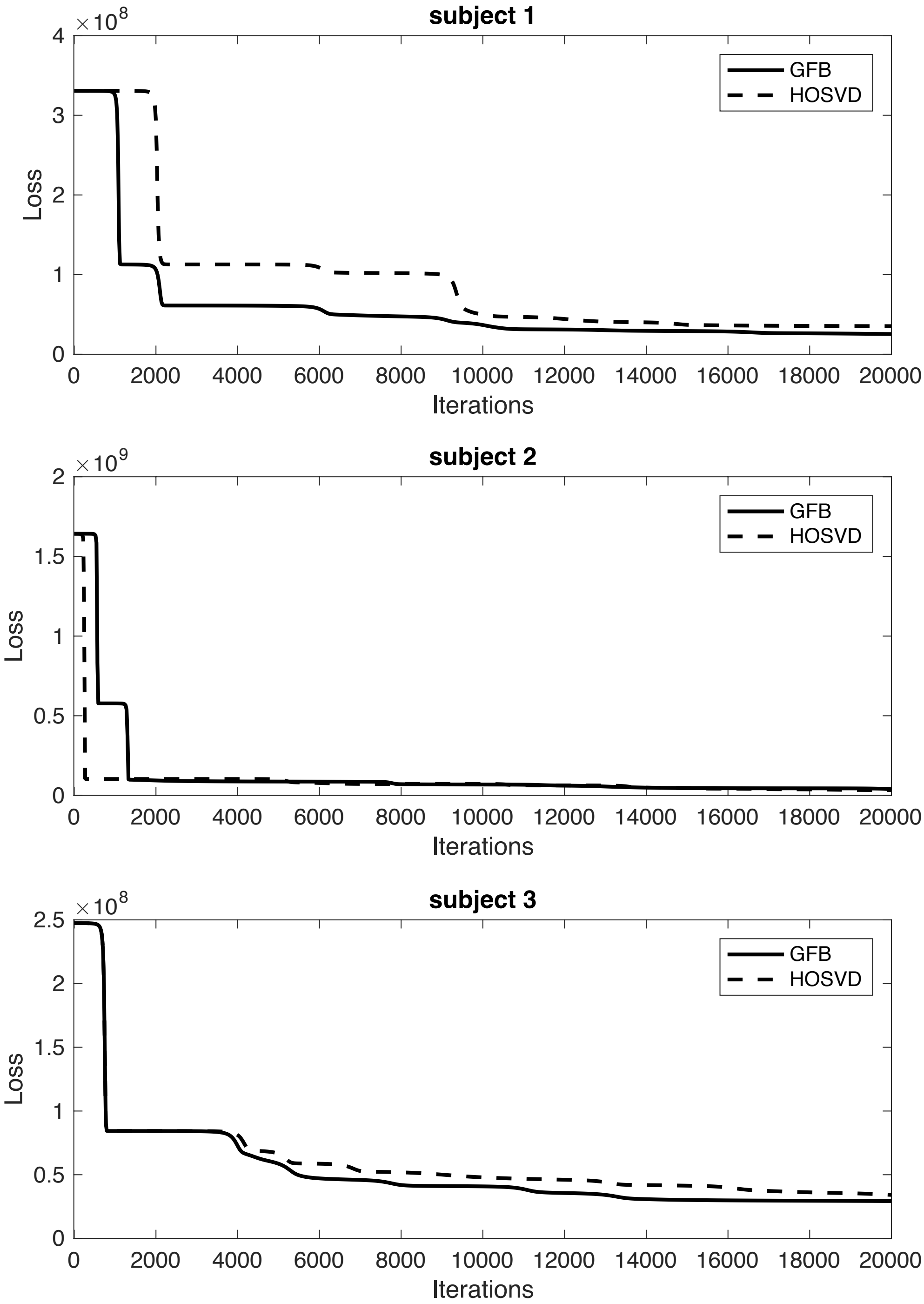


Figure 3.5: Convergence results of GCPD initialized by GFB (solid line) and HOSVD (dotted line). Horizontal axis indicates the number of iterations of the gradient descent algorithm, and vertical axis does the value of the loss function.

Table 3.1: RMSE of all subjects

	RMSE		
	Subject 1	Subject 2	Subject 3
CPD	3.96 (s.d. 3.15)	7.12 (s.d. 6.29)	6.13 (s.d. 4.57)
GCPD + HOSVD	3.41 (s.d. 3.03)	7.12 (s.d. 6.29)	6.39 (s.d. 6.50)
GCPD + GFB	2.87 (s.d. 2.05)	6.88 (s.d. 4.60)	4.05 (s.d. 4.36)
Spatial Wiener filter	9.22 (s.d. 18.63)	11.27 (s.d. 14.64)	11.42 (s.d. 14.46)
Conventional averaging	2.47 (s.d. 1.18)	3.14 (s.d. 2.02)	2.01 (s.d. 1.04)

3.5 Summary

This Chapter proposed GCPD, a new tensor factorization method that extends the standard CPD so that it incorporates the geometrical structure of the EEG electrode location.

An adjacency matrix for each mode encodes the geometric structure of the mode whose elements represents the distances between entries of the mode bases, and was used for for two purposes, regularization and initialization.

The regularization term is defined using the graph Laplacian matrix that represents the smoothness on its graph. The initialization was performed using the graph Fourier bases, or eigen vectors of the graph Laplacian matrix. They are orthogonal to each other and consider the graph structure.

The noise-removal experiment from ERP data demonstrated that GCPD removed background EEGs from single-trial ERPs more effectively than the conventional CPD, especially with initialization by graph Fourier bases although the performance of GCPD was not compatible with the conventional averaging.

Chapter 4

Noise-Removal via Spatial Covariance Modeling

4.1 Introduction

In this chapter, a noise-removal problem from EEGs is tackled using a different approach from Chapter 3. The goal of this chapter is to construct an adaptive filter that separates an observed EEG signal into multiple components. An probabilistic generative model of EEG with multiple modes is defined, and an adaptive filter is created to maximize posterior probability of an observed signal. Because not only filter coefficients but also model parameters have to be estimated from observed data, they are iteratively updated. To create such a filter, spatial covariance matrices (covariance matrix between data captured by multiple electrodes) play a key role. While GCPD introduced in Chapter 3 performs signal separation without specific knowledge of components, prior distribution of spatial covariance matrices are considered to improve the performance of parameters estimation. It is also shown that tensor factorization and maximum a posteriori approaches can be comprehensively understood from the point of divergence minimization.

This chapter is constructed as follows. First, related researches are surveyed. Second, our framework for signal separation employed in this chapter is explained. Third, a probabilistic generative model of EEG sources and maximum likelihood estimation of the model parameters are introduced. Next, a method to incorporate prior knowledge of objective components is presented. Finally, a theoretical connection between the tensor factorization and the adaptive filtering approaches that are presented in Chapter 3 and 4 is discussed.

4.2 Related Works

In most cases, an EEG recording experiment is conducted using more than one electrode, which allows an observed signal to have spatial characteristics. In the literature of the multi-channel sound signal separation research, several methods have been investigated that model the spatial spread of an observed signal using a covariance matrix between microphones [57, 82–84]. In these methods, coefficients of short-time Fourier transform (STFT) of observed signal is assumed to follow a phase-invariant multivariate complex Gaussian distribution with zero mean vector as follows:

$$\begin{aligned} \boldsymbol{x}(n,f) &\sim \mathcal{N}_c\left(\boldsymbol{x}(n,f) \mid \mathbf{0}, \mathrm{R}_{n,f}\right) \\ &= \frac{1}{\pi^J \det\left(\mathrm{R}_{n,f}\right)} \exp\left(-\boldsymbol{x}^{\mathrm{H}}(n,f)\mathrm{R}_{n,f}^{-1}\boldsymbol{x}(n,f)\right), \end{aligned} \tag{4.1}$$

where $\boldsymbol{x}(n,f) \in \mathbb{C}^J$ is the complex valued activity of the observed signal captured by J microphones, n and f are indexes of the time frame and frequency bin, respectively.

Our method is strongly inspired by these methods. While multi-channel sound is modeled in the time-frequency domain, EEG signal often have additional modes, for example, trial, experimental condition, subject, and etc. Moreover, the sources of the EEG signal are numerous neurons, which makes the problem under-determined signal separation explained in Section 2.2.1. Therefore, we created a model that can handle an observed data that has more than two modes and an under-determined signal separation, which is explained in the following sections.

4.3 Signal Separation Framework

4.3.1 Observational Model

Let us assume EEG signals are observed using J electrodes in the the time-frequency domain and can have additional modes. In general, we assume the observed signal has N modes and denote the complex-valued observational activity by $\boldsymbol{x}(n_1, n_2, \dots, n_N)$, where n_m is the index of the m-th mode. We call the index set $(n_1, n_2, \dots, n_N)$ *multi-domain frame*. To avoid a clunky notation, we use a single index i that corresponds to a unique multi-domain frame as:

$$\boldsymbol{x}(i) := \boldsymbol{x}(n_1, n_2, \dots, n_N), \tag{4.2}$$

$$i = 1, 2, \dots, I, \quad n_m = 1, 2, \dots, I_m, \tag{4.3}$$

$$I = \prod_{m=1}^{N} I_m. \tag{4.4}$$

Observed signal is assumed to be a sum of K components (ERPs, background EEGs, eye blinks, and others) as:

$$\boldsymbol{x}(i \mid \theta(i)) = \mathbf{c}_1(i \mid \theta_1(i)) + \mathbf{c}_2(i \mid \theta_2(i)) + \cdots + \mathbf{c}_K(i \mid \theta_K(i)), \tag{4.5}$$

$$\theta(i) = \{\theta_1(i), \theta_1(i), \dots, \theta_K(i)\}, \tag{4.6}$$

where $\mathbf{c}_k(i)$ is the k-th component and $\theta_k(i)$ is the parameter set of the probabilistic generative model that generates $\mathbf{c}_k(i)$.

4.3.2 Adaptive Filtering

The log-likelihood of the observed signal $L_1(\boldsymbol{x}, \theta)$ is written as:

$$L_1(\boldsymbol{x}, \theta) = \log p(\boldsymbol{x} \mid \theta) = \log \sum_{i=1}^{I} p(\boldsymbol{x}(i) \mid \theta(i)), \tag{4.7}$$

$$\boldsymbol{x} = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \dots, \boldsymbol{x}(I)\},$$

$$\theta = \{\theta(1), \theta(2), \dots, \theta(I)\},$$

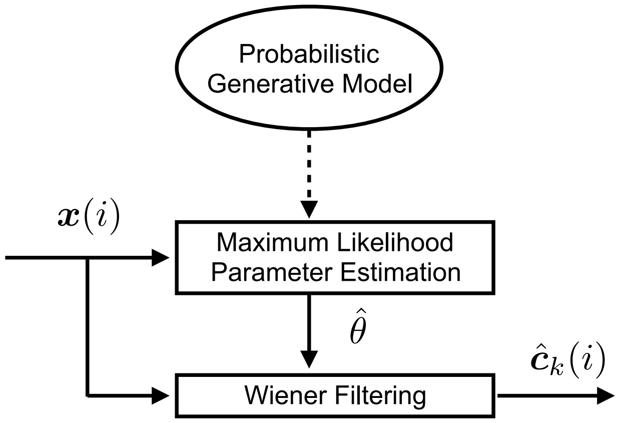

Probabilistic Generative Model
$\boldsymbol{x}(i)$
Maximum Likelihood Parameter Estimation
$\hat{\theta}$
$\hat{\boldsymbol{c}}_k(i)$
Wiener Filtering

Figure 4.1: Overview of the Wiener filtering based on the probabilistic generative model.

and the maximum likelihood estimator of the parameter set $\hat{\theta}$ is:

$$\hat{\theta} = \arg\max_{\theta} L_1(\boldsymbol{x}, \theta). \tag{4.8}$$

An adaptive filter using the optimal parameter set $\mathrm{M}(\hat{\theta})$ is constructed and an estimator of the k-th component $\hat{\boldsymbol{c}}_k(i)$ in the multi-domain frame (i) is obtained as:

$$\hat{\boldsymbol{c}}_k(i) = \mathrm{M}(\hat{\theta}_k(i))\boldsymbol{x}(i). \tag{4.9}$$

Notice that this is a time- and trial- variant filtering in contrast that ICA shown in Equation (2.2) is a time- and trial- invariant filtering. If the filter is estimated by minimizing mean square error assuming all components are independent from each other, it is called *a multi-channel Wiener filter* and obtained as:

$$\hat{\mathrm{M}}_k(i) = \mathrm{R}_{\boldsymbol{c}_k(i)} \mathrm{R}^{-1}_{\boldsymbol{x}(i)}, \tag{4.10}$$

where

$$\mathbb{E}_i\left[\boldsymbol{x}(i)\boldsymbol{x}(i)^{\mathrm{H}}\right] = \mathrm{R}_{\boldsymbol{x}(i)}, \quad \mathbb{E}_i\left[\boldsymbol{c}_k(i)\boldsymbol{c}_k(i)^{\mathrm{H}}\right] = \mathrm{R}_{\boldsymbol{c}_k(i)}. \tag{4.11}$$

The schematic image of the overall procedure of the component extraction is shown in Figure 4.3.2.

4.4 Probabilistic Generative Model

In this section, we explain a source generative model at first. Second, a component generative model is explained. Third, an observation generative model is derived and an observational likelihood to maximize is formulated. Generative models of a source, a component, and an observation are linked to each other due to the linearity of the Gaussian distribution.

4.4.1 Source Generative Model

Each component is assumed to be generated from L sources whose contributions are instaneously mixed as:

$$\begin{gathered}\boldsymbol{c}_k(i) = \mathrm{B}_k \boldsymbol{s}_k(i) \in \mathbb{C}^J, \\ \boldsymbol{s}_k(i) \in \mathbb{C}^L, \\ \mathrm{B}_k = \begin{pmatrix} b_k(1,1) & \dots & b_k(1,L) \\ \vdots & \ddots & \vdots \\ b_k(J,1) & \dots & b_k(J,L) \end{pmatrix} \in \mathbb{C}^{J \times L},\end{gathered} \tag{4.12}$$

where B_k is the lead field matrix of the k-th component and $\boldsymbol{s}_k(i)$ is the source contribution to the k-th component. A schematic image of the observational model is depicted in Figure 4.4.1.

Assumption 1: Independent and Individual Gaussian Distributions

The source activation *independently* follows the multi-variate proper complex Gaussian with the zero mean vector and the *individual variances* that depend on the index of multi-domain frame (i):

$$\boldsymbol{s}_k(i) \sim \mathcal{N}_c\left(\boldsymbol{s}_k(i) \mid \mathbf{0}, \sigma_k(i)\mathrm{I}\right), \tag{4.13}$$

where I is the L–by–L identity matrix.

Assumption 2: Variance factorization

Each mode has its own variance in each multi-domain frame and the source

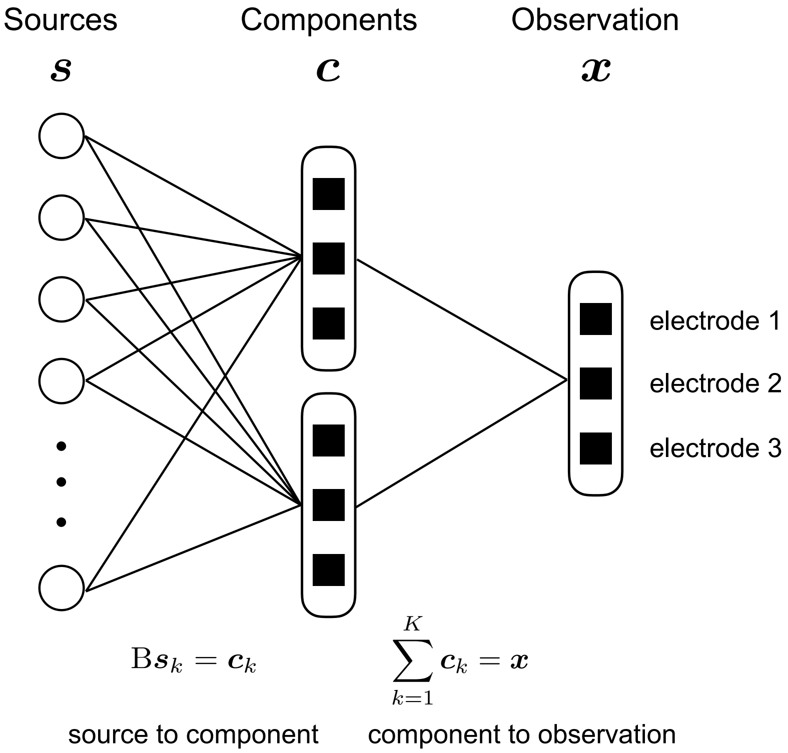


Figure 4.2: Schematic image of the observational model.

variance is the product of the mode variances. Therefore, source variances can be expressed as an outer product of vectors:

$$\begin{aligned}\sigma_k(i) &= \sigma_k(n_1, n_2, \ldots, n_N)\\ &= \mathrm{A}^{(1)}_{n_1,k}\mathrm{A}^{(2)}_{n_2,k}\ldots\mathrm{A}^{(N)}_{n_N,k}\\ &= \left(\boldsymbol{a}^{(1)}_k \circ \boldsymbol{a}^{(2)}_k \circ \cdots \circ \boldsymbol{a}^{(N)}_k\right)_{n_1,n_2,\ldots,n_N},\end{aligned} \tag{4.14}$$

where $\mathrm{A}^{(n)}$ is called *a variance factor matrix* and defined as:

$$\mathrm{A}^{(n)} = \begin{pmatrix} \mathrm{A}^{(n)}_{1,1}, & \ldots & ,\mathrm{A}^{(n)}_{1,K} \\ \vdots & \ddots & \vdots \\ \mathrm{A}^{(n)}_{I_n,1}, & \ldots & ,\mathrm{A}^{(n)}_{I_n,K} \end{pmatrix} = \left(\boldsymbol{a}^{(n)}_1, \quad \boldsymbol{a}^{(n)}_2, \quad \ldots \quad ,\boldsymbol{a}^{(n)}_K\right). \tag{4.15}$$

4.4.2 Component Generative Model

Thanks to the linearity of the Gaussian (See Appendix), component activations also follow the multi-variate proper complex Gaussian, whose covariance matrix is calculated as follows:

$$\boldsymbol{c}_k(i) \sim \mathcal{N}_c\left(\boldsymbol{c}_k(i) \mid \mathbf{0}, \mathrm{R}_{\boldsymbol{c}_k(i)}\right), \tag{4.16}$$

$$\mathrm{R}_{\boldsymbol{c}_k(i)} = \mathbb{E}_i[\boldsymbol{c}_k(i)\boldsymbol{c}_k(i)^{\mathrm{H}}] = \mathrm{B}_k\mathbb{E}_i[\boldsymbol{s}_k(i)]\mathrm{B}_k^{\mathrm{H}} = \sigma_k(i)\mathrm{R}_k, \tag{4.17}$$

where $\mathrm{R}_k = \mathrm{B}_k\mathrm{B}_k^{\mathrm{H}}$ is called *the spatial covariance matrix* of the k-th component.

4.4.3 Observation Generative Model

Using the linearity of the Gaussian, we can finally derive the probability density function distribution that the observed signal follows:

$$\boldsymbol{x}(i) \sim \mathcal{N}_c\left(\boldsymbol{x}(i) \mid \mathbf{0}, \mathrm{R}_{\boldsymbol{x}(i)}\right), \tag{4.18}$$

$$\mathrm{R}_{\boldsymbol{x}(i)} = \sum_{k=1}^{K} \mathrm{R}_{\boldsymbol{c}_k(i)}. \tag{4.19}$$

4.4.4 Sparsity Constraint

To make the maximum likelihood estimation simpler, the components are assumed to be generated sparsely in each multi-domain frame. To modify Equation (4.5) introducing the sparsity, we use the latent factors $z_k(i)$, which takes the 1-of-K representation as follows:

$$\boldsymbol{x}(i) = \sum_{k=1}^{K} z_k(i)\boldsymbol{c}_k(i), \tag{4.20}$$

$$\sum_{k=1}^{K} z_k(i) = 1, \quad z_k(i) \in \{0,1\}. \tag{4.21}$$

We call this observational model the sparse model. It restricts the number of components that activate in a multi-domain frame to one, and makes the conditional distribution of $\boldsymbol{x}(i)$ expressed as:

$$p(\boldsymbol{x}(i) \mid \sigma, \mathrm{R}, [z_k(i) = 1]) = \mathcal{N}_c(\boldsymbol{x} \mid \mathbf{0}, \sigma_k(i)\mathrm{R}_k). \tag{4.22}$$

If the probability that $z_k(i)$ will take the value 1 is π, it is written:

$$p(z_k(i) = 1) = q_k(i), \tag{4.23}$$

$$0 \leq q_k(i) \leq 1, \quad \sum_{k=1}^{K} q_k(i) = 1. \tag{4.24}$$

Denoting the latent factors by the vector form

$$\boldsymbol{z}(i) = \Big(z_1(i), z_2(i), \ldots, z_K(i)\Big)^{\top}, \tag{4.25}$$

the marginal distribution of the latent factors can be written:

$$p(\boldsymbol{z}(i)) = \prod_{k=1}^{K} (q_k(i))^{z_k}. \tag{4.26}$$

Then, the conditional distribution of the observed signal Equation (4.22) can be rewritten with the latent factors:

$$p(\boldsymbol{x}(i) \mid \theta, \boldsymbol{z}(i)) = \mathcal{N}_c(\boldsymbol{x}(i) \mid \mathbf{0}, \sigma_k(i)\mathrm{R}_k)^{z_k(i)}, \tag{4.27}$$
$$\theta = \{\sigma, \mathrm{R}, q\}.$$

The marginal distribution of $\boldsymbol{x}(i)$ is obtained by summing the joint distribution over all possible state of $\boldsymbol{x}$ to obtain the following Gaussian mixture model (GMM):

$$p(\boldsymbol{x}(i) \mid \theta) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}(i)) p(\boldsymbol{x}(i) \mid \theta, \boldsymbol{z}(i)) = \sum_{k=1}^{K} q_k(i) \mathcal{N}_c(\boldsymbol{x} \mid \mathbf{0}, \sigma_k(i)\mathrm{R}_k). \tag{4.28}$$

4.5 Maximum Log-Likelihood Estimation

Maximization of the data log-likelihood shown in Equation (4.7) can be iteratively solved by the expectation maximizing (EM) algorithm [85]. Let us denote the posterior distribution of the latent factors $\{z_k(i)\}_{k=1}^{K}$ given observed data and current parameter set θ_{old} by $\{m_k(i)\}_{k=1}^{K}$. In the E step, the posterior distribution is calculated for each multi-domain frame (i):

$$\begin{aligned} m_k(i) = p(z_k(i) = 1 \mid \boldsymbol{x}(i), \theta_{\text{old}}) &= \frac{p(z_k(i) = 1)p(\boldsymbol{x}(i) \mid z_k(i) = 1, \theta_{\text{old}})}{\sum_{k'=1}^{K} p(z_{k'}(i) = 1)p(\boldsymbol{x}(i) \mid z_{k'}(i) = 1, \theta_{\text{old}})} \\ &= \frac{q_k(i)\mathcal{N}_c\left(\boldsymbol{x} \mid \mathbf{0}, \sigma_k^{\text{old}}(i)\mathrm{R}_k^{\text{old}}\right)}{\sum_{k'=1}^{K} q_{k'}(i)\mathcal{N}_c\left(\boldsymbol{x} \mid \mathbf{0}, \sigma_{k'}^{\text{old}}(i)\mathrm{R}_{k'}^{\text{old}}\right)}. \end{aligned} \tag{4.29}$$

Under this posterior distribution of the latent variables found, the expected value of the complete data log-likelihood evaluated for some parameter set θ, denoted by $Q_1(\theta, \hat{\theta})$, is calculated:

$$Q_1(\theta, \hat{\theta}) = \mathbb{E}_{\boldsymbol{z}}(\boldsymbol{x}, \boldsymbol{z} \mid \theta^{\text{old}}, \theta) = \sum_{\boldsymbol{z}} p(\boldsymbol{z} \mid \boldsymbol{x}, \theta^{\text{old}}) \log p(\boldsymbol{x}, \boldsymbol{z} \mid \theta). \tag{4.30}$$

In the M step, this expected value is maximized. The update formulas are obtained by setting its derivative with respect to each parameter zero:

$$q_k = \frac{I_k}{I}, \tag{4.31}$$

$$\mathrm{R}_k = \frac{1}{I_k} \sum_{i=1}^{I} \frac{m_k(i)}{\sigma_k(i)} \hat{\mathrm{R}}_{\boldsymbol{x}(i)}, \tag{4.32}$$

$$\sigma_k(i) = \frac{1}{J} \mathrm{Tr}\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)} \mathrm{R}_k^{-1}\right), \tag{4.33}$$

where

$$I_k = \sum_{i=1}^{I} m_k(i), \tag{4.34}$$

$$\hat{\mathrm{R}}_{\boldsymbol{x}(i)} = \boldsymbol{x}^{\mathrm{H}}(i)\boldsymbol{x}(i). \tag{4.35}$$

Note that the determinant of each spatial covariance matrix is normalized in each iterative step, and R_k and $\sigma_k(i)$ are iteratively updated since their update formulas depend on each other.

After convergence of the EM algorithm, the expected value of the covariance matrix of each component is given by:

$$\hat{\mathrm{R}}_{\boldsymbol{c}_k(i)} = m_k(i)\hat{\sigma}_k(i)\hat{\mathrm{R}}_k, \tag{4.36}$$

and Wiener filtering is performed using Equation (4.10) to extract each component in each multi-domain frame. Overall algorithm is shown in Algorithm 2.

4.5.1 Initialization of model parameters

To perform the EM algorithm, model parameters need to be initialized. Initialization can be performed with random numbers or principal component analysis (PCA). Let us denote the eigen vectors and their corresponding eigen values of observational covariance matrix $\mathrm{R}_{\boldsymbol{x}(i)}$ by $\boldsymbol{v}_1(i), \boldsymbol{v}_2(i), \ldots, \boldsymbol{v}_J(i)$ and $u_1(i), u_2(i), \ldots, u_J(i)$ respectively, and assume they are sorted by descending order of their eigen values. Spatial covariance matrices and mixture ratio are initialized as:

$$\mathrm{R}_k^{\text{init}} = \sum_{i=1}^{I} \boldsymbol{v}_k(i)\boldsymbol{v}_k(i)^\top, \tag{4.37}$$

$$q_k = \frac{1}{K}, \tag{4.38}$$

and the initialization of $\sigma_k(i)$ is performed using Equation (4.33).

4.6 Maximum Posteriori Estimation Using Reference Signals

Well known EEG responses including ERPs, motor imagery, SSVEP are often used to pursue a research goal instead of exploring unseen EEG responses. Therefore, it is reasonable to assume that objective components of EEG can be recorded to obtain prior knowledge and more effective signal separation can be obtained by incorporating such prior information. In this section, we extend the framework that is just explained in the previous section to incorporate prior knowledge of objective

Algorithm 2 Wiener filtering based on maximum likelihood estimation

```
Require:
    $\boldsymbol{x}_k(i) \in \mathbb{C}^J$ for all $k$ and $i$                ▷ Observed activities
    $K > 1$ is an integer.                                  ▷ Number of components
Ensure:
    $\hat{\boldsymbol{c}}_k(i)$ is estimator of $\boldsymbol{c}_k(i)$
function EM_ML($\boldsymbol{x}$)
    for all $k$ do                               ▷ Initialization of model parameters
        calculate $\mathrm{R}_k^{\mathrm{old}}$ using Equation (4.37)
        calculate $q_k^{\mathrm{old}} \leftarrow \frac{1}{K}$
        for all $i$ do
            calculate $\sigma_k^{\mathrm{old}}(i)$ using Equation  (4.33)
        end for
    end for
    repeat
        for all $k$ and $i$ do
            update $m_k(i)$ using Equation (4.29)                          ▷ E step
        end for
        for all $k$ do                                                 ▷ M step
            update $\hat{\mathrm{R}}_k$ using Equation (4.32)
            for all $i$ do
                update $\sigma_k(i)$ using Equation (4.33)
            end for
        end for
        for all $k$ and $i$ do                                         ▷ M step
            update $q_k$ using Equation (4.31)
        end for
    until convergence                                      ▷ End of EM algorithm
    for all $k$ and $i$ do                                      ▷ Wiener filtering
        calculate $\hat{\mathrm{R}}_{\boldsymbol{c}_k(i)}$ using the Equation (4.36)
        calculate $\hat{\boldsymbol{c}}_k(i)$ using the Equation (4.10)
    end for
    return $\hat{\boldsymbol{c}}_k(i)$
end function
```

components by setting prior distributions of spatial covariance matrices whose hyper parameter is calculated using pre-recorded objective components.

4.6.1 Spatial Prior Distribution

Let us assume that J-dimensional random complex vectors that are independent from each other with a known mean and an unknown covariance matrix A^{-1}. Then, the conjugate prior distribution of the precision matrix A is *the Wishart distribution*, whose probability density function is provided as:

$$\mathrm{A} \sim \mathscr{W}_J(\mathrm{A} \mid \Psi, \phi) = B(\Psi, \phi)\mathrm{det}\,(\mathrm{A})^{(\phi-J-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\mathrm{A}\Psi^{-1}\right)\right), \tag{4.39}$$

where

$$B(\Psi, \phi) = \mathrm{det}\,(\Psi)^{-\phi/2}\left(2^{\frac{\phi J}{2}}\pi^{\frac{D(D-1)}{4}}\prod_{d=1}^{D}\Gamma\left(\frac{\phi+1-d}{2}\right)\right)^{-1},$$

$$\Gamma(\phi) = \int_0^\infty \exp(-t)t^{\phi-1}dt.$$

Γ is called the gamma function, Ψ is a positive definite Hermitian matrix, and $\phi > J-1$.

Maximum likelihood estimation is extended to maximum a posteriori (MAP) estimation, where posterior probability of parameters is maximized instead of likelihood. From the Bayes' theorem, the log posterior probability of spatial covariance matrices $L_2(\boldsymbol{x}, \sigma, \Psi, \phi)$ can be expressed as:

$$L_2(\boldsymbol{x}, \sigma, \Psi, \phi) \propto \log \prod_i^I p(\boldsymbol{x}(i) \mid \theta) \prod_k^K \mathscr{W}_J\left(\mathrm{R}_k^{-1} \mid \Psi_k, \phi_k\right), \tag{4.40}$$

$$\sigma = \{\sigma(i)\}_{i=1}^I, \quad \Psi = \{\Psi_k\}_{k=1}^K, \quad \phi = \{\phi_k\}_{k=1}^K. \tag{4.41}$$

By maximizing this posterior probability, a parameter set to construct Wiener filtering can also be found using the expectation maximization (EM) algorithm [86]. E step is the same to Equation (4.29). The function to be maximized in the M step can be expressed as a sum of $Q_1(\theta, \hat{\theta})$, the expected value under the posterior probability of the latent factors shown in Equation (4.30), and log prior distribution of parameters:

$$Q_2(\theta, \hat{\theta}) = Q_1(\theta, \hat{\theta}) + \sum_{k=1}^{K} \log \mathscr{W}_J \left(\mathrm{R}_k^{-1} \mid \Psi_k, \phi_k \right). \tag{4.42}$$

The update formulas of $q_k(i)$ and $\sigma_k(i)$ in the M step are same to Equation (4.31) and (4.33) respectively, and that of spatial covariance is obtained:

$$\hat{\mathrm{R}}_k = \frac{\Psi_k + 2\sum_{i=1}^{I} \frac{m_k(i)}{\sigma_k(i)} \hat{\mathrm{R}}_{\boldsymbol{x}(i)}}{\phi_k + J - 1 + 2I_k}. \tag{4.43}$$

Overall procedure is shown in Algorithm 3.

4.6.2 Obtaining Prior Information

Prior information of components are provided via the hyper parameters of the Wishart distribution $\{\Psi_k\}_{k=1}^{K}$ and $\{\phi_k\}_{k=1}^{K}$ that are calculated using pre-recorded components. Component activities can not be recorded alone because EEG signals are always mixture of multiple components. However, EEG signals in which a specific component is supposed to appear and not to appear can be recorded respectively to establish contrast between them. For example, synchronously averaged ERP can be used as a prototype of single-trial ERPs although they have different waveform trial-to-trial.

From Equation (4.42), effect of setting prior distribution for k-component can be written:

$$\log \mathscr{W}_J(\mathrm{R}_k^{-1} \mid \Psi_k, \phi_k) = -\frac{1}{2}\mathrm{Tr}(\Psi_k \mathrm{R}_k^{-1}). \tag{4.44}$$

This term takes the maximum value when $\Psi_k = \mathrm{R}_k$. Therefore, Ψ_k can be regarded as a prototype of R_k and obtained as:

$$\Psi_k = \frac{\sum_{i=1}^{I} \bar{\boldsymbol{c}}_k^{\mathrm{H}}(i) \bar{\boldsymbol{c}}_k(i)}{\det\left(\sum_{i=1}^{I} \bar{\boldsymbol{c}}_k^{\mathrm{H}}(i) \bar{\boldsymbol{c}}_k(i) \right)}, \tag{4.45}$$

Algorithm 3 Wiener filtering based on MAP estimation

```
Require:
    $\boldsymbol{x}_k(i) \in \mathbb{C}^J$ for all $k$ and $i$                          ▷ Observed activities
    $K > 1$ is an integer.                                          ▷ Number of components
    $\Psi_k$ is positive definite symmetric matrix whose elements are more than zero and
    determinant is one for all $k$.                                    ▷ Prototype of $\mathrm{R}_k$
Ensure:
    $\hat{\boldsymbol{c}}_k(i)$ is estimator of $\boldsymbol{c}_k(i)$
function EM_MAP($\boldsymbol{x}$)
    for all $k$ do                                  ▷ Initialization of model parameters
        $\mathrm{R}_k^{\mathrm{old}} \leftarrow \Psi_k$
        $q_k \leftarrow \frac{1}{K}$
        for all $i$ do
            calculate $\sigma_k(i)$ using Equation (4.33)
        end for
    end for
    repeat
        for all $k$ and $i$ do
            update $m_k(i)$ using Equation (4.29)                              ▷ E step
        end for
        for all $k$ do                                                         ▷ M step
            update $\hat{\mathrm{R}}_k$ using Equation (4.43)
            for all $i$ do
                update $\sigma_k(i)$ using Equation (4.33)
            end for
        end for
        for all $k$ and $i$ do                                                 ▷ M step
            update $q_k$ using Equation (4.31)
        end for
    until convergence                                       ▷ End of EM algorithm
    for all $k$ and $i$ do                                      ▷ Wiener filtering
        calculate $\hat{\mathrm{R}}_{\boldsymbol{c}_k(i)}$ using the Equation (4.36)
        calculate $\hat{\boldsymbol{c}}_k(i)$ using the Equation (4.10)
    end for
    return $\hat{\boldsymbol{c}}_k(i)$
end function
```

where a signal that contains the k-th component is denoted by $\bar{\boldsymbol{c}}_k(i)$.

4.7 Experimental Evaluation

In this section, ocular artifact removal experiment is described. The oddball paradigm experiments are employed again to evaluate the proposed and the ICA-based methods.

4.7.1 Data Acquisition

All EEG signals were recorded in a sound proof room with 25 scalp electrodes placed according to International 10-20 system at sampling rate of 1000 Hz. Three healthy subjects aged from 23 to 26 years old without any neurological disorders participated in the experiment. All experimental procedures were approved by the Ethical Review Board of Nara Institute of Science and Technology. The recorded EEG signals were down-sampled to 200 Hz, and a band-pass filter was applied between 0.01 Hz and 30 Hz.

The experiment consists of the four sessions. In the first session, a standard oddball paradigm paradigm experiment was performed in the same manner as Chapter 3. In the second session, an modified auditory oddball paradigm was performed where, as same as the precious session, subjects counted the number of target-stimuli presence, and intentionally made an eye blink every time they heard the either types of stimuli. In the third session, the resting state EEG of the subjects was recorded. In the last session, subjects did not do cognitive tasks and just made eye blinks every 1.5 seconds.

The data obtained in the first, third and fourth session were used for calculating the prior information, and the data obtained in the second session was used for test. If the potential produced by the intentional eye blinks are successfully removed, the N1 and P3 components are expected to be seen. Moreover, the P3 component should be seen larger at the target-trials than non-target trials.

The proposed method separated the test data into three components ($K = 3$), which have eye blink, ERP, and resting state prior, respectively. An ICA-based automatic artifact-removal method called ADJUST [50] was compared with the proposed method.

4.7.2 Result

The result is shown in Figure 4.7.2, where the black and white lines are the averaged target and the non-target trials at the channel F3, respectively. The left column shows the trial averaged data that were processed by only a band pass filter, which have extremely high amplitude and no ERP components are observable except the N1 component of the subject 3. The second column shows the data cleaned by ICA, where the waveforms are almost identical to the band pass filtered data. The third column shows the cleaned signal by GCPD introduced in Chapter 3. The fourth column shows the data processed by the proposed method. The N1 component can be seen at the subject 2 and 3, and the P3 component can be seen at the subject 1 and 3. However, the magnitude of P3s are almost identical in the target and non-target trials.

4.8 Unifying View of Tensor factorization and Maximum Likelihood Approach

The two noise removal approaches, tensor factorization described in Chapter 3 and maximum likelihood described in Chapter 4 can be comprehensively understood by revisiting the log-likelihood of observed data shown in Equation (4.7). From

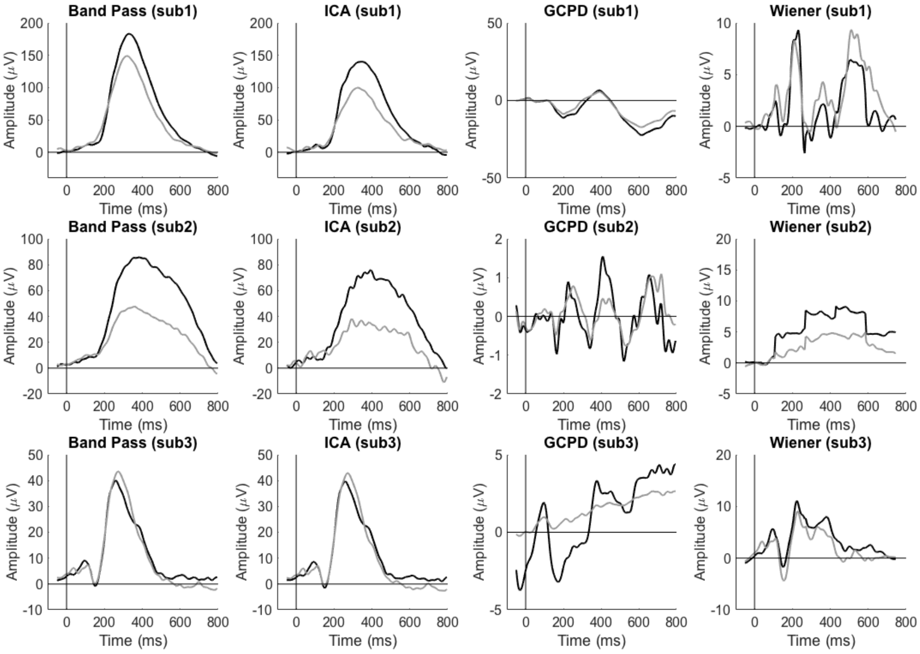


Figure 4.3: Result of ocular artifact removal.

Equation (4.7), (4.17), and (4.19) we obtain :

$$\begin{aligned} L_1(\boldsymbol{x} \mid \theta) &= \log \sum_{i=1}^{I} \mathcal{N}_c\left(\boldsymbol{x}(i) \mid \mathbf{0}, \mathrm{R}_{\boldsymbol{x}(i)}\right) \\ &\stackrel{\mathrm{c}}{=} \sum_{i=1}^{I} \log\left(\det\left(\mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right)\right) - \boldsymbol{x}(i)^{\mathrm{H}} \mathrm{R}_{\boldsymbol{x}(i)}^{-1} \boldsymbol{x}(i) \\ &\stackrel{\mathrm{c}}{=} -\sum_{i=1}^{I} -\log\left(\det\left(\mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right)\right) + \mathrm{Tr}\left(\boldsymbol{x}(i)\boldsymbol{x}(i)^{\mathrm{H}} \mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right) \\ &\stackrel{\mathrm{c}}{=} -\sum_{i=1}^{I} -\log\left(\det\left(\mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right)\right) + \mathrm{Tr}\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)} \mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right) + \log\left(\det\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)}\right)\right) \\ &\stackrel{\mathrm{c}}{=} -\sum_{i=1}^{I} -\log\left(\det\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)} \mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right)\right) + \mathrm{Tr}\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)} \mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right), \end{aligned} \tag{4.46}$$

where

$$\log\left(\det\left(\mathrm{R}_{\boldsymbol{x}(i)}^{-1}\right)\right) + \log\left(\det\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)}\right)\right) = \log\left(\det\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)}^{-1} \mathrm{R}_{\boldsymbol{x}(i)}\right)\right). \tag{4.47}$$

Equation (4.47) holds because both of $\hat{\mathrm{R}}_{\boldsymbol{x}(i)}$ and $\mathrm{R}_{\boldsymbol{x}(i)}$ are positive definite (See Appendix). Consequently, the following equation holds:

$$\begin{aligned} L_1(\mathrm{X} \mid \theta) \stackrel{\mathrm{c}}{=} -\sum_{i=1}^{I} D_{gIS}\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)}, \mathrm{R}_{\boldsymbol{x}(i)}\right) = -\sum_{i=1}^{I} D_{gIS}\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)}, \sum_{k=1}^{K} \mathrm{R}_{\boldsymbol{c}_k(i)}\right) \\ = -\sum_{i=1}^{I} D_{gIS}\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)}, \sum_{k=1}^{K} \left(\boldsymbol{a}_k^{(1)} \circ \boldsymbol{a}_k^{(2)} \circ \cdots \circ \boldsymbol{a}_k^{(N)}\right)_{n_1, n_2, \ldots, n_N} \boldsymbol{b}_k \circ \boldsymbol{b}_k\right). \end{aligned} \tag{4.48}$$

Therefore, maximum likelihood estimation of spatial covariance matrices under the probabilistic generative model defined in Section 4.4 is mathematically equivalent to performing matrix factorization for spatial covariance matrices in each multi-domain frame based on gIS divergence shown in Equation 2.22.

4.8.1 Spatially Independent Model

The spatial covariance matrices are generally expressed as a sum of outer products of array manifold vectors:

$$\mathrm{R}_k = \mathrm{B}_k\mathrm{B}_k^{\mathrm{H}} = \sum_{d=1}^{D} \begin{pmatrix} |b_k(1,d)|^2 & \dots & b_k(J,d)b_k^*(1,d) \\ \vdots & \ddots & \vdots \\ b_k(J,d)b_k^*(J,d) & \dots & |b_k(J,d)|^2 \end{pmatrix}$$

$$= \begin{pmatrix} |b_k(1)|^2 & \dots & b_k(J)b_k^*(1) \\ \vdots & \ddots & \vdots \\ b_k(1)b_k^*(J) & \dots & |b_k(J)|^2 \end{pmatrix}, \tag{4.49}$$

where complex conjugate is denoted by *.

Let us assume that component activations captured by J electrodes are spatially independent, which means the spatial covariance matrices are diagonal matrices:

$$\mathrm{R}_k = \mathrm{B}_k\mathrm{B}_k^{\mathrm{H}} = \mathrm{diag}(\boldsymbol{b}_k), \tag{4.50}$$

$$\boldsymbol{b}_k = \left(|b_k(1)|^2, |b_k(2)|^2, \dots, |b_k(J)|^2\right), \tag{4.51}$$

where $\mathrm{diag}(\boldsymbol{b}_k)$ is the diagonal matrix, whose diagonal elements are the elements of $\boldsymbol{b}_k$. Under this assumption, the observational likelihood is measured ignoring non-diagonal component:

$$
\begin{aligned}
L_3(\boldsymbol{x} \mid \theta) &\overset{c}{=} -\sum_{i=1}^{I} D_{gIS}\left(\hat{\mathrm{R}}_{\boldsymbol{x}(i)}, \mathrm{R}_{\boldsymbol{x}(i)}\right) \\
&\overset{c}{=} -\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{|\hat{x}_j(i)|^2}{|x_j(i)|^2} - \log\frac{|\hat{x}_j(i)|^2}{|x_j(i)|^2}\right) \\
&\overset{c}{=} -\sum_{i=1}^{I}\sum_{j}^{J} d_{IS}\left(|\hat{x}_j(j)|^2, |x_j(j)|^2\right) \\
&\overset{c}{=} -\sum_{i=1}^{I}\sum_{j}^{J} d_{IS}\left(|\hat{x}_j(j)|^2, \sum_{k=1}^{K}(\sigma_k(i)\,|b_k(j)|)^2\right) \\
&= -\sum_{i=1}^{I}\sum_{j=1}^{J} d_{IS}\left(|\hat{x}_j(i)|^2, \sum_{k=1}^{K}\left(\boldsymbol{a}_k^{(1)} \circ \boldsymbol{a}_k^{(2)} \circ \cdots \circ \boldsymbol{a}_k^{(N)} \circ \boldsymbol{b}_k\right)_{n_1,n_2,\ldots,n_N,j}\right).
\end{aligned}
\tag{4.52}
$$

Therefore, in contrast to the general model shown in Equation (4.48), maximum likelihood estimation of the spatially independent model is mathematically equivalent to performing tensor factorization based on the IS divergence shown in Equation (2.20).

4.8.2 Difference between Unconstrained and Diagonal Models

Equation (4.48) and (4.52) illustrates the difference between the two model in terms of outer product.

The spatially independent model treats all modes equally, which means outer product is calculated using all modes including the electrode mode. The element-wise distances between the estimated and observed tensor are measured by IS divergence, and all of element-wise divergences are summed as a the total divergence to be minimized.

On the other hand, the general model regards the spatial mode as special one, and define the divergence to be minimized as follows: it excludes the electrode mode from outer product calculation, instead calculates the covariance matrix of the mode (spatial covariance matrix) , each element of the outer product tensor is multiplied to the covariance matrix, the distances between the estimated and observed covariance matrices are measured in each multi-domain frame by the gIS divergence shown in, and then the distances of all frames are summed to obtain the total divergence.

Because the general model is more general than the other, it is inherently more computationally expensive and largely depends on the initial values of the iterative update. This fact can be understood in terms of the numbers of parameters to be estimated. Let us denote the number of parameters by the general model by v_1 and the spatially independent model by v_2, respectively. Then, they are calculated as:

$$v_1 = K\left(\frac{KJ(J-1)}{2} + I_1 I_2 \dots I_N\right), \tag{4.53}$$

$$v_2 = K\left(J + I_1 + I_2 + \cdots + I_N\right). \tag{4.54}$$

This indicates the number of the general model parameters increases exponentially and it suffers from the curse of dimensionality while that of the spatially independent model does linearly.

4.9 Discussion About Probabilistic Modeling of Amplitude Generation

We assumed STFT coefficients of each component follow complex Gaussian so far. This section gives its theoretical validation and further discussion about probabilistic modeling of amplitude generation

As same as most of signal separation methods, we have assumed *the additivity of components*:

$$|x(i)|^\alpha = \sum_{k=1}^{K} |c_k(i)|^\alpha, \tag{4.55}$$

where $|c(i)|^\alpha$ is called fractional spectrum [87], amplitude spectrum with $\alpha = 1$, or power spectrum with $\alpha = 2$. In general, the above equality does not hold because phase difference of components causes canceling out each other. However, the additivity can be justified in terms of the expected value if the component STFT coefficient $c(i)$ is modeled by a probabilistic distribution with stability.

4.9.1 Symmetric Alpha Distribution

Let x_1, x_2, and x_3 be random variables that are independently and identically distributed (i.i.d). They are said to be strictly stable if there exists a positive number c for any positive numbers a and b such that:

$$ax_1 + bx_2 \overset{\text{dist.}}{=} cx_3, \tag{4.56}$$

where $\overset{\text{dist.}}{=}$ denotes equality in distribution [87, 88]. We say that a random vector x_3 satisfying Equation 4.56 is α-stable because there exists a constant $\alpha \in]0\,2]$ called the characteristic exponent that satisfies:

$$c = (a^\alpha + b^\alpha)^{1/\alpha}. \tag{4.57}$$

If $x_3 \overset{\text{dist.}}{=} -x_3$, it is called symmetric α-stable and denoted by $S\alpha S$.

We say a complex random variable $z = x_1 + ix_2$ is complex $S\alpha S$ if the random vector $\left(x_1^\top, x_2^\top\right)^\top$ is $S\alpha S$. Moreover, z is said to be isotopic complex $S\alpha S$ if the following relation holds:

$$\forall \theta \in [0\,2\pi[, \quad \exp(i\theta)z \overset{\text{dist.}}{=} z. \tag{4.58}$$

Regarding to signal processing application, isotopic complex $S\alpha S$ is particularly important and parameterized by a scale parameter σ^α. We denote it by $S\alpha S_c(\sigma^\alpha)$. Although there is no closed-form expression for $S\alpha S_c(\sigma^\alpha)$ except the case of $\alpha = 1$ or 2, the characteristic function ϕ is given by:

$$\begin{aligned} z = x_1 + ix_2 &\sim S\alpha S_c(\sigma^\alpha), \\ \phi(z) = \mathbb{E}[\exp(i(\theta_1 x_1 + \theta_2 x_2))] &= \exp(-\sigma^\alpha |\boldsymbol{\theta}|^\alpha), \end{aligned} \tag{4.59}$$

where $|\boldsymbol{\theta}|^\alpha$ is the Euclidian norm of the vector $\left(\theta_1\,\theta_2\right)$.

Table 4.1: The value of α, additive spectrum, generative distribution, and divergence.

α	Additive Spectrum	Generative Distribution	Divergence
1	amplitude spectrum	Cauchy distribution	Cauchy divergence
2	power spectrum	Gaussian distribution	IS divergence

4.9.2 Distribution Stability and Component Additivity

Let $z_1(i)$ and $z_2(i)$ are STFT coefficients in the multi-domain frame (i), and follow $S\alpha S_c$ independently as:

$$z_1(i) \sim S\alpha S_c(|z_1(i)|^\alpha), \quad z_2(i) \sim S\alpha S_c(|z_2(i)|^\alpha),$$

then, the distribution reproductivity holds:

$$z_1(i) + z_2(i) \sim S\alpha S_c(|z_1(i)|^\alpha + |z_2(i)|^\alpha), \tag{4.60}$$

which means the additivity of component activation is justified in terms of the expected value.

When $\alpha = 1$, $S\alpha S_c(|z(i)|)$ is equivalent to the complex Cauchy distribution, and component additivity is satisfied in expected amplitude spectrum. If the generation of the component activity $c_k(i)$ is assumed to follow the complex Cauchy distribution, maximum likelihood estimation is equivalent to minimizing the Cauchy divergence (Equation (2.21)) between $x(i)$ and $\sum_k c_k(i)$ [89]. Therefore, application of tensor factorization to amplitude spectrum based on the Cauchy divergence is theoretically justified. Similarly, when $\alpha = 2$, $S\alpha S_c(|z(i)|^2)$ is equivalent to the univariate complex Gaussian distribution, and component additivity is satisfied in expected power spectrum, which is the case of Equation (4.52). As already seen, if the generative distribution is the complex Gaussian distribution, maximum likelihood estimation is equivalent to minimizing the IS divergence [90]. Therefore, tensor factorization to power spectrum with IS-divergence is also verified. Relations between the value of α, the expression of component additivity that holds additivity, a probabilistic distribution to model a generation of the complex value of c_k, and the corresponding divergence are summarized in Table 4.1.

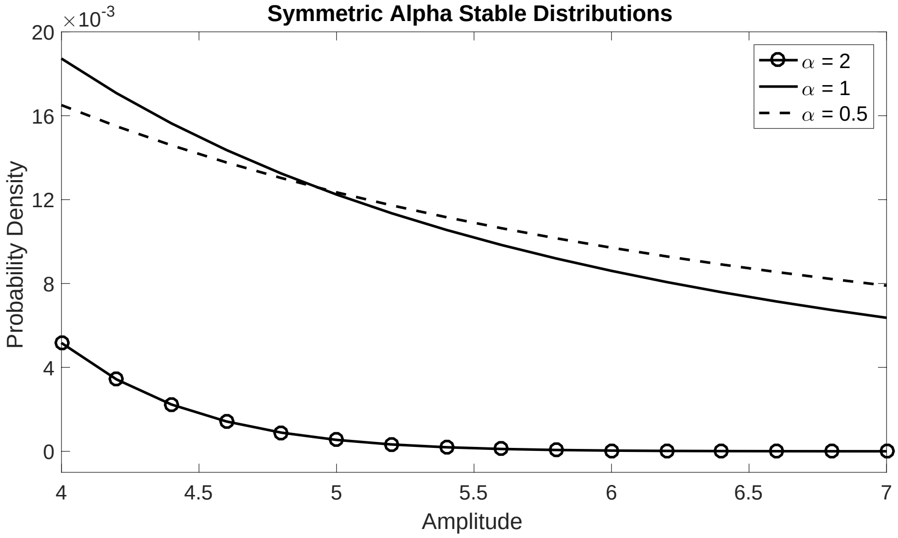


Figure 4.4: Tails of symmetric alpha stable distribution with different value of α.

Although stability of multivariate distribution is much more complicated and difficult to compute, similar discussion is available [91].

An important property of $S\alpha S_c(\sigma^\alpha)$ is that the constant α controls the heaviness of the density tail. As shown in Figure 4.9.2, smaller α makes the distribution tail heavier. Signal separation based on a distribution with smaller α is worth investigating because an EEG signal is often impulsive and include outliers and heavy-tailed distribution is more suitable for modeling such a kind of signal.

4.10 Summary

In this Chapter, we proposed a noise-removal method based on spatial Wiener filtering. A probabilistic generative model of EEG components in the multi-domain was created and the filter coefficients are calculated to maximize the posterior probability of the observed EEG signal.

Spatial spread of each component was encoded in the spatial covariance matrix and their prior knowledge were incorporated to estimate model parameters setting their prior distribution.

Discussion about the equivalence to tensor factorization was also provided, where maximum likelihood estimation assuming Gaussian distribution is equivalent to minimization of IS divergence. Furthermore, we discussed the theoretical validity of the component additivity under the Gaussian distribution assumption. We explained the Gaussian distribution falls under the umbrella of the symmetric alpha distribution, which gives the validity in the terms of the expected value of power spectrum.

Chapter 5

Quality Prediction of Synthesized Speech Using EEG

5.1 Introduction

Text-to-Speech (TTS) systems, which convert a written text into speech, and are becoming more widely implemented in mobile phones, car navigation systems, and other consumer electronics. Such systems play a critical role in many applications because speech is the most fundamental and easiest communication tool for human beings. Therefore, synthesized speeches must sound natural for good machine-to-human communications.

The research of TTS systems needs reasonable criteria that evaluate the qualities of synthesized speeches. Several current evaluation methods have their own advantages and disadvantages: (1) subjective ratings [92–94], (2) speech feature analysis [95–97], and (3) physiological response analysis [98–105].

Naturalness and intelligibility are the most commonly aspects for subjective quality judgment of TTS. Naturalness describes how close synthesized speech is to human speech, and intelligibility reflects how well the speech content can be heard. The former is usually measured by a mean opinion score (MOS) test [92], and the latter is gauged by semantically unpredictable sentences (SUS) [93]. In addition, valence and arousal are often used to evaluate the subjective impressions of speech [103, 105–107] and to model emotions [108–111]. Valence reflects a positive or a negative emotion. Arousal reflects the degree of intensity or activation. In a MOS test, subjects listen to speech and rate its relative perceived quality on some kind of a scale, for example, “excellent,” “good,” “fair,” “poor,” “bad.” Then the scores are averaged across subjects. This is well established method for which references on how to perform it are available [94], making it the only standard way to evaluate the naturalness quality of synthesized speech. However, their appropriateness has not been fully proven because high inter- and intra-subject inconsistencies are often observed in the ratings, resulting in poor reproductivity [112]. Another problem is that a MOS test can only provide an overall impression without any further detailed information about the speech.

Speech feature analysis automatically evaluates speech quality at its signal level by software that inputs a speech file and outputs the estimated speech quality. The fundamental frequency or the mel-feats are usually extracted as representative features from synthesized and human speeches, and their root mean square error (RMSE) is used as an evaluation metric. Advantages of these methods include complete reproductivity and less time consumption after such software is developed. However, appropriateness is difficult to prove because the exact relationship between the acoustic features and the perceived quality of speech by a listener is not well understood [112]. In fact, speech quality must be evaluated not only physically but also psychologically because it is commonly defined as an assessment result within which a listener compares his/her perceptions with expectations [113, 114].

Physiological response analysis methods for multi media including speech are emerging recently [115]. Even though these methods have not been established yet, they are worth investigating because physiological signals can be recorded automatically and continuously to provide insight about listeners’ cognitive states without interruptions caused by directly asking them to answer questions. Among existing non-invasive physiological response measures, EEG has especially great potential to estimate a listener’s perceived speech quality for the following reasons. EEGs can be recorded at a higher temporal resolution, which can be important to evaluate speech quality since the temporal structure of speech largely affects its perceived quality. In addition, as already explained, EEG recording equipment is relatively small and portable. Measuring physiological responses to speech in daily environments is critical because speech is everywhere. Despite the above advantages, the main disadvantage of physiological measures is the difficulty of data gathering. The amount of data that can be collected from a subject is limited for practical and ethical reasons. Conducting experiments is usually time-consuming and labor-intensive. Furthermore, the recording of multiple EEG channels are usually highly correlated to each other, which makes the features extracted from them less informative compared to the height of their dimensions. These aspects of EEG (limited amount of data, noise, and high correlation and dimension) complicate training a predictive model with EEG data and require a sophisticated dimension reduction or regularization techniques [116].

Existing researches have analyzed EEG responses to speech stimuli using event-related potentials (ERP), which are time-locked responses to external or internal events in terms of a voltage change that are usually visualized and quantified after synchronous averaging of multiple epochs [99–101]. Due to its definition, measuring ERP needs the instantaneous time-locking points at which an event occurs, complicating the use of ERP if stimuli onsets are gradual or unclear [33]. Therefore, ERP is not suitable for our purpose of the predicting perceived quality of speech whose length exceeds a second because it is usually unclear which time points affect a listener's perceived quality. Other research used power spectral density [102, 117] and difference between scalp EEG channels [103, 105] at multiple frequency bands. Neuroscience studies reported that EEG spectral changes in distinct regions and between hemispheres are related to emotions [118–121]. Other studies used EEG phase synchronization between EEG channel pairs and found a correlation to emotions [122, 123].

The purpose of this research is to predict the perceived qualities of synthesized speeches just using EEG. Interest is growing in the development of a machine learning algorithm that uses an input/output data structure as tensor formats [124–126]. Such tensor structured features were investigated in this study because EEG signals can have structures in time, frequency, space, experimental condition, and other modalities.

Table 5.1: Subjective rating scales.

Rating Scale	Abbreviation	Description
Overall Impression	MOS	1 (Bad) to 5 (Excellent)
Valence	VAL	1 (Negative) to 9(Positive)
Arousal	ARL	1 (Unexcited) to 9 (Excited)

5.2 Materials

In this study, we used the Physyqx data set [98], which consists of speech files, their subjective rating scores from 21 subjects, and EEG signals from the same subjects that recorded while they listened to the speech. The data recording protocol was approved by the INRS Research Ethics Office, and participants gave informed consent for their participation and to make their data anonymous and freely available online. The details of the data set and the experimental procedures are available in [98]. We obtained it by an e-mail request.

5.2.1 Speech Stimuli

The speech stimuli presented to the subjects in the data set consist of speech collected from four humans and seven commercially available TTS systems. From each human and each TTS system, four English sentences were collected whose durations ranged from 13 to 22 seconds. The 44 human and synthesized speeches were presented randomly to each subject.

5.2.2 Experimental Procedure

The experiment's timeline is shown in Figure 5.1. A 15-second rest period was provided before each stimulus presentation followed by a subjective rating period during which the subjects evaluated the speech to which they had just listened. The subjective rating scales used in this study are shown in Table 5.1 and include overall impression (MOS), valence (VAL), and arousal (ARL). MOS was evaluated with a 5-scale rating and the others with a 9-scale using self-assessment manikin [127].

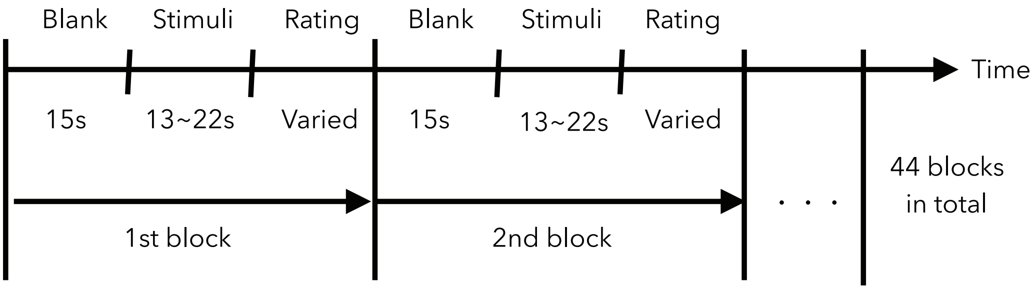


Figure 5.1: Timeline of EEG and subjective evaluation data recording experiment.

5.2.3 EEG Recording and Preprocess

EEG data were recorded throughout the experiment with 64 scalp channels. The sampling rate was 512 Hz, which was down-sampled to 256 Hz. All the channels were placed on scalp according to the modified 10/20 system [128]. Some channels were removed from further analysis by visual inspection because they were noisy. We applied a band-pass filter to all the data between 0.5-50 Hz and applied an independent component analysis based semi-artifact removal technique using the ADJUST toolbox [50]. After preprocessing, the EEG signal of each subject was cut into 44 epochs corresponding to the stimuli listening periods.

5.3 Methods

Multivariate regression analysis was performed to predict three response variables, the overall impression, the valence, and arousal using the features extracted from the EEG signals.

5.3.1 Feature Extraction and Construction

All of the features for the regression analysis performed in this study were extracted at five frequency bands from a channel or a channel pair. The frequency bands include delta (δ : 1–4 Hz), theta (θ : 4–8 Hz), alpha (α : 8–12 Hz), beta (β : 12–30 Hz), and gamma (γ : 30–45 Hz). Let us denote the Fourier transformation at a frequency of f_k of the n-th training trial recorded by the m-th channel by $\boldsymbol{x}_{n,m}(f_k)$

where k is the index of the frequency bins. An estimator of the power spectrum density and a phase spectrum denoted by p_k and h_k can be calculated using the periodogram method as follows:

$$p_{n,m}(f_k) = \frac{1}{T}\left|\boldsymbol{x}_{n,m}(f_k)\right|^2, \tag{5.1}$$

$$h_{n,m}(f_k) = \text{angle}(\boldsymbol{x}_{n,m}(f_k)), \tag{5.2}$$

where T is the number of time samples. Then, we averaged the power spectrum density over the frequency bins within the range of each frequency band to define channel-based features $\text{PSD}_n(c, f)$ as follows:

$$\text{PSD}_n(m, f) = \frac{1}{|D_f|}\sum_{f_k \in D_f} p_{n,m}(f_k), \tag{5.3}$$

where D_f is the index set of the frequency bins that are included in the range of the f-th frequency band and $|D_f|$ is the number of the elements in D_f. The channel-pair-based features are also defined using the averaged power spectrum density and the phase spectrum as follows:

$$\text{PWD}_n(m_1, m_2, f) = \text{PSD}(t, m_1, f) - \text{PSD}_n(m_2, f), \tag{5.4}$$

$$\text{PHD}_n(m_1, m_2, f) = \frac{1}{|D_f|}\sum_{f_k \in D_f} h_{n,m_1}(f_k) - h_{n,m_2}(f_k). \tag{5.5}$$

If M EEG channels and F frequency bands are used (F=5 in this study), $I = F\left(M(M-1) + M\right)$ features are calculated. The feature matrix X can be expressed as:

$$\text{X} = \left(\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(N)\right)^{\top} \in \mathbb{R}^{N \times I}, \tag{5.6}$$

where N is the number of training trials and $\boldsymbol{x}(n)$ is a feature vector of the n-th trial and has all the features PSD_n, PWD_n, and PHD_n.

To exploit structures of the features, we organized the features as a tensor $\mathscr{X} \in$

$\mathbb{R}^{N\times M\times M\times F}$ as follows:

$$\mathcal{X}(n,m_1,m_2,f)=\begin{cases}\mathrm{PWD}_n(m_1,m_2,f) & (m_1>m_2)\\ \mathrm{PHD}_n(m_1,m_2,f) & (m_1<m_2)\\ \mathrm{PSD}_n(m_1,f) & (m_1=m_2)\end{cases}. \tag{5.7}$$

The feature matrix and tensor are depicted in Figure 5.2.

5.3.2 Regression Analysis

Among dimension reduction or regularization techniques, higher order partial least square (HOPLS) [124] and standard partial least square (PLS) [129, 130] simultaneously perform dimension reduction and regression and are used for regression analysis because the former is a natural extension of the latter so that features can be structured as a tensor.

Let us denote the feature matrix and the response matrix by X and Y, respectively. X has all the features of all training trials while Y has all the response variables of all training trials:

$$\mathrm{Y}=\left(\boldsymbol{y}(1),\boldsymbol{y}(2),\ldots,\boldsymbol{y}(N)\right)^\top\in\mathbb{R}^{N\times J}, \tag{5.8}$$

where N, and J are the number of training trials, features, and response variables to be predicted, respectively. $\boldsymbol{x}(n)$ is a feature vector of the n-th trial and has all the features explained above, and $\boldsymbol{y}(n)$ is the response vector that has all the response variables of the n-th trial.

PLS performs a simultaneous decomposition of X and Y to find common latent variables $\boldsymbol{t}_r\in\mathbb{R}^N$ as:

$$\mathrm{X}=\sum_{r=1}^{R_1}\boldsymbol{t}_r\boldsymbol{p}_r^\top+\mathrm{E}, \tag{5.9}$$

$$\mathrm{Y}=\sum_{r=1}^{R_1}\boldsymbol{t}_r\boldsymbol{q}_r^\top+\mathrm{F}, \tag{5.10}$$

where E and F are the residual matrices, and R_1 is called the number of the components.

On the other hand, HOPLS can be similarly formulated as the problem to find latent variables as follows:

$$\mathcal{X} = \sum_{r=1}^{R_2} \mathcal{G}_r \times_1 \boldsymbol{t}_r \times_2 \mathrm{P}_r^{(1)} \times_3 \mathrm{P}_r^{(2)} \times_4 \mathrm{P}_r^{(3)} + \mathcal{E}, \tag{5.11}$$

$$\mathcal{G}_r \in \mathbb{R}^{1\times M\times M\times F}, \quad \mathrm{P}_r^{(k)} \begin{cases} \in \mathbb{R}^{M\times L_k} & (k=1,2) \\ \in \mathbb{R}^{F\times L_k} & (k=3) \end{cases}, \tag{5.12}$$

$$\mathrm{Y} = \sum_{r=1}^{R_2} \boldsymbol{t}_r \boldsymbol{q}_r^{\top} + \mathrm{F}, \tag{5.13}$$

where $\mathcal{G}_r$ is called the core tensor, $\mathcal{E}$ and V are the residuals, R_2 is the number of the components. $\mathrm{P}_r^{(n)}$ is called the loading matrix of the r-th component, and L_k is called the number of the k-mode loadings.

If data are plentiful, which is rare in EEG studies, the best approach for training and evaluating the performance of a predictive model is to randomly divide the dataset into three parts: training, validation, and test sets, which are respectively used to train a model, tune hyper-parameters or select a model, and evaluate the generalization error [131]. However, since the amount of data in this study is too small to exploit such an ideal protocol, we instead used leave-one-out cross-validation for each subject. The hyper-parameter R_1 of PLS varied from 1 to 43, loadings of the channel-1 L_1 and the channel-2 L_2 ranged from 1 to 7. The loading gs of the frequency band L_3 and the number of components R_2 of HOPLS ranged from 1 to 5. The result of the models that achieved the best performance was reported in Section 5.4.

5.3.3 Evaluation Metrics

Root mean squared error (RMSE) was used to quantify the predictability of the trained predictive models for their overall impression, valence, and arousal.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{5.14}$$

where N is the number of test samples, $\hat{y}_i$ is the predicted value for the i-th test data, and y_i is the actually observed value. All were calculated individually for each subject.

5.4 Results

Table 5.2 summarizes RMSE and the numbers of latent factors identified by PLS and HOPLS. Predictions with tensorial features generally made smaller errors than the vectorized ones for all the rating scales. Figure 5.3 summarizes which features contributed to the 3 calculated by taking the magnitude of the regression coefficients. The number of features of each type in the 100 most contributed features across subjects is shown. PHD contributed the most; PSD rarely appeared in the list of the top 100 features list for all of the rating scales. Among the five frequency bands, the α band contributed the most to the MOS prediction, followed by the β band. For the VAL and ARL predictions, the β band contributed the most, followed by the α band. The top ten channel pairs, which contributed the most to the MOS prediction of subjects 1, 2, 3, and 4, are shown in Figure 5.4. The numbers of latent factors identified by PLS and HOPLS are also shown in Table 5.2.

5.5 Discussion

Channel-pair-based features (PWD and PHD) contributed more to the predictions than channel-based features (PSD). This result agrees with a previous study [121] and suggests the importance of considering scalp EEG dynamics between brain regions and that graph theory based features and functional connectivity analysis can be effective [132, 133]. The association of the lateral (left-right) spectral difference (DLAT) is well documented [118, 120]. In addition, spectral differences in caudality (DCAU) between the anterior and posterior [104, 134] or the front-posterior brain regions [121] have been investigated. In this study, both DLAT and DCAU contributed to the predictions (Figure 5.4) although their effectiveness was dependent on the subjects.

Quality prediction models were independently trained for each subject because

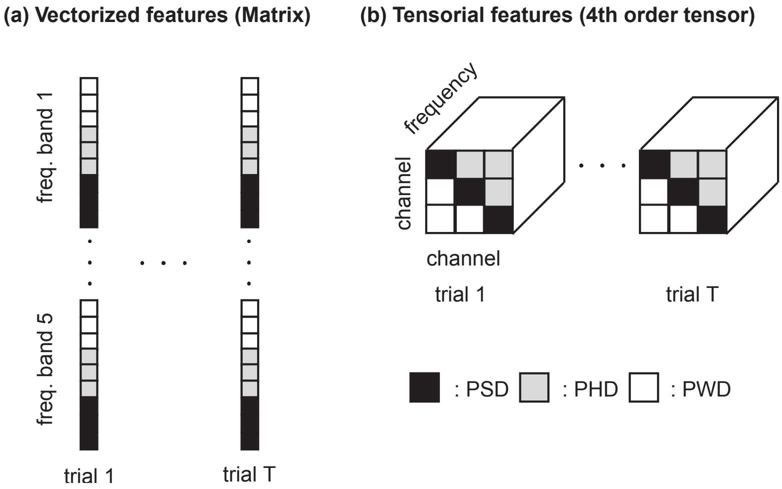


Figure 5.2: Schematic image of vectorized and tensorial features. (a) Vectorized feature is included in a matrix whose first and second modes are trials and features. PWD, PHD, and PSD at each frequency band are lined as a vector in each trial. (b) Tensor structured dependent variable with four modes: trials, channel-1, channel-2, and frequency bands. Tensor elements with a larger channel-1 index than channel-2 are PWD, and a smaller channel-1 index than channel-2 are PHD, and identical channel indexes are PSD.

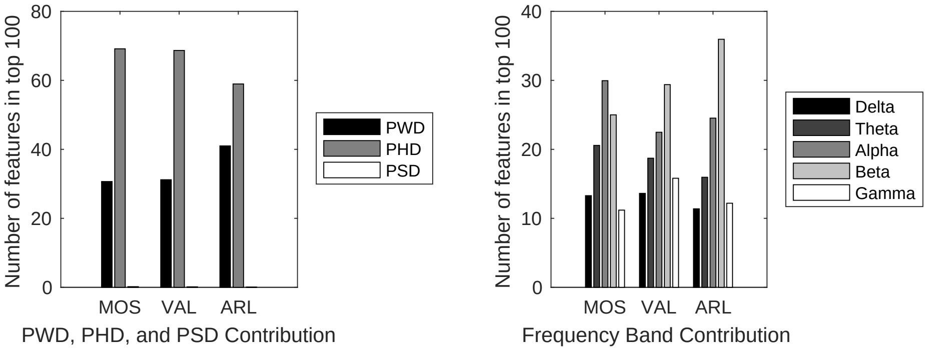
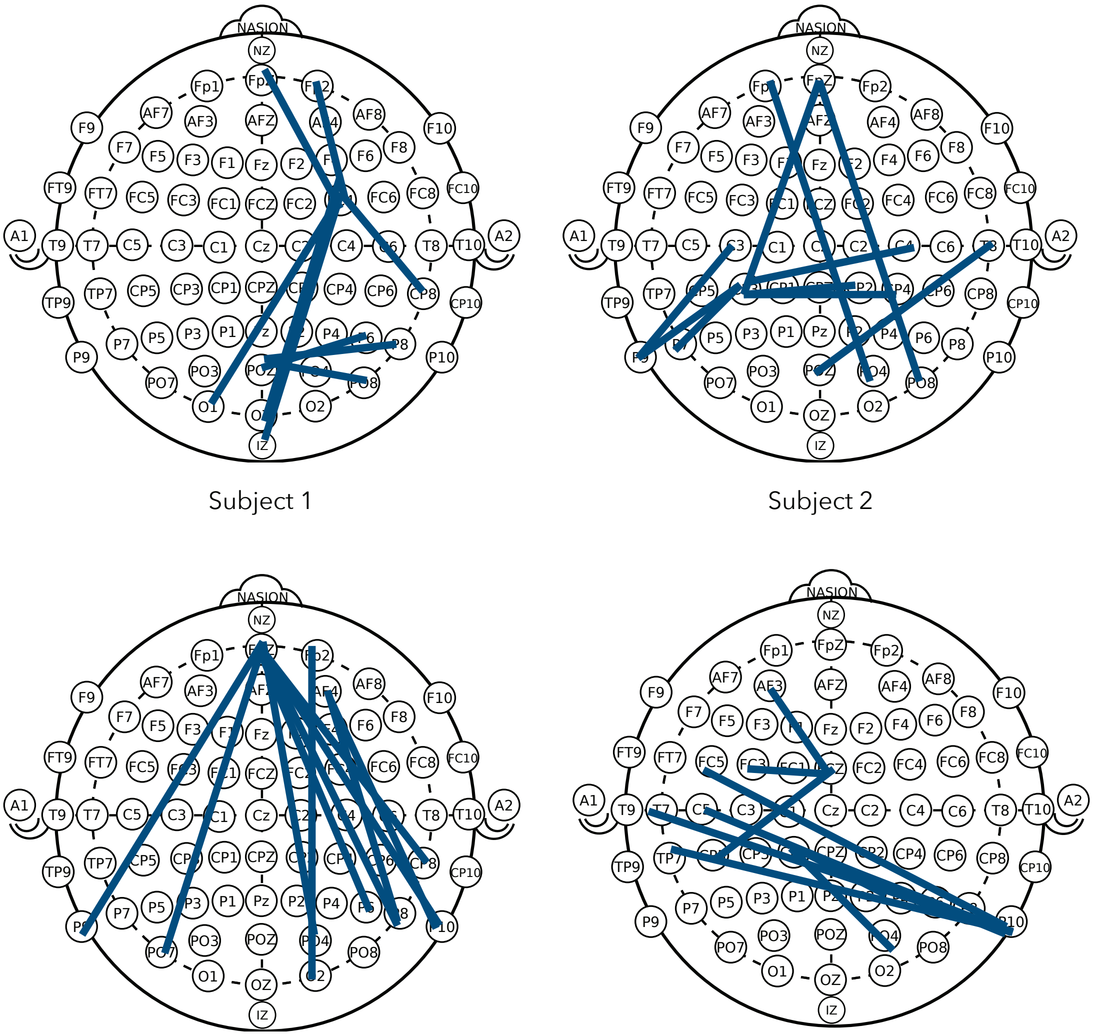


Figure 5.3: Contribution of features. Feature contributions were calculated by taking magnitude of their regression coefficients. (Left) Numbers of PWD, PHD, and PSD among 100 features that most greatly contributed to the prediction of each rating scale among all features. (Right) Number of features of each frequency band among 100 features that most greatly contributed to each rating scale.



Subject 1

Subject 2

Subject 3

Subject 4

Figure 5.4: Contribution of channel pairs. Top ten channel pairs that contributed the most to MOS predictions of subjects 1, 2, 3, and 4. This figure was made by modifying the original one, which is distributed under the public domain dedication [79]

Table 5.2: Prediction results and the number of latent factors identified.

subject	RMSE (vector)			RMSE (tensor)			R_1	L_1	L_2	L_3	R_2
	MOS	VAL	ARL	MOS	VAL	ARL					
1	1.091	1.090	1.086	0.961	0.957	0.962	1	3	4	5	3
2	1.042	1.057	1.058	0.987	0.958	0.975	1	4	4	4	3
3	0.997	1.018	1.020	0.907	0.936	0.931	2	7	3	4	2
4	1.148	1.110	1.043	0.949	0.944	0.994	2	7	5	2	2
5	1.187	1.225	1.229	1.082	1.177	1.150	1	6	7	4	5
6	1.260	1.292	1.151	1.000	2.176	1.941	4	7	4	4	4
7	0.979	0.981	1.007	0.869	0.935	1.167	1	2	3	5	5
8	1.155	1.186	1.125	0.970	0.989	1.017	1	1	5	1	1
9	1.215	1.221	1.160	0.996	0.994	1.023	2	7	7	1	2
10	1.111	1.112	1.022	0.957	0.985	1.144	1	5	4	3	4
11	1.125	1.243	1.0.19	1.013	1.047	0.920	3	7	7	1	1
12	0.996	1.193	1.177	0.641	0.912	0.680	29	4	2	3	4
13	1.258	1.227	1.234	1.051	1.050	1.035	3	4	2	5	4
14	0.991	1.102	1.040	0.940	0.980	0.929	12	7	1	2	3
15	1.022	0.989	0.969	0.965	0.927	0.934	1	7	3	2	5
16	1.196	1.206	1.087	1.058	1.047	1.027	4	7	6	2	5
17	1.021	1.083	1.087	0.884	0.886	0.924	3	4	7	4	3
18	1.055	1.027	1.092	0.915	0.920	0.969	2	1	3	4	4
19	1.130	1.142	1.126	1.021	1.081	1.020	1	5	1	5	4
20	0.995	1.028	0.944	0.887	0.990	1.057	1	4	5	2	5
21	1.121	1.157	1.103	0.900	0.969	0.997	1	1	1	4	2

emotion regulation is reportedly dependent on individuals [135]. The commonality of the channels/channel pairs, which greatly contributed to the predictions, was actually rather small (Figure 5.4). Therefore, creating subject-independent features is an interesting future work. However, note that the alpha and beta bands commonly contributed to the predictions while the effective channels/channel pairs differed depending on the subjects. The alpha and beta bands contributed more largely to the predictions than the other frequency bands, which is in line with previous neurophysiological studies. The relationship between alpha band asymmetry and the withdrawal or disengagement from a stimulus or negative valence is well documented in response to a variety of stimuli, including pictures [136, 137], music [121, 134, 138], movies [139], and speech [103, 105]. The beta band, which contributed the most to the ARL predictions, is reportedly associated with arousal and emotional experiences [140, 141].

Gupta et. al. [105] predicted the MOS values using the same data set that we used in this study (they are the creators of the data set). Their study used a simple linear regression model with not only EEG but also speech features. They reported the RMSE of their model was 0.117, which is much lower than our model, and can suggest that speech features are much more informative than EEG features to predict subjective quality ratings although they failed to report the detailed procedure of model training and test.

Neither previous work nor our current study advocate that physiological assessment methods of speech quality should replace subjective rating methods or signal analysis methods because, as stated in Section 5.1, each method has its own advantages and disadvantages and they can complement each other.

Features were extracted and constructed as tensors as described in Section 5.3.1 and 5.3.2, but other features and construction ways are also possible. For example, if time-frequency analysis is employed, times frames can be treated as one of the tensor modes.

5.6 Summary

This chapter performed prediction of subjective quality ratings for synthesized speech solely based on EEG. We created tensorial features that consider all of the channel-based and channel-pair-based features in terms of power and phase spectrum at multiple frequency bands and used them as regression features. The experimental result showed that tensorial features more effectively predicted the subjective ratings than vectorized ones, and the trained predictive models were neurophysiologically plausible.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

In this doctoral dissertation, methods to enable single-trial analysis of EEG have been presented.

Chapter 3 and 4 discussed single-trial noise removal from EEG signals, where different methods are respectively proposed, namely tensor factorization with graph structure and spatial Wiener filtering with spatial correlation prior. In the both of the method, incorporating prior knowledge played a key role. In Chapter 3, EEG electrode location was given as a prior knowledge to regularize CPD, where prior knowledge of an objective component (for example, ERP) was not given. On the other hand, in Chapter 4, prior knowledge of an objective component was given explicitly as spatial covariance matrices. Moreover, they can be understood in an unified manner in the terms of divergence minimization.

In Chapter 5, we tackled the prediction of perceived quality of synthesized speech. To our best of our knowledge, this study is the first study which attempts to predict subjective rating scales of naturalness to synthesized and natural speeches solely based on EEG.

6.2 Future Directions

We have come to the end of this dissertation. It was a long way, but there are so much more to do. Here are a part of them.

Obtaining adjacency matrix

Electrode similarities have to be given as a prior knowledge, and we calculated them solely based on their relative distances. This similarity measure is particularly effective when amount of EEG data is small because it is calculated without any EEG recording. However, if a plenty of data is available it is worth considering to calculate electrode similarities based on recorded EEG data. For example, in [142], the functional connectivities of the brain [143, 144], which were synchronization of different brain regions, were estimated to define a graph adjacency matrix based on recorded EEG data and their correlation coefficients and mutual information. Graph structure learning is gathering much attention from researchers recently [145, 146].

Using graph structure in other tensor decomposition model

We used the EEG electrode location to regularize CPD. Tucker decomposition, tensor train decomposition [147], and HOPLS can be extended in a similar way.

Component selection

It is important to automatically select a subset of the decomposed components of interest. We did component selection in a heuristic manner using the difference of the trial mode bases between experimental conditions. Even though this approach worked well in this study, a better component selection method should be invented. Especially, a selection method that don't use information of stimuli classes is desired, whereas our method does.

Initialization

We proposed an initialization method using graph Fourier bases, which are orthogonal each other as same as HOSVD initialization. However, it is pointed out recently that orthogonality might not be a good property for a basis of matrix factorization [148]. It is worth investigate an initialization method based on a property other than orthogonality.

Probabilistic distribution to model EEG signals

In Chapter 4, amplitude generation of EEG sources are expressed using a probabilistic generative model, and its covariance matrix parameters were exploited to construct an adaptive filter to extract objective components. Discussion about the equivalence of tensor factorization and maximum likelihood estimation was given, which shows a selection of probabilistic distribution is the same as a selection of divergence. Moreover, the theoretical validity of component additivity were also discussed. As shown in Figure 4.9.2, the complex symmetric alpha stable distribution has a heavy tail to be robust to outliers when the characteristic exponent α is small, which can be desirable to model an impulsive and noisy signal such as EEG. However, the optimal value of α to model EEG signals is not known. Although there is an experimental study about the optimal value of α to model sound signals [87], similar discussions in an EEG case are not available to the best of our knowledge.

Convex Optimization

The optimization of CPD is a non-convex problem, which is only guaranteed to converge locally and can suffer from a local minima. However, there are endeavors to re-formulate it as a convex problem [149, 150]. A method of convex relaxation based on the general alpha stable distribution with a graph structure should be investigated.

Detecting degraded speech parts

In Chapter 5, we predicted the subjective rating scales of speech. Analysis unit was by entire of speech. However, it is possible that particular parts of speech affect the perceived qualities. Detecting such degraded parts exploiting high temporal resolution of EEG can contribute to the future TTS research.

Perceptual dimensions

We predicted the subjective rating scale of the overall impression, valence, and arousal as the previous works did [103, 105–107], various perpetual dimensions other than them have been also proposed recently to model emotions or perceived quality-of-experiences [151, 152], which were not investigated in the current research. For example, voice peasantness [94], prosody [112], intonation [153], and naturalness [154].

Online analysis

Throughout this dissertation, EEG analysis is limited to offline manners. Online analysis should be investigated because it is important for some EEG application including brain-machine interfaces.

This paper ends here, but I hope these points are investigated in the future by someone or me.

References

[1] Hans Berger. “Ueber das Elektrenkephalogramm des Menschen”. *Archives fur Psychiatrie Nervenkrankheiten* 87 (1929).

[2] Sabine Weiss and Horst M Mueller. “The contribution of EEG coherence to the investigation of language”. *Brain and Language* 85.2 (2003), pp. 325–343.

[3] Marcel Bastiaansen and Peter Hagoort. “Oscillatory neuronal dynamics during language comprehension”. *Progress in Brain Research* 159 (2006), pp. 179–196.

[4] Maria Pefkou, Luc H Arnal, Lorenzo Fontolan, and Anne-Lise Giraud. “θ-Band and β-Band Neural Activity Reflects Independent Syllable Tracking and Comprehension of Time-Compressed Speech”. *Journal of Neuroscience* 37.33 (2017), pp. 7930–7938.

[5] Marcela Peña and Lucia Melloni. “Brain oscillations during spoken sentence processing”. *Journal of Cognitive Neuroscience* 24.5 (2012), pp. 1149–1164.

[6] Ashley G Lewis and Marcel Bastiaansen. “A predictive coding framework for rapid neural dynamics during sentence-level language comprehension”. *Cortex* 68 (2015), pp. 155–168.

[7] Marta Kutas, Gregory McCarthy, and Emanuel Donchin. “Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time”. *Science* 197.4305 (1977), pp. 792–795. DOI: `10.1126/science.887923`.

[8] David Carmel and Shlomo Bentin. “Domain specificity versus expertise: factors influencing distinct processing of faces”. *Cognition* 83.1 (2002), pp. 1–29.

[9] Geraldine Dawson, Leslie Carver, Andrew N Meltzoff, Heracles Panagiotides, James McPartland, and Sara J Webb. "Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development". *Child Development* 73.3 (2002), pp. 700–717.

[10] Renana H Ofan, Nava Rubin, and David M Amodio. "Situation-based social anxiety enhances the neural processing of faces: evidence from an intergroup context". *Social Cognitive and Affective Neuroscience* 9.8 (2013), pp. 1055–1061.

[11] Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu. "EEG-based emotion recognition during watching movies". *Proceeding of the fifth International IEEE EMBS Conference on Neural Engineering* (2012), pp. 1097–1105.

[12] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "DEAP: A database for Emotion Analysis using Physiological Signals". *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 18–31.

[13] James A Coan and John J. B. Allen. "Frontal EEG asymmetry as a moderator and mediator of emotion". *Biological Psychology* 67.1 (2004), pp. 7–50.

[14] Jaeseung Jeong. "EEG dynamics in patients with Alzheimer's disease". *Clinical Neurophysiology* 115.7 (2004), pp. 1490–1505.

[15] Joseph C McBride, Xiaopeng Zhao, Nancy B Munro, Gregory A Jicha, Frederick A Schmitt, Richard J Kryscio, Charles D Smith, and Yang Jiang. "Sugihara causality analysis of scalp EEG for detection of early Alzheimer's disease". *NeuroImage: Clinical* 7 (2015), pp. 258–265.

[16] Andrzej Cichocki, Sergei L Shishkin, Toshimitsu Musha, Zbigniew Leonowicz, Takashi Asada, and Takayoshi Kurachi. "EEG filtering based on blind source separation (BSS) for early detection of Alzheimer's disease". *Clinical Neurophysiology* 116.3 (2005), pp. 729–737.

[17] Oliver Faust, U Rajendra Acharya, Hojjat Adeli, and Amir Adeli. "Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis". *Seizure* 26 (2015), pp. 56–64.

[18] Mohamad A Mikati, Edwin Trevathan, Kalpathy S Krishnamoorthy, and Cesare T Lombroso. “Pyridoxine-dependent epilepsy: EEG investigations and long-term follow-up”. *Electroencephalography and Clinical Neurophysiology* 78.3 (1991), pp. 215–221.

[19] U Rajendra Acharya, S Vinitha Sree, G Swapna, Roshan Joy Martis, and Jasjit S Suri. “Automated EEG analysis of epilepsy: a review”. *Knowledge-Based Systems* 45 (2013), pp. 147–165.

[20] David Gloss, Jay K Varma, Tamara Pringsheim, and Marc R Nuwer. “Practice advisory: The utility of EEG theta/beta power ratio in ADHD diagnosis Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology”. *Neurology* 87.22 (2016), pp. 2375–2379.

[21] Valery A Ponomarev, Andreas Mueller, Gian Candrian, Vera A Grin-Yatsenko, and Juri D Kropotov. “Group independent component analysis (gICA) and current source density (CSD) in the study of EEG in ADHD adults”. *Clinical Neurophysiology* 125.1 (2014), pp. 83–97.

[22] Adam R Clarke, Robert J Barry, RORY McCARTHY, and Mark Selikowitz. “Electroencephalogram differences in two subtypes of attention-deficit/hyperactivity disorder”. *Psychophysiology* 38.2 (2001), pp. 212–221.

[23] Milos Matouek, Peder Rasmussen, and Christopher Gillberg. “EEG frequency analysis in children with so-called minimal brain dysfunction and related disorders”. *Advances in Biological Psychiatry* 15 (1984), pp. 102–108.

[24] Jonathan Wolpaw and Elizabeth Winter Wolpaw. *Brain-computer interfaces: principles and practice*. OUP USA, 2012.

[25] Hiroshi Morioka, Atsunori Kanemura, Jun-ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii. “Learning a common dictionary for subject-transfer decoding with resting calibration”. *NeuroImage* 111 (2015), pp. 167–178.

[26] Lynn M McCane, Susan M Heckman, Dennis J McFarland, George Townsend, Joseph N Mak, Eric W Sellers, Debra Zeitlin, Laura M Tenteromano, Jonathan R Wolpaw, and Theresa M Vaughan. "P300-based brain-computer interface (BCI) event-related potentials (ERPs): People with amyotrophic lateral sclerosis (ALS) vs. age-matched controls". *Clinical Neurophysiology* 126.11 (2015), pp. 2124–2131.

[27] Lawrence Ashley Farwell and Emanuel Donchin. "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials". *Electroencephalography and Clinical Neurophysiology* 70.6 (1988), pp. 510–523. DOI: `10.1016/0013-4694(88)90149-6`.

[28] Eric W Sellers, Dean J Krusienski, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. "A P300 event-related potential brain–computer interface (BCI): the effects of matrix size and inter stimulus interval on performance". *Biological Psychology* 73.3 (2006), pp. 242–252. DOI: `10.1016/j.biopsycho.2006.04.007`.

[29] George Townsend, BK LaPallo, CB Boulay, DJ Krusienski, GE Frye, CKea Hauser, NE Schwartz, TM Vaughan, Jonathan R Wolpaw, and EW Sellers. "A novel P300-based brain–computer interface stimulus presentation paradigm: moving beyond rows and columns". *Clinical Neurophysiology* 121.7 (2010), pp. 1109–1120. DOI: `10.1016/j.clinph.2010.01.030`.

[30] Klaus-Robert Müller, Michael Tangermann, Guido Dornhege, Matthias Krauledat, Gabriel Curio, and Benjamin Blankertz. "Machine learning for real-time single-trial EEG-analysis: From brain–computer interfacing to mental state monitoring". *Journal of Neuroscience Methods* 167.1 (2008), pp. 82–90.

[31] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks". *Aviation, space, and environmental medicine* 78.5 (2007), B231–B244.

[32] Khald Ali I Aboalayon, Miad Faezipour, Wafaa S Almuhammadi, and Saeid Moslehpour. "Sleep Stage Classification Using EEG Signal Analysis: A Com-

prehensive Survey and New Investigation”. *Entropy* 18.9 (2016). DOI: 10.3390/e18090272.

[33] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[34] Emily S. Kappenman and Steven J. Luck, eds. *The Oxford Handbook of Event-Related Potential Components*. Oxford University Press, Dec. 2011. DOI: 10.1093/oxfordhb/9780195374148.001.0001.

[35] Paul L Nunez and Ramesh Srinivasan. *Electric Fields of the Brain: The Neurophysics of EEG*. second. Oxford University Press, 2006.

[36] Emanuel Donchin. “Discriminant analysis in average evoked response studies: the study of single trial data”. *Electroencephalography and Clinical Neurophysiology* 27.3 (1969), pp. 311–314.

[37] Tzyy-Ping Jung, Scott Makeig, fMarissa Westerfield, Jeanne Townsend, Eric Courchesne, and Terrence J Sejnowski. “Analysis and visualization of single-trial event-related potentials”. *Human brain mapping* 14.3 (2001), pp. 166–185.

[38] Roger Ratcliff, Marios G Philiastides, and Paul Sajda. “Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG”. *Proceedings of the National Academy of Sciences* 106.16 (2009), pp. 6539–6544.

[39] Guillaume A Rousselet, Carl M Gaspar, Kacper P Wieczorek, and Cyril R Pernet. “Modeling single-trial ERP reveals modulation of bottom-up face visual processing by top-down task constraints (in some subjects)”. *Frontiers in Psychology* 2 (2011).

[40] Tim R Mullen, Christian AE Kothe, Yu Mike Chi, Alejandro Ojeda, Trevor Kerth, Scott Makeig, Tzyy-Ping Jung, and Gert Cauwenberghs. “Real-time neuroimaging and cognitive monitoring using wearable dry EEG”. *IEEE Transactions on Biomedical Engineering* 62.11 (2015), pp. 2553–2567.

[41] Emily S. Kappenman and Steven J. Luck. “The effects of electrode impedance on data quality and statistical significance in ERP recordings”. *Psychophysiology* 47.5 (2010), pp. 888–904. DOI: 10.1111/j.1469-8986.2010.01009.x.

[42] Steven J Luck. “Sources of dual-task interference: Evidence from human electrophysiology”. *Psychological Science* 9.3 (1998), pp. 223–227.

[43] Marta Kutas, Steven A Hillyard, et al. “Reading senseless sentences: Brain potentials reflect semantic incongruity”. *Science* 207.4427 (1980), pp. 203–205. DOI: `10.1126/science.7350657`.

[44] Yen Na Yum, Katherine J. Midgley, Phillip J. Holcomb, and Jonathan Grainger. “An ERP study on initial second language vocabulary learning”. *Psychophysiology* 51.4 (2014), pp. 364–373. DOI: `10.1111/psyp.12183`.

[45] James W Tanaka and Tim Curran. “A neural basis for expert object recognition”. *Psychological Science* 12.1 (2001), pp. 43–47.

[46] Bruno Rossion, Chun-Chia Kung, and Michael J Tarr. “Visual expertise with nonface objects leads to competition with the early perceptual processing of faces in the human occipitotemporal cortex”. *Proceedings of the National Academy of Sciences of the United States of America* 101.40 (2004), pp. 14521–14526. DOI: `10.1073/pnas.0405613101`.

[47] Emanuel Donchin and Michael GH Coles. “Is the P300 component a manifestation of context updating?” *Behavioral and Brain Sciences* 11.3 (1988), pp. 357–374. DOI: `10.1017/S0140525X00058027`.

[48] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Vol. 46. John Wiley & Sons, 2004. DOI: `10.1002/0471221317`.

[49] Christian Jutten and Jeanny Herault. “Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture”. *Signal Processing* 24.1 (1991), pp. 1–10. DOI: `10.1016/0165-1684(91)90079-X`.

[50] Andrea Mognon, Jorge Jovicich, Lorenzo Bruzzone, and Marco Buiatti. “ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features”. *Psychophysiology* 48.2 (2011), pp. 229–240. DOI: `10.1111/j.1469-8986.2010.01061.x`.

[51] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. “Independent component analysis of electroencephalographic data”. *Advances in Neural Information Processing Systems* (1996), pp. 145–151.

[52] Tamara G Kolda and Brett W Bader. "Tensor decompositions and applications". *SIAM review* 51.3 (2009), pp. 455–500.

[53] Age Smilde, Rasmus Bro, and Paul Geladi. *Multi-Way Analysis: Applications in the Chemical Sciences*. John Wiley & Sons, 2005.

[54] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multiway data analysis and blind source separation*. John Wiley & Sons, 2009.

[55] William James and Charles Stein. "Estimation with Quadratic Loss". *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability* 1.1961 (1961), pp. 361–379.

[56] Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. "Log-Determinant Divergences revisited: Alpha-Beta and Gamma Log-Det Divergences". *Entropy* 17.5 (2015), pp. 2988–3034. DOI: 10.3390/e17052988.

[57] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. "Multichannel Extensions of Non-negative Matrix Factorization with Complex-valued Data". *IEEE Transactions on Audio, Speech, and Language Processing* 21.5 (2013), pp. 971–982. DOI: 10.1109/TASL.2013.2239990.

[58] J. Douglas Carroll and Jih-Jie Chang. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition". *Psychometrika* 35.3 (1970), pp. 283–319. DOI: 10.1007/BF02310791.

[59] Richard A Harshman. "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis". *UCLA Working Papers in Phonetics* (1970).

[60] Frank L Hitchcock. "The expression of a tensor or a polyadic as a sum of products". *Studies in Applied Mathematics* 6.1-4 (1927), pp. 164–189.

[61] Joseph B Kruskal. "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics". *Linear algebra and its applications* 18.2 (1977), pp. 95–138.

[62] André Uschmajew. "Local convergence of the alternating least squares algorithm for canonical tensor approximation". *SIAM Journal on Matrix Analysis and Applications* 33.2 (2012), pp. 639–652.

[63] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "A multilinear singular value decomposition". *SIAM journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[64] Steven Lemm, Gabriel Curio, Yevhen Hlushchuk, and Klaus-Robert Muller. "Enhancing the signal-to-noise ratio of ICA-based extracted ERPs". *IEEE Transactions on Biomedical Engineering* 53.4 (2006), pp. 601–607.

[65] Fengyu Cong, Qiu-Hua Lin, Li-Dan Kuang, Xiao-Feng Gong, Piia Astikainen, and Tapani Ristaniemi. "Tensor decomposition of EEG signals: a brief review". *Journal of neuroscience methods* 248 (2015), pp. 59–69.

[66] Fengyu Cong, Anh Huy Phan, Qibin Zhao, Tiina Huttunen-Scott, Jukka Kaartinen, Tapani Ristaniemi, Heikki Lyytinen, and Andrzej Cichocki. "Benefits of multi-domain feature of mismatch negativity extracted by non-negative tensor factorization from EEG collected by low-density array". *International journal of neural systems* 22.06 (2012), p. 1250025.

[67] Morten Mørup, Lars Kai Hansen, Christoph S Herrmann, Josef Parnas, and Sidse M Arnfred. "Parallel factor analysis as an exploratory tool for wavelet transformed event-related EEG". *NeuroImage* 29.3 (2006), pp. 938–947. DOI: `10.1016/j.neuroimage.2005.08.005`.

[68] Fumikazu Miwakeichi, Eduardo Martnez-Montes, Pedro A Valdés-Sosa, Nobuaki Nishiyama, Hiroaki Mizuhara, and Yoko Yamaguchi. "Decomposing EEG data into space–time–frequency components using parallel factor analysis". *NeuroImage* 22.3 (2004), pp. 1035–1045.

[69] Martin Weis, Florian Romer, Martin Haardt, Dunja Jannek, and Peter Husar. "Multi-dimensional space-time-frequency component analysis of event related EEG data using closed-form PARAFAC". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2009), pp. 349–352. DOI: `10.1109/ICASSP.2009.4959592`.

[70] Andre Nikolaevich Tikhonov and Vasili Arsenin. *Solutions of ill-posed problems*. Vol. 14. Winston Washington, DC, 1977.

[71] Hiroshi Higashi, Andrzej Cichocki, and Toshihisa Tanaka. "Regularization using geometric information between sensors capturing features from brain signals". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2012), pp. 721–724.

[72] Fabien Lotte and Cuntai Guan. "Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms". *IEEE Transactions on Biomedical Engineering* 58.2 (2011), pp. 355–362.

[73] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. "Graph regularized nonnegative matrix factorization for data representation". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8 (2011), pp. 1548–1560.

[74] Robert Oostenveld and Peter Praamstra. "The five percent electrode system for high-resolution EEG and ERP measurements". *Clinical neurophysiology* 112.4 (2001), pp. 713–719.

[75] Marc R Nuwer. "Recording electrode site nomenclature." *Journal of Clinical Neurophysiology* 4.2 (1987), pp. 121–134.

[76] Jaakko Malmivuo and Robert Plonsey. *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields*. Oxford University Press, USA, 1995.

[77] Fan RK Chung. *Spectral graph theory*. 92. American Mathematical Soc., 1997.

[78] Mikhail Belkin and Partha Niyogi. "Laplacian eigenmaps and spectral techniques for embedding and clustering". *Proceedings of the Advances in Neural Information Processing Systems* (2002), pp. 585–591.

[79] URL: `https://commons.wikimedia.org/wiki/File:International_10-20_system_for_EEG-MCN.svg`.

[80] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains". *IEEE Signal Processing Magazine* 30.3 (2013), pp. 83–98.

[81] Piia Astikainen, Fengyu Cong, Tapani Ristaniemi, and Jari K Hietanen. "Event-related potentials to unattended changes in facial expressions: detection of regularity violations or encoding of emotions?" *Frontiers in human neuroscience* 7 (2013).

[82] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval. "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model". *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (2010), pp. 1830–1840. DOI: 10.1109/TASL.2010.2050716.

[83] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. "Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization". *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.9 (2016), pp. 1626–1641. DOI: 10.1109/TASLP.2016.2577880.

[84] Alexey Ozerov and Cédric Févotte. "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation". *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2010), pp. 550–563.

[85] Christopher M Bishop. *Pattern Recognition and Machine Learning*. springer, 2006.

[86] Aathur P. Dempster, Nan M. Laird, and Donald B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38.

[87] Antoine Liutkus and Roland Badeau. "Generalized Wiener filtering with fractional power spectrograms". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2015), pp. 266–270. DOI: 10.1109/ICASSP.2015.7177973.

[88] Panayiotis G. Georgiou, Panagiotis Tsakalides, and Chris Kyriakakis. "Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise". *IEEE Transactions on Multimedia* 1.3 (1999), pp. 291–301. DOI: DOI:10.1109/6046.784467.

[89] Antoine Liutkus, Derry Fitzgerald, and Roland Badeau. “Cauchy nonnegative matrix factorization”. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2015), pp. 1–5. DOI: `10.1109/WASPAA.2015.7336900`.

[90] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis”. *Neural Computation* 21.3 (2009), pp. 793–830. DOI: `10.1162/neco.2008.04-08-771`.

[91] John P Nolan. “Multivariate elliptically contoured stable distributions: theory and estimation”. *Computational Statistics* 28.5 (2013), pp. 2067–2089. DOI: `10.1007/s00180-013-0396-7`.

[92] CCITT. “Absolute category Rating (ACR) method for subjective testing of digital processors, Redbook” (1984).

[93] Christian Benoît, Martine Grice, and Valérie Hazan. “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences”. *Speech Communication* 18.4 (1996), pp. 381–392. DOI: `10.1016/0167-6393(96)00026-X`.

[94] ITU-T Recommendation P.85. *A Method for Subjective Performance Assesment of the Quality of the Speech Voice Output Devices*. International Telecommunication Union Telecommunication Standardization Sector, 1996.

[95] Arnd Mariniak. “A Global Framework for the Assessment of Synthetic Speech Without Subjects”. *Proceedings of the third European Conference on Speech Communication and Technology* (1993), pp. 1683–1686.

[96] Christoph R. Norrenbrock, Florian Hinterleitner, Ulrich Heute, and Sebastian Möller. “Quality prediction of synthesized speech based on perceptual quality dimensions”. *Speech Communication* 66 (2015), pp. 17–35. DOI: `10.1016/j.specom.2014.06.003`.

[97] ITU-T Recommendation P.862. *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*. International Telecommunication Union Telecommunication Standardization Sector, 2001.

[98] Rishabh Gupta, Hubert J. Banville, and Tiago H. Falk. "PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience". *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2015), pp. 1–5. DOI: `10.1109/WASPAA.2015.7336888`.

[99] Jan-Niklas Antons. *Neural Correlates of Quality Perception for Complex Speech Signals*. Springer, 2015.

[100] Jan-Niklas Antons, Robert Schleicher, and Sebastian Arndt. "Analyzing Speech Quality Perception Using Electroencephalography". *IEEE Journal of Selected Topics in Signal Processing* 6.6 (2012), pp. 721–731. DOI: `10.1109/JSTSP.2012.2191936`.

[101] Jan-Niklas Antons, Benjamin Blankertz, Gabriel Curio, Sebastian Möller, Anne K Porbadnigk, and Robert Schleicher. "Subjective Listening Tests and Neural Correlates of Speech Degradation in Case of Signal-Correlated Noise". *Audio Engineering Society Convention 129* (2010).

[102] Jan-Niklas Antons, Robert Schleicher, Sebastian Arndt, Sebastian Möller, and Gabriel Curio. "Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations". *Proceedings of the fourth International Workshop on Quality of Multimedia Experience* (2012), pp. 63–67. DOI: `10.1109/QoMEX.2012.6263840`.

[103] Rishabh Gupta, Khalil Laghari, Hubert Banville, and Tiago H Falk. "Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling". *Human-centric Computing and Information Sciences* 6.1 (2016), p. 5. DOI: `10.1186/s13673-016-0062-5`.

[104] Michela Sarlo, Giulia Buodo, Silvia Poli, and Daniela Palomba. "Changes in EEG alpha power to different disgust elicitors: the specificity of mutilations". *Neuroscience Letters* 382.3 (2005), pp. 291–296. DOI: `10.1016/j.neulet.2005.03.037`.

[105] Sebastian Arndt, Jan-Niklas Antons, Rishabh Gupta, Khalil ur Rehman Laghari, Robert Schleicher, Sebastian Möller, and Tiago H. Falk. "The effects of text-to-speech system quality on emotional states and frontal alpha band power".

Proceedings of the sixth International IEEE/EMBS Conference on Neural Engineering (2013), pp. 489–492. DOI: 10.1109/NER.2013.6695978.

[106] Meysam Asgari, Géza Kiss, Jan van Santen, Izhak Shafran, and Xubo Song. "Automatic measurement of affective valence and arousal in speech". *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing* (2014), pp. 965–969. DOI: 10.1109/ICASSP.2014.6853740.

[107] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space". *IEEE Transactions on Affective Computing* 2.2 (2011), pp. 92–105. DOI: 10.1109/T-AFFC.2011.9.

[108] James A. Russell and Albert Mehrabian. "Evidence for a three-factor theory of emotions". *Journal of Research in Personality* 11.3 (1977), pp. 273–294. DOI: 10.1016/0092-6566(77)90037-X.

[109] James A. Russell. "A circumplex model of affect". *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178.

[110] James A. Russell and Lisa Feldman Barrett. "Prototypical emotional episodes, and other things called emotion: dissecting the elephant". *Journal of Personality and Social Psychology* 76.5 (1999), pp. 805–819. DOI: 10.1037/0022-3514.76.5.805.

[111] James A. Russell. "Core affect of and the phychological construction of emotion". *Psychological Review* 110.1 (2003), pp. 145–172. DOI: 10.1037/0033-295X.110.1.145.

[112] Catherine Mayo, Robert A. J. Clark, and Simon King. "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis". *Speech Communication* 53.3 (2011), pp. 311–326. DOI: 10.1016/j.specom.2010.10.003.

[113] Ute Jekosch. *Voice and Speech Quality Perception: Assessment and Evaluation*. Springer Science and Business Media, 2006.

[114] Sebastian Möller, Florian Hinterleitner, Tiago H Falk, and Tim Polzehl. “Comparison of approaches for instrumentally predicting the quality of text-to-speech systems”. *Proceedings of the 11th annual conference of the International Speech Communication Association* (2010).

[115] Sebastian Arndt, Kjell Brunnström, Eva Cheng, Ulrich Engelke, Sebastian Möller, and Jan-Niklas Antons. “Review on using physiology in quality of experience”. *Electronic Imaging* 2016.16 (2016), pp. 1–9. DOI: `10.2352/ISSN.2470-1173.2016.16.HVEI-125`.

[116] Benson Mwangi, Tian Siva Tian, and Jair C. Soares. “A Review of Feature Reduction Techniques in Neuroimaging”. *Neuroinformatics* 12.2 (2014), pp. 229–244. DOI: `10.1007/s12021-013-9204-3`.

[117] Jan-Niklas Antons, Friedemann Köster, Sebastian Arndt, Sebastian Möller, and Robert Schleicher. “Changes of vigilance caused by varying bit rate conditions”. *Proceedings of the fifth International Conference on Quality of Multimedia Experience* (2013), pp. 148–151. DOI: `10.1109/QoMEX.2013.6603228`.

[118] Richard J. Davidson. “Anterior cerebral asymmetry and the nature of emotion”. *Brain Cognition* 20 (1992), pp. 125–151.

[119] L. I Aftanas, N. V Reva, A. A Varlamov, S. V Pavlov, and V. P. Makhnev. “Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics”. *Neuroscience and Behavioral Physiology* 34.8 (2004), pp. 859–867. DOI: `10.1023/B:NEAB.0000038139.39812.eb`.

[120] James A Coan and John J. B Allen. “Frontal EEG asymmetry as a moderator and mediator of emotion”. *Biological Psychology* 67.1 (2004), pp. 7–50. DOI: `10.1016/j.biopsycho.2004.03.002`.

[121] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. “EEG-based emotion recognition in music listening”. *IEEE Transactions on Biomedical Engineering* 57.7 (2010), pp. 1798–1806. DOI: `10.1109/TBME.2010.2048568`.

[122] You-Yun Lee and Shulan Hsieh. “Classifying Different Emotional States by Means of EEG-Based Functional Connectivity Patterns”. *PLoS ONE* 9 (2014). DOI: 10.1371/journal.pone.0095415/e95415.

[123] Tommaso Costa, Elena Rognoni, and Dario Galati. “EEG phase synchronization during emotional response to positive and negative film stimuli”. *Neuroscience Letters* 406.3 (2006), pp. 159–164. DOI: 10.1016/j.neulet.2006.06.039.

[124] Qibin Zhao, Cesar F Caiafa, Danilo P Mandic, Zenas C Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, and Andrzej Cichocki. “Higher Order Partial Least Squares (HOPLS): A Generalized Multilinear Regression Method”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.7 (2012), pp. 1660–1673. DOI: 10.1109/TPAMI.2012.254.

[125] Guillaume Rabusseau and Hachem Kadri. “Low-Rank Regression with Tensor Responses”. *Proceedings of the Advances in Neural Information Processing Systems* (2016), pp. 1867–1875.

[126] Hua Zhou, Lexin Li, and Hongtu Zhu. “Tensor regression with applications in neuroimaging data analysis”. *Journal of the American Statistical Association* 108.502 (2013), pp. 540–552. DOI: 10.1080/01621459.2013.776499.

[127] Margaret M Bradley and Peter J Lang. “Measuring emotion: The self-assessment manikin and the semantic differential”. *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (1994), pp. 49–59. DOI: 10.1016/0005-7916(94)90063-9.

[128] “American electroencephalographic society guidelines for standard electrode position nomenclature”. *Journal of Clinical Neurophysiology* 8.2 (1991), pp. 200–202.

[129] Svante Wold, Michael Sjöström, and Lennart Eriksson. “PLS-regression:a basic tool of chemometrics”. *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[130] Qibin Zhao, Liqing Zhang, and Andrzej Cichocki. “Multilinear and and non-linear generalizations of partial least squares: an overview of recent advances”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.2 (2014), pp. 104–115.

[131] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning, Second Edition*. Springer, 2009.

[132] Rishabh Gupta, Khalil ur Rehman Laghari, and Tiago H. Falk. “Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization”. *Neurocomputing* 174 Part B.22 (2016), pp. 875–884. DOI: 10.1016/j.neucom.2015.09.085.

[133] Ramesh Srinivasan, William R Winter, Jian Ding, and Paul L Nunez. “EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics”. *Journal of Neuroscience Methods* 166.1 (2007), pp. 41–52. DOI: 10.1016/j.jneumeth.2007.06.026.

[134] Louis A. Schmidt and Laurel J. Trainor. “Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions”. *Cognition and Emotion* 15.4 (2001), pp. 487–500. DOI: 10.1080/02699930126048.

[135] James J. Gross and Oliver P. John. “Individual differences in two emotion regulation processes: implications for affect relationships, and well-being”. *Journal of Personality and Social Psychology* 85.2 (2003), pp. 348–362. DOI: 10.1037/0022-3514.85.2.348.

[136] Michela Balconi and Guido Mazza. “Lateralisation effect in comprehension of emotional facial expression: a comparison between EEG alpha band power and behavioural inhibition (BIS) and activation (BAS) systems”. *Laterality: Asymmetries of Body, Brain and Cognition* 15.3 (2010), pp. 361–384. DOI: 10.1080/13576500902886056.

[137] Rone J. Huster, Stephan Stevens, Alexander L. Gerlach, and Fred Rist. “A spectralanalytic approach to emotional responses evoked through picture presentation”. *International Journal of Psychophysiology* 72.2 (2009), pp. 212–216. DOI: 10.1016/j.ijpsycho.2008.12.009.

[138] Eckart Altenmüller, Kristian Schürmann, Vanessa K Lim, and Dietrich Parlitz. "Hits to the left, flops to the right: different emotions during listening to music are reflected in cortical lateralisation patterns". *Neuropsychologia* 40.13 (2002), pp. 2242–2256. DOI: 10.1016/S0028-3932(02)00107-0.

[139] Nancy Aaron Jones and Nathan A Fox. "Electroencephalogram asymmetry during emotionally evocative films and its relation to positive and negative affectivity". *Brain and Cognition* 20.2 (1992), pp. 280–299. DOI: 10.1016/0278-2626(92)90021-D.

[140] K Luan Phan, Stephan F Taylor, Robert C Welsh, Laura R Decker, Douglas C Noll, Thomas E Nichols, Jennifer C Britton, and Israel Liberzon. "Activation of the medial prefrontal cortex and extended amygdala by individual ratings of emotional arousal: a fMRI study". *Biological Psychiatry* 53.3 (2003), pp. 211–215. DOI: 10.1016/S0006-3223(02)01485-3.

[141] Elise S. Dan Glauser and Klaus R. Scherer. "Neuronal Processes involved in Subjective Feeling Emergence: Oscillatory Activity During an Emotional Monitoring Task". *Brain Topography* 20.4 (2008), pp. 224–231. DOI: 10.1007/s10548-008-0048-3.

[142] Hiroshi Higashi, Tomasz M Rutkowski, Toshihisa Tanaka, and Yuichi Tanaka. "Multilinear discriminant analysis with subspace constraints for single-trial classification of event-related potentials". *IEEE Journal of Selected Topics in Signal Processing* 10.7 (2016), pp. 1295–1305. DOI: 10.1109/JSTSP.2016.2599297.

[143] Ed Bullmore and Olaf Sporns. "Complex brain networks: graph theoretical analysis of structural and functional systems". *Nature Reviews Neuroscience* 10.3 (2009), p. 186. DOI: 10.1038/nrn2575.

[144] Mahmoud Hassan, Olivier Dufor, Isabelle Merlet, Claude Berrou, and Fabrice Wendling. "EEG source connectivity analysis: from dense array recordings to brain networks". *PloS One* 9.8 (2014), e105041. DOI: 10.1371/journal.pone.0105041.

[145] Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. "Learning Laplacian matrix in smooth graph signal representations". *IEEE Transactions on Signal Processing* 64.23 (2016), pp. 6160–6173. DOI: 10.1109/TSP.2016.2602809.

[146] Liu Rui, Hossein Nejati, Seyed Hamid Safavi, and Ngai-Man Cheung. "Simultaneous low-rank component and graph estimation for high-dimensional graph signals: Application to brain imaging". *IEEE International Conference on Acoustics, Speech and Signal Processing* (2017), pp. 4134–4138. DOI: 10.1109/ICASSP.2017.7952934.

[147] Ivan V Oseledets. "Tensor-train decomposition". *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2295–2317. DOI: 10.1137/090752286.

[148] Daichi Kitamura and Nobutaka Ono. "Efficient initialization for nonnegative matrix factorization based on nonnegative independent component analysis". *IEEE International Workshop on Acoustic Signal Enhancement* (2016), pp. 1–5.

[149] Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. "Estimation of low-rank tensors via convex optimization". *arXiv preprint:1010.0789* (2010).

[150] Marco Signoretto, Quoc Tran Dinh, Lieven De Lathauwer, and Johan AK Suykens. "Learning with tensors: a framework based on convex optimization and spectral regularization". *Machine Learning* 94.3 (2014), pp. 303–351. DOI: 10.1007/s10994-013-5366-3.

[151] David C. Rubin and Jennifer M. Talarico. "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words". *Memory* 17.8 (2009), pp. 802–808. DOI: 10.1080/09658210903130764.

[152] Rishabh Gupta and Tiago H. Falk. "Latent factor analysis for synthesized speech quality-of-experience assessment". *Quality and User Experience* 2.2 (2017). DOI: 10.1007/s41233-017-0005-6.

[153] Florian Hinterleitner, Georgina Neitzel, Sebastian Möller, and Christoph Norrenbrock. "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks". *Procedings of the Blizzard Challenge Workshop* (2011).

[154] Florian Hinterleitner, Christoph Norrenbrock, Sebastian Möller, and Ulrich Heute. "What makes this voice sound so bad? A multidimensional analysis of state-of-the-art text-to-speech systems". *Proceedings of the IEEE Spoken Language Technology Workshop* (2012), pp. 240–245. DOI: 10.1109/SLT.2012.6424229.

Publication List

Reviewed Internatinal Jounals

Related to Chapter 4

1. Hayato Maki, Tomoki Toda, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. “Enhancing Event-Related Potentials Based on Maximum a Posteriori Estimation with a Spatial Correlation Prior”, *IEICE TRANSACTIONS on Information and Systems*, Volume E99-D, Number 6, pages 1437-1446, 2016, DOI: 10.1587/transinf.2015CBP0008

Related to Chapter 5

2. Hayato Maki, Sakriani Sakti, Hiroki Tanaka, and Satoshi Nakamura. “Quality Prediction of Synthesized Speech Based on Tensor Structured EEG Signals”, *Plos One*, 10.1371/journal.pone.0193521

Reviewed Internatinal Conferences

Related to Chapter 4

1. Hayato Maki, Tomoki Toda, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. “EEG signal enhancement using multi-channel Wiener filter with a spatial correlation prior,” *in Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP’15), Brisbane, April, 2015, pages 2639-2643, DOI: 10.1109/ICASSP.2015.7178449

2. Hayato Maki, Tomoki Toda, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. “An evaluation of EEG ocular artifact removal with a multi-channel wiener filter based on probabilistic generative model,” *in Proceedings of the 37th International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC’15), Milan, August, 2015, pages 2775-2778,

   DOI: 10.1109/EMBC.2015.7318967

3. Hayato Maki, Tomoki Toda, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. “Removing noise from event-related potentials using a probabilistic generative model with grouped covariance matrices,” *in Proceedings of the 38th International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC’16), Orland, August, 2016, pages 3728–3731,

   DOI: 10.1109/EMBC.2016.7591538

Related to Chapter 3

4. Hayato Maki, Hiroki Tanaka, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. “Graph Regularized Tensor Factorization for Single-Trial EEG Analysis,” *in Proceedings of the 43th IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP’18), Calgary, April, 2018, to appear