Improving Group Decision Making with Collaborative Brain-Computer Interfaces

Davide Valeriani

A thesis submitted for the degree of

Doctor of Philosophy

School of Computer Science and Electronic Engineering

University of Essex

February 2017

To all those who believe in
education as a means to achieve
freedom and peace.

Abstract

Groups are generally superior to individuals in making decisions. However, time constraints and authoritarian leaders could nullify the potential advantages provided by groups.

This thesis proposes a hybrid collaborative Brain-Computer Interface (cBCI) for improving performance in group decision-making. Neural signals recorded via electroencephalography are integrated with other physiological and behavioural measures to predict the likelihood of the user being correct in a decision, i.e., decision confidence. Behavioural responses from multiple users are then weighed according to these confidence estimates to obtain group decisions.

The proposed cBCI has been tested with a variety of decision-making tasks, including visual matching, visual search with traditional and realistic stimuli, face recognition from multiple viewpoints, and speech perception. Groups assisted by the cBCI were significantly superior in making decisions than both individuals and traditional equally-sized groups making decisions using the majority method.

This thesis also investigates the impact that a constrained form of communication has on individual and group performance in a visual-search experiment. When decision makers are able to exchange in-

formation during the experiment, their performance dramatically decreases. However, the cBCI yields superior group decisions even in this context.

The confidence estimated by the cBCI is also a more reliable predictor of correctness than the confidence reported by participants after making a decision. When group members were allowed to communicate during visual search, their reported confidence was totally unrelated to the decision correctness, while in a speech perception task reported confidences were very good predictors of correctness. On the contrary, the cBCI's confidence estimates correlated with correctness in all experiments.

When critical decisions involving substantial risks have to be made (e.g., in defence), the proposed cBCI could be a useful tool to reduce the number of erroneous group decisions, thereby saving money and lives.

Acknowledgements

Many people have accompanied me in the long journey of submitting this thesis. I could not imagine arriving here without them.

It is not easy to find the proper words to express the gratitude that my supervisors Riccardo Poli and Caterina Cinel deserve. Their continuous guidance and support, from my arrival at Essex till the final stages of my PhD, even during stressful periods, have made this journey possible. Their expertise and enthusiasm during the long brainstorming sessions have stimulated my interest in research. Finally, their friendly supervision and trust have made this adventure really enjoyable.

Special thanks go to Ana, the best mentor one could ever desire, especially for the support with the English and her constant presence in the office (with coffee and cakes). Thanks for being a loyal companion in all the projects and activities we have done together in these years, from organising conferences to starting a business. There are definitely people in our department still believing we are siblings.

A big thank you to all the other friends and colleagues who have shared part of this journey with me, most especially Diego, Spyros, Florian, Miguel, Louis and Javi. Thanks to the Brainstormers, in particular to David and Hilary Rose, for their hard work and deter-

mination that allowed us to get a bronze medal in the Cybathlon 2016. Thanks to the BCI-NE group and my department for creating a stimulating environment where doing research, especially to Francisco Sepulveda, for his invaluable suggestions during the supervisory board meetings, and Luca Citi, for the support in machine learning and the opportunity of continuing doing research in this amazing group. Also, thanks to Demba Ba for having hosted me in the CRISP lab at Harvard University while working on state-space modelling.

This research would not have been possible without the funding provided by DSTL via its Defence and Security National PhD programme. Special thanks to our technical partner Colin Corbridge for his enthusiasm and invaluable feedback throughout the years. This thesis is also a tribute to his memory. Thanks also to Annalise Whittaker, for her feedback on this thesis and her support during the years, and to Mike Potter, for recording the stimuli used in one of my experiments.

Last, but not least, thank you to my family for always standing by my side, even after deciding to leave my country to follow my dreams. Despite the distance, I still feel your support everyday in whatever I do. And finally, thanks to Elena, for her support during the writing of this thesis and for having added sweetness to this adventure.

Contents

List of Figures

Chapter 1

Introduction

This chapter introduces the motivation of this thesis, summarises its main contributions and research questions addressed, and describes its organisation. A list of papers published during this research is also provided.

1.1 Motivation

Decision making has been studied for decades by a broad range of disciplines for its direct impact on everyday life. Cognitive neuroscientists have been trying to decipher what exactly is happening in our mind when we make decisions, while social scientists have been investigating which external factors influence our decisions and how. One of the objectives of studying decision making is to understand what leads human to make incorrect choices, in order to find strategies to reduce the number of erroneous decisions, as their consequences could be very dramatic in certain contexts. For example, in finance, where deciding to buy/sell the wrong stock can cause significant loss of money, or in medicine, where a wrong

therapy prescribed to a patient could cause serious issues, or in defence, where not identifying a threat in pictures taken from a security camera could cause loss of human lives.

Frequently, making the correct decision depends on several factors, including the level of knowledge of the person and the time available. Moreover, the human brain has some capacity limitations that restrict our ability of processing information and perceive properly [108]. These flaws of the conscious perception could make people decide on the basis of incorrect information gathered from the senses, leading to suboptimal decisions.

Research on decision making has shown that a solution to partially solve individual misjudgement is making decisions in groups. Groups have augmented capabilities and intelligence that are the result of integrating different views and percepts through the interaction of their members [181]. For these reasons, group decisions are usually more accurate than those made by individuals [6]. This is why organisations such as universities are run by boards and panels, and why democratic institutions such as the parliaments are organised in committees and assemblies.

However, there are circumstances in which involving other people in a decision could be deleterious [12]. For example, having strict time constraints or in the presence of leaders can nullify most of the advantages provided by groups. Moreover, the traditional approach to group decision making includes communication and discussion between the group's members, which could reduce or even nullify the contribution of some people (e.g., people who are naturally shy) to the group decision, as well as slowing down the decision process. In contexts where decisions have to be taken rapidly, group discussion is not possible and its absence could

lead to suboptimal decisions [6].

Brain-Computer Interfaces (BCIs) are devices that convert brain signals into commands that can be used to operate external devices, such as a prosthetic arm. BCIs have traditionally been used to provide an alternative communication channel to people with disabilities, allowing them to act on the world. In recent years, the promising results obtained by BCIs have pushed researchers to apply these technologies to other fields, such as human augmentation, hence increasing the number of potential BCI end-users. One of these promising new areas of applications of BCIs is decision making. Research has shown that it is possible to decode the choice of the user from his/her brain signals, allowing to develop BCIs that can accelerate decisions in tasks such as the classification of images [9] or the detection and localisation of planes in aerial images [111]. However, EEG signals are noisy and require the averaging of multiple recordings over time to be able to provide reasonable performance, which, in turns, reduces the responsiveness of the system. This trade-off between performance and speed makes single-user BCIs difficult to be applied in contexts like critical decision making, where an error caused by the system not being able to correctly detect the intentions of the user could have serious consequences.

With the aim of improving BCI performance without reducing speed, researchers have started investigating the possibility of aggregating brain signals from multiple users as an alternative approach for reducing the noise that affects neural recordings. When compared to single-user BCIs, these collaborative BCIs (cBCIs) have been able to significantly boost performance. For example, when applied to decision making, cBCIs make better and faster decisions than individuals [212]. However, these systems have only been applied to a very limited

number of simple tasks. Moreover, when critical decisions are involved, reducing the number of erroneous decisions is usually more important than making faster (but less accurate) decisions.

This thesis explores the possibility of using a *hybrid* cBCI to support and augment group decision making in a variety of critical, difficult target-detection tasks, involving visual or auditory stimuli. The cBCI uses a hybrid approach as it combines behavioural responses, acquired via traditional means (i.e., mouse clicks), and decision confidence, estimated using the brain signals and other physiological and behavioural measures. This approach allows cBCI-assisted groups to perform better not only than individuals, but also than equally-sized groups making decisions using the majority rule, in contrast with traditional cBCIs based only on neural signals which required up to seven participants to perform better than individuals [36].

Group decisions could also be obtained using confidence estimates reported by the observers themselves after each decision. This thesis shows how these subjective estimates may be unreliable as their reliability is highly influenced by the task at hand, the participants, and other external factors, such as the interaction between group members. In some circumstances, group decisions made using these confidence estimates are even worse than those made using the simple majority. Conversely, the hybrid cBCI is able to provide a consistent advantage for groups over majority across tasks.

Finally, this thesis investigates which are the best conditions for groups to make decisions. These factors include (a) the presence or absence of communication between group's members, (b) the exposure of observers within a group to the same or different sources of information, and (c) the modality of stimulating

the decision makers (e.g., visual or audio).

1.2 Contributions

The main scientific contributions of this thesis are:

1. A hybrid cBCI framework to enhance group decision making (Chapter 3). The framework uses a combination of physiological and behavioural measures to estimate the confidence level of each decision maker, which represents the likelihood of the user making a correct decision. Individual decisions acquired with traditional means (e.g., mouse clicks) are then integrated together according to these confidence estimates to obtain group decisions. Since the hybrid cBCI is based on individual decisions, it does not require the extra time generally needed by traditional groups to discuss and agree on a decision, hence making cBCI decisions faster than traditional groups ones.

2. The identification of the best set of physiological and behavioural correlates of decision confidence amongst a number of indicators analysed. Previous cBCIs were focused on predicting the intentions of the users (decisions) rather than their validity (decision confidence). Therefore, more research was needed in order to identify confidence correlates. We analysed (a) brain signals recorded via electroencephalography, (b) eye movements and eye blinks, and (c) response times (RTs). The experimental work (Chapters 4, 5, 6, 8 and 9) shows that a few neural features and RTs provide most of the information available on decision confidence in all experiments, while

eye features seemed to be informative only in tasks using visual stimuli.

3. An evaluation of the performance obtained using the proposed hybrid cBCI in a variety of decision-making tasks of increasing realism involving uncertainty. These tasks include (a) visual matching (Chapter 4), (b) visual search with traditional stimuli (Chapter 5), (c) visual search with realistic stimuli (Chapters 5 and 6), (d) face recognition from realistic pictures recorded from multiple security cameras (Chapter 8), and (e) speech perception with real radio communication messages affected by noise (Chapter 9). A total of 76 participants have taken part in the seven experiments described in this thesis, hence providing evidence of the superiority of the proposed approach for group decision making.

4. A comparison between the decision confidence estimated by the cBCI using physiological and behavioural measures and the confidence reported by the participants after making a decision. The results (Chapters 6, 8 and 9) show that the cBCI is able to provide an estimate that correlates with decision correctness in all experiments, while the confidence reported by the participants is generally less reliable, working well in some cases and really badly in others.

5. An investigation on the impact that a constrained form of communication has on individual and group performance (Chapter 6), both when groups are assisted by the hybrid cBCI or when they are not. In the visual search experiment with naturalistic stimuli, participants were paired while undertaking the same decision tasks. After individual decisions, they were given feedback about the decision and confidence level of the other member. Results

show that this constrained communication negatively affects the individual (and, therefore, the group) performance when compared to experiments where participants undertake the task in isolation. The communication had also a negative impact on the correlation between the confidence reported by the users and the correctness in the decision.

6. A study on how the exposure of different observers to different sources of information (e.g., pictures of the same scene taken from various viewpoints) affects the performance of non-BCI and BCI-assisted groups. Previous studies have shown that traditional groups are effective when individual opinions are not correlated [181], which is more likely to happen when each participant is exposed to different sources of information. However, little was known about the effects of this multi-viewpoint approach on cBCI performance. Chapter 8 analyses the performance of groups undertaking a face recognition task where group's members were exposed to images of the same scene taken from three different viewpoints. Results show that the multi-viewpoint groups are superior to groups where members are exposed to the same stimuli.

1.3 Research Questions

This thesis addresses the following research questions:

Q1. Can group decision making based on neural, physiological and behavioural features achieve better levels of accuracy than traditional majority voting across a range of tasks?

Q2. What is the best set of physiological and behavioural features acting as confidence indicators?

Q3. What are the neural features that are the most relevant for the proposed hybrid cBCI for group decision making?

Q4. Is the confidence estimate provided by the cBCI more reliable than a confidence reported by the user?

Q5. Can collaborative BCIs lead to faster decisions than average human reaction times?

Q6. Are there optimal scenarios for which BCI group decision making is most suited?

Q7. What is the impact of group interaction on cBCI performance?

Q8. In what ways does the exposure of different observers to various sources of information modify optimal group sizes, accuracy, and speed of decisions?

1.4 Structure

The concepts of BCI and cBCIs are introduced in Chapter 2, which also reviews relevant literature related to decision making and neural signal processing.

Chapter 3 describes the hybrid cBCI framework that will be used in most other parts of the thesis to improve group decisions. This chapter also discusses which features a decision-making experiment should have to be suitable for the proposed cBCI. Moreover, it provides an overview on how the physiological signals

are recorded, processed and used to estimate the decision confidence and obtain group decisions.

The proposed framework has firstly been applied to a simple visual matching task, described in Chapter 4, where the experimental part of this thesis starts. The results obtained with 10 participants are presented and discussed, showing how, for the first time, the proposed cBCI was able to beat not only non-BCI users but also equally-sized non-BCI groups.

Chapter 5 analyses the performance of the hybrid cBCI in two visual search experiments, one using standard stimuli (i.e., coloured bars) and one using realistic ones (i.e., pictures of Arctic environments). This chapter also describes the performance obtained by the cBCI when using a more advanced technique for extracting neural correlates of decision confidence.

The analyses of performance in visual search continues in Chapter 6, where the impact of a constrained form of communication between pairs is studied. This chapter also discusses whether or not a decision confidence reported by participants would be more accurate than the confidence obtained by the cBCI.

Chapter 7 explores the possibility of using state-space models to estimate the cognitive state of the decision maker by means of behavioural and physiological measures. This model could then be used by the hybrid cBCI to temporarily exclude from the group individuals that are tired or not focused, hence improving group performance.

In an attempt to make another step towards applying the proposed cBCI to real decision-making problems, Chapter 8 describes the performance of cBCI-assisted groups carrying out a face recognition task using pictures gathered from three surveillance cameras. This chapter also discusses the variations on perfor-

mance when participants are exposed to different sources of information.

While the previous chapters were focused on tasks based on visual stimuli, Chapter 9 analyses the performance of groups undertaking a speech recognition task using auditory stimuli. Here, the cBCI used only a small subset of the electrodes to estimate the decision confidence, hence promoting generalisation and practicality of the system. However, in this experiment participants seemed to be very good in estimating the confidence themselves, therefore making the cBCI not needed. The chapter discusses the risks of using the reported confidence for obtaining group decisions and analyses the limitations of the cBCI in that particular task.

The thesis ends with Chapter 10, where the major achievements of this doctoral work are summarised and ideas for future work are discussed.

1.5 List of Publications

This thesis is partially based on the papers listed in the following subsections. The chapters based on each paper are indicated in bold face.

1.5.1 Peer-Reviewed Journal Papers

- Davide Valeriani, Riccardo Poli and Caterina Cinel. Enhancement of Group Perception via a Collaborative Brain-Computer Interface. *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 6, August 2016. **Chapters 5 and 6**

- Riccardo Poli, Davide Valeriani and Caterina Cinel. Collaborative brain-

computer interface for aiding decision-making. *PLOS ONE*, vol. 9, no. 7, July 2014. **Chapters 3 and 4**

1.5.2 Peer-Reviewed Conference Papers

- Davide Valeriani, Caterina Cinel and Riccardo Poli. Augmenting Group Performance in Target-Face Recognition via Collaborative Brain-Computer Interfaces for Surveillance Applications. 8^{th} *International IEEE EMBS Conference on Neural Engineering*, May 2017. **Chapter 8**

- Davide Valeriani, Caterina Cinel and Riccardo Poli. Hybrid Collaborative Brain-Computer Interfaces to Augment Group Decision Making. 1^{st} *International Conference on Neuroergonomics*, October 2016. **Chapter 9**

- Davide Valeriani, Caterina Cinel and Riccardo Poli. Improving Speech Perception with Collaborative Brain-Computer Interfaces. 38^{th} *Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, August 2016. **Chapter 9**

- Davide Valeriani, Riccardo Poli and Caterina Cinel. A Collaborative Brain-Computer Interface to Improve Human Performance in a Visual Search Task. *Proceedings of the* 7^{th} *International IEEE EMBS Neural Engineering Conference*, pp. 218-223, April 2015. **Chapter 5**

- Davide Valeriani, Riccardo Poli and Caterina Cinel. A Collaborative Brain-Computer Interface for Improving Group Detection of Visual Targets in Complex Natural Environments. *Proceedings of the* 7^{th} *International IEEE EMBS Neural Engineering Conference*, pp. 25–28, April 2015. **Chapter 5**

1.5.3 Book Chapters

- Davide Valeriani and Ana Matran Fernandez. Past and Future of Multi-Mind Brain-Computer Interfaces. *Brain-Computer Interfaces Handbook: Technological and Theoretical Advances*, C. S. Nam, A. Nijholt and F. Lotte, Eds. CRC Press. 2017 (in press). **Chapter 2**

1.5.4 Other Publications

The following papers were published during the writing of this thesis, although they do not directly contribute to it:

- Ana Matran-Fernandez, Davide Valeriani and Riccardo Poli. Toward BCIs Out of the Lab: Impact of Motion Artifacts on Brain-Computer Interface Performance. *Wireless Medical Systems and Algorithms*, P. Salvo and M. Hernandez-Silveira, Eds. CRC Press, pp. 219-240, 2016.

- Davide Valeriani, Ana Matran Fernandez, Diego Perez Liebana, Javier Asensio Cubero, Christian O'Connell and Andrei Iacob. A Comparison of Ensemble Methods for Motor Imagery Brain-Computer Interfaces. *Proceedings of the European Conference on Data Analysis*, 2015.

- Davide Valeriani and Ana Matran Fernandez. Towards a Wearable Device for Controlling a Smartphone with Eye Winks. *Proceedings of the* 7^{th} *Computer Science and Electronic Engineering Conference (CEEC15)*, pp. 41-46, 2015.

Chapter 2

Background

This chapter presents an overview of the main literature published in the research areas related to this thesis, spanning from single and collaborative brain-computer interfaces to biomedical signal processing and group decision making. The main elements required in a collaborative brain-computer interface, such as the signal acquisition and the methods for data processing, are also introduced.

2.1 Neuroimaging Techniques

The human brain is one of the most powerful and complex machines in the world. Despite advances in research and technology, no computer is able to perform all the activities of the brain with the same accuracy. Its largest part, the cerebrum, is divided into four lobes, each of which is in charge of many different functions. Figure 2.1 summarises the main functions associated to each lobe [47].

The human brain is far from being perfect. Phenomena such as inattentional blindness [106] could lead to individuals failing to recognise unexpected stimuli

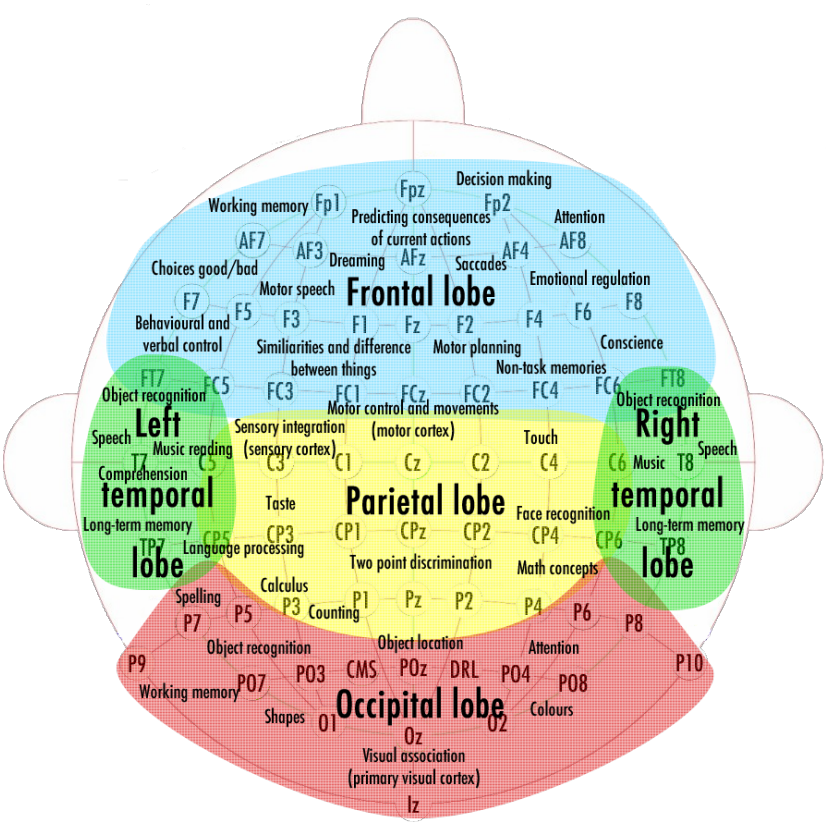

Figure 2.1: Summary of the main functions associated to each brain lobe.

that are in plain sight. Moreover, our brain has some intrinsic speed limitations, for example in visual processing [188].

Some of the limitations of the human brain could be overcome with the use of technology. Computers are incredibly fast and accurate in doing complex calculation or, in general, in performing tasks that can be translated into algorithms. For these reasons, for many years scientists have looked into the possibility of integrating brain and computers to enhance human capabilities. In order to do so, methods to observe the brain activity of a user and transform it in signals that are processable by a machine are required. This section presents an overview of the main *noninvasive* techniques for neuroimaging, most of which are adaptations

of corresponding technologies used in the medical sector. Particular attention is devoted to electroencephalography, the technique that will be used in this thesis.

2.1.1 Electroencephalography (EEG)

The human brain is composed of billions of neurons, cells that process and transmit information through electrical and chemical signals. A neuron transfers information by "firing", i.e., generating trains of pulses along its axon. The currents produced by this electrical activity are generally too small to be measured, but when many neurons fire at the same time, they become measurable via EEG [102].

To record this electrical activity, various electrodes are placed on the scalp of the user, usually following the 10-20 international system. Active electrodes are generally the most used. These include an additional pre-amplifier located inside the electrode to amplify the small signal before it gets contaminated by electromagnetic environmental noise, while passive electrodes only rely on the EEG system amplifier. Also, electrodes may be "wet" or "dry". The former require placing a small amount of electrically conductive gel between each electrode and the scalp to ensure good electrical contact, which extends preparation time. Dry electrodes are quicker to set up since they use different mechanical methods to ensure acceptable contact, but the quality of the signals recorded is generally inferior to that obtained with wet electrodes.

EEG is one of the cheapest and most portable techniques to measure neural activity, together with fNIRS (see Section 2.1.4). It also has an excellent temporal resolution (milliseconds) and is totally noninvasive and safe for the user. The main drawbacks of EEG are the low spatial resolution (mainly due to the skull and

skin between the electrodes and the brain, which are not perfect conductors [107]) and the poor signal-to-noise ratio, which require sophisticated data processing to extract useful information from the brain signals acquired. Also, EEG mainly records neural activity that occurs in the upper layers of the brain.

The information obtained from the EEG signals can be used to understand the brain activity [203], diagnose pathological conditions or for human augmentation. The application of this technology to humans dates back to 1929, when the German psychiatrist Hans Berger recorded the first human EEG [8]. Since then, EEG has been broadly used in neuroscience and its popularity has also pushed the development of commercial portable EEG devices [118] that can be bought and used by the end-user for different applications [59].

The low cost and non-invasiveness of EEG have made this method the most popular for data recording for human augmentation. For this reason, EEG is also the method adopted in this thesis for observing the brain activity of decision makers.

2.1.2 Magnetoencephalography (MEG)

The electrical activity produced by firing neurons generates magnetic fields. MEG is a technique that uses special sensors (SQUIDs, i.e., Superconducting Quantum Interference Devices) to detect the very tiny magnetic fields (a few fT in strength) generated by the neurons. This technology has been broadly used to determine the function of various parts of the brain, localise regions affected by a pathology, and other medical applications [63]. One of the main drawbacks of MEG is that it requires complex and expensive devices for signal acquisition, including a

magnetically-shielded room, making it not practical for most human enhancement applications.

2.1.3 Functional Magnetic Resonance Imaging (fMRI)

Neurons are active cells that require energy (sugar) and oxygen to perform their functions. fMRI is a noninvasive technique that measures brain activity by detecting changes in the blood flow (hemodynamic response). The primary form of fMRI uses the blood-oxygen-level dependent contrast to associate changes in blood flow to neural activity in the brain. When blood is rich in contrast, it produces a stronger electromagnetic response to the spin-altering waves emitted by the MRI scanner than when it is poor in contrast, making it possible for fMRI to measure differential brain activity.

Like MEG, fMRI does not require the contact with the body but it needs big and expensive devices for signal acquisition. For these reasons, it is generally unsuitable for applications in human augmentation [202].

2.1.4 Functional Near-Infrared Spectroscopy (fNIRS)

Similarly to fMRI, fNIRS uses hemodynamic responses to measure the brain activity. Instead of measuring chemical concentrations, fNIRS sends beams of near-infrared (NIR) light into the scalp and measures how much light is reflected back. The transmission and absorption of NIR light in human body tissues is related to changes of oxygen concentration.

NIR beams are sent via several probes placed on the scalp at different locations, making this technology more portable [163] and cheaper than fMRI, and

less susceptible to electrical noise than EEG. However, the quality of the signals recorded is quite poor due to low spatial and temporal resolution. For these reasons, its applications to human augmentation are still quite limited [120].

2.2 Event-Related Potentials

One of most interesting uses of EEG signals is the study of relationships between external events (e.g., the presentation of a stimulus) and the corresponding brain activity recorded, in order to understand how the brain reacts to a single event or a category of events. These brain responses to external events are named *event-related potentials* (ERPs).

External stimuli usually generate the activation of multiple areas of the brain and the corresponding elicitation of many ERPs. Literature has introduced the term *ERP component* to identify the scalp-recorded voltage change that reflects a specific psychological process. However, this assumption is an approximation. In fact, an ERP is generated by a neural activation that, usually, lasts for tens or hundreds of milliseconds. Therefore, as it happens frequently, when an ERP signal is generated (after a particular event) the tails of old ERPs are still present. This means that an overlap between different neural processes could happen, making the precise mapping between ERP components and specific psychological processes almost impossible [104].

ERPs are usually represented through their waveforms. An ERP waveform is a depiction of the changes over time in the scalp-recorded voltage that reflect the sensory, cognitive, affective, and motor processes elicited by a stimulus [104]. Multiple ERP components are generally represented in a waveform.

Recorded ERPs are generally affected by noise. The high impedance of the skull makes the electric signals travelling from the neurons to the electrodes spread laterally. Therefore, the EEG signal recorded at a particular location is the result of a weighted sum of ERP components and noise, where the weights depend on the distance between the sensor and the firing neurons. A metaphor often used to explain this phenomenon is that of a cocktail party, where several people are chatting together in small groups. If a person (EEG recorder) enters into the room and wants to understand what a particular person is saying (ERP component), he/she will hear sounds originated by a mix of the different conversations held in the room (waveform). However, if the person moves around the room, the sound changes because the contribution of each person to the mix changes.

Several signal processing techniques have been employed in the literature to reduce the noise of ERP recordings. The most used consists in *averaging* several ERP recordings belonging to many repetitions of the same stimulus [102]. By using enough repetitions, a robust EEG waveform describing how the brain reacts to a particular stimulus can be obtained.

The number of different ERP components reported in cognitive neuroscience and psychophysiology (some of which are used in BCI research) is quite high – see [162] for a review. It includes components associated to visual responses, such as C1 and P1, auditory responses, such as N1, and so on. The following section will describe a particular component called P300 which is the most used ERP in this thesis.

2.2.1 The P300 ERP

One of the main ERPs used in BCI is the P300 [43], a parietocentral positive peak occurring between 300 and 600 ms after the onset of a stimulus. This component is also known as P3 (the third positive peak after stimulus' onset) as the latency of its peak could vary between subjects [147, 74] and trials.

The P300 component is associated with the detection and recognition of interesting, rare, deviant or target stimuli [152, 66, 145]. Its amplitude can reach 40 μV, which is large for an ERP, making it easy to use in several BCI applications. The P300 ERP seems to correlate better with stimulus task relevance than with conscious perception [124].

Generally, the P300 component is employed in tasks where users have to discriminate between different stimuli [145]. These tasks usually follow the *oddball paradigm* [42], characterised by a number of low-probability "target" and high-probability "non-target" stimuli presented to the user. When a stimulus containing the target is shown, the brain of the user generates a P300 wave in response to this rare event.

P300-based BCIs, such as a speller [42] or a mouse [22], use a display where different locations are associated with different stimuli, each of which represents a "command" (e.g., a character to spell). The stimuli are flashed in turn (typically in random order) and the user is asked to focus on one of them (i.e., "target"). The P300 ERPs are generated only after the flashing of target stimuli and no other, making it possible for the BCI to determine which stimulus is being attended to, i.e., which command the user intends to issue. The process of focusing attention can be made easier by assigning a mental task, such as counting the flashes or

mentally naming the colour of the target stimulus [164]. P300-based BCIs have also been used to control external devices other than computers [32].

Some studies [178, 175] have proposed to split the P300 into two subcomponents: P3a and P3b. In a modification of the oddball paradigm using a third type of stimuli similar to the target ("distractors"), research suggested that the P3a subcomponent is generally associated to distractors, while the P3b (differing in latency from the P3a [102]) is the ERP associated to the target [23, 61, 145]. In a modification of the inattentional blindness paradigm [138], Pitts *et al.* found that, while ERP negativities could be elicited in presence of awareness, regardless the task relevance, the P3b component seems to be elicited only by task-relevant stimuli [139]. The P3a subcomponent is generated with both auditory and visual stimulus modalities [23].

Several researchers have shown that the P300 is also elicited in the process of decision-making [127, 160, 137], e.g., the brain process responsible to determine the presence or absence of a particular target in a stimulus and to map this decision to a particular response. For example, there seems to be a correlation between P300 amplitude and the uncertainty of a user in a decision [182]. This suggests possible BCI applications of the P300 other than those used for communication purposes.

P300 is not the only component used in BCI. A recent study [80] compared the reaction of participants in an oddball paradigm experiment by considering the components P300 and N200. They found that 30% of participants achieved better results using the N200 component instead of the P300. However, currently the P300 seems to be the most reliable and easy-to-use component in BCI. Recent advances have also allowed to further push the performance of BCIs based on this

component [191].

2.3 Decision Making

The process of decision making has been studied for centuries in several fields, such as psychology, political sciences and government. A particular focus has been group decision making, with several investigations about voting structures in democracies.

Several studies [6, 25, 82, 89, 83] have shown how group decisions can be superior compared to individual ones in many different contexts, including settings where individuals are involved in visuals tasks [177]. An earlier study by Barnlund [6] showed that the main reason why group decisions are superior is the discussion taking place within the group that leads people to be more cautious and focused on the task.

However, there are circumstances in which the discussion cannot take place properly and thereby group decision-making can be disadvantageous [75, 12]. For example, sometimes an agreed decision is difficult to be achieved because of lack of interaction between group members; also, a strong leadership can make the decision unfair for some members [82, 83, 177].

Another reason why groups seem to be superior to individuals in the decision-making process is that they can represent a larger set of perspectives and points of view. The decision made is the result of a process of mediation and discussions where members share information and get to know other members' opinions [190]. However, more communication and feedback is not necessarily better. A recent study [5], for example, has found that when there are time constraints or if lead-

ership prevails, the process of combining information from freely-communicating individuals can be an obstacle to optimal decision-making. Moreover, even when there is an advantage in the decision made by a group, the optimal group-size depends on the task at hand [90]. In other words, a decision made by a group of three people could be better than a decision made by an individual but also better than the decision made by a group of five people.

2.3.1 Neural Correlates of Decision Making

Neuroimaging techniques such as EEG can reveal important information about the different cognitive stages that lead to a decision. For example, the timing of the N1 – a large negative ERP occurring between 80 and 120 ms after the onset of an unpredictable stimulus in the absence of task demands – is sensitive to the difficulty of the task, while its amplitude decreases with the attentional level [105, 65]. The difficulty of a task also affects amplitude and timing of the P300 [61, 102]. For example, the differences in P300 responses have been used to make rapid decisions when determining whether a soldier is under fire from only auditory perception [170, 171].

While the aforementioned ERPs are typically associated with early perceptual and cognitive processing of events, other, later ERPs are instead associated with decision processes preceding, for example, the overt response of a decision maker. For instance, the contingent negative variation is a slow negative deflection related to the preparation for a motor response and stimulus anticipation. This ERP is smaller before incorrect responses than before correct ones in a task where information necessary to identify a target letter (e.g., its colour) is conveyed to

participants only a few hundred milliseconds before two potential targets are presented [131]. The error related negativity – an ERP occurring 50–80 ms after an incorrect response – is affected by confidence in own performance [168]. This happens even when participants are unaware of the error [128]. Moreover, neural correlates of individual decisions can be detected hundreds of milliseconds before an explicit response is given [192].

The observation of the brain activity during decision making does not provide only an insight on the choice itself, but on the decision-making process as a whole. This includes the estimation of the "decision confidence", our feeling about the validity of the response provided (metacognition) [57]. To this extent, several models have been proposed in the literature, including those using signal detection theory [237] and Bayesian inference [117]. All these models were based on the assumption that confidence estimates are built during the formation of our decision. More recent theories, however, have proposed that our sense of confidence is determined by brain processes occurring well after making a choice [228, 122, 220]. This sometimes leads participants to desire to reverse their initial choice [157], especially when their confidence is low [45]. Navajas *et al.* [125] used eye tracking to show that later stimuli are assigned greater confidence and that, therefore, confidence does not only measure the accumulated intensity of a stimulus [206], but varies reflecting an endogenous integration process. These studies suggest that in a behavioural experiment where participants report their choices, it is reasonable to observe the EEG activity both before and after the participants' responses.

Several studies have shown how the decision confidence estimated by participants is far from being perfect. In an ideal case, we would like this quantity to

reflect the probability our decision being correct ("metacognitive accuracy") [148], that is, having high values of confidence only when the decision is likely to be correct. However, humans are often miscalibrated [122]. For example, when the task is hard we tend to underestimate our confidence, while when the task is easy we usually overestimate it [96, 132]. Moreover, confidence estimates seem to be dependent on the stimulus features, including the motion direction in visual tasks [27], and on the amount of time between making a decision and giving the confidence estimate of that decision [122, 1].

2.4 Brain-Computer Interfaces

A brain-computer interface is a system that converts the brain activity (observed using one of the techniques described in Section 2.1) into commands for external devices or textual messages for communication [223]. They, therefore, allow users to affect the world without moving any muscle.

BCIs tend to be divided into two groups: (1) *continuous BCIs*, where the BCI transforms the user intentions into continuous outputs (i.e., real-valued quantities that can have many different values), and (2) *discrete BCIs*, where the BCI outputs categorical values. Examples from the first class are BCIs for cursor control [225, 136, 41, 224, 22, 210] or robotic control [18, 48, 67, 55]. Discrete BCIs include the P300 speller developed by Farwell and Donchin [42] and used in many other studies [10, 191], as well as BCIs for playing video games [200] and image classification [9, 111].

The typical structure of an EEG-based BCI system is depicted in Figure 2.2. It is composed by the following steps:

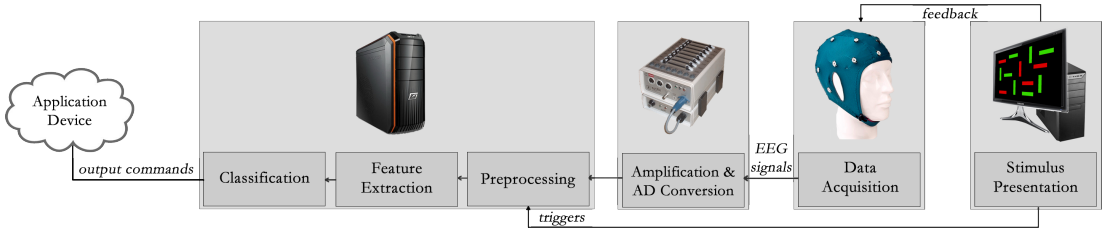

Figure 2.2: The main steps of an EEG-based BCI.

- *data acquisition*: usually performed with EEG [202] via electrodes mounted on a headcap;
- *amplification*: the small currents produced by neurons are amplified at this stage; an AD converter then converts the analogue signals to digital signals that could be interpreted by the computer;
- *preprocessing*: operations such as artefact removal, noise reduction and band-pass filtering are performed to improve the quality of the signals; also, the synchronisation of the signals with the occurrence of external events (such as visual stimuli) is performed;
- *feature extraction and selection*: signal processing and machine learning methods are used to isolate the components of the brain signals that carry the most information related to the task;
- *classification*: a classifier maps the set of features extracted from the brain signals at the previous step to a command/decision;
- *output*: the command produced at the previous step is sent to the external device, such as a wheelchair or a computer.

A BCI could be based on various paradigms. In this thesis, we will use ERP-based BCIs (see Section 2.2) as they have the advantage of requiring little training

from the user. Moreover, decision-making tasks usually include the presentation of a stimulus (i.e., evidence on which the user has to make the decision), which is a requirement for this type of BCIs. Another popular BCI paradigm is that based on mental tasks that the user has to perform to trigger the activation of the BCI. In this case, the BCI has to identify from the brain signals which cognitive task the user is performing (e.g., imagining the movement of a limb) and convert it to a specific output [34].

2.5 Collaborative BCI

The encouraging results obtained by BCIs have triggered the idea of using neural data from *multiple* brains to enhance BCI performance. The terms *collaborative BCIs* and *multi-mind BCIs* were introduced to identify systems that use the brain activity of at least two participants to perform a common task [197, 212]. Before that, the brain activity of multiple users participating in a common activity was analysed only for monitoring purposes with the *hyperscanning* technique [3]. The development of collaborative BCIs has also allowed to improve the accuracy of single-user BCIs, making it possible to use such systems as tools to enhance human performance for able-bodied users, as well as for people with disabilities.

Occasionally, the name "collaborative BCIs" has been associated to systems where the output depends on a combination of artificial intelligence and single-user BCIs [55, 79] and not on the brain signals of multiple users. We prefer to identify such systems with the term "shared-control BCIs", as there collaboration occurs between the computer and *one* user.

In the rest of this section, we review the main research conducted in the area

of collaborative BCIs [197].

2.5.1 Implementing a Collective Brain

Collaborative BCIs have been introduced back in 2010, when Wang and Jung [211] proposed a collaborative framework for BCIs to integrate brain signals recorded from multiple participants performing a movement planning task. The same authors also discussed the possible ways to implement a cBCI via fusing the brain activity of multiple users [212].

As explained in the previous section, a traditional single-user BCI is usually composed by a *signal acquisition* module, a *feature extraction* module, and a *decision* module. The brain activity of multiple users can thus be combined at four different levels: signal, feature, decision and application levels – see Figure 2.3.

Collaborative BCIs fusing brain recordings at the signal level have been studied to a significant extent. Generally, the brain signals of multiple users are averaged (an operation that also reduces the noise) and fed into a unique classifier directly, without extracting any feature [141, 112, 17, 16, 109, 72, 78, 86]. Some studies have also used the averaged brain signals to perform multi-user analyses [28, 110].

In a second scenario, features extracted from each user's EEG signals are merged. The fusion can be done by simple concatenation to form a unique feature vector or any other combination [212, 36], so that only one classifier is used to obtain the BCI output.

In the first two scenarios, the cBCI follows a "centralised" paradigm [212]: the neural data from multiple participants are collected by one machine and used

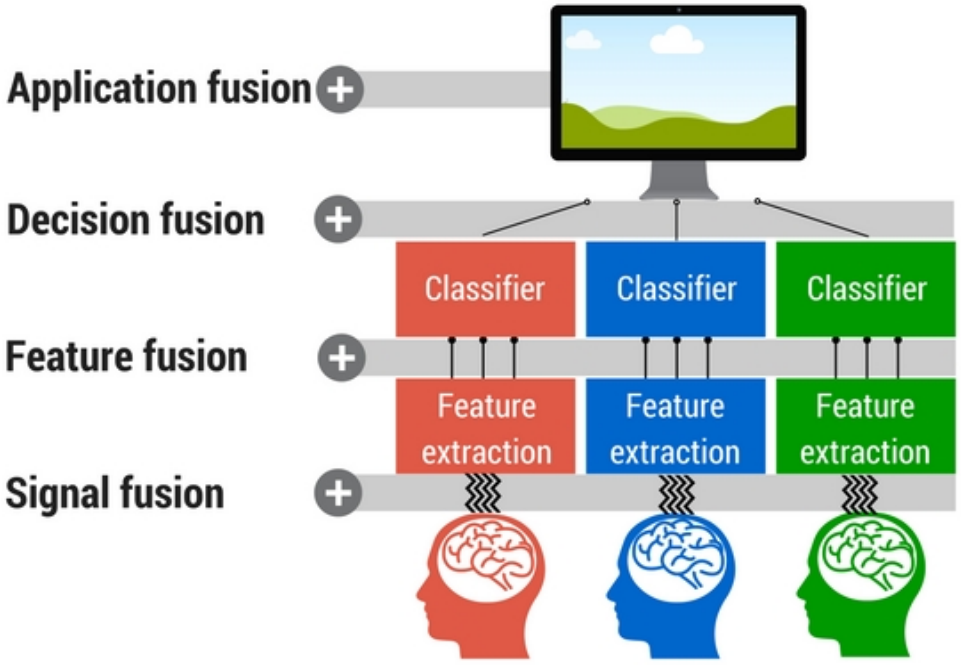

Figure 2.3: Main techniques to fuse neural signals from multiple users.

as inputs of a conventional BCI module. However, the large amount of data generated by several users and high computational costs for signal processing make this approach not suitable for many applications [212].

In the last two levels, the data acquisition, processing and classification steps are done on a participant-by-participant basis ("distributed" paradigm [212]). The outputs of these single-user BCIs are then aggregated by a separated module (e.g., another computer over a network). This approach is more efficient as it reduces significantly the amount of data travelling towards the central module.

Multi-mind BCIs at the decision level integrate the outputs of individually-tailored classifiers. At this level, we should emphasise the work by Cecotti and Rivet [17, 16], who studied different modes of combining the BCI decisions on a P300-based cBCI and a steady-state visual evoked potential multi-brain BCI.

Their strategies for merging the classifiers outputs included majority voting, average, maximum and minimum values. They found that averaging the classifiers' outputs provided the best performance. However, Eckstein *et al.* [36] found that the similarity in performance across observers affects the optimal strategy to integrate their decisions. Three different rules for integrating the classifier output of multiple observers discriminating pictures of cars and faces were compared: the optimal linear, the standard majority and the extreme opinion. The majority rule with its simplicity seemed to provide the best balance between performance and computational cost.

Finally, more recently an additional level of integration of brain signals called the "application level" has been proposed [11]. In this case, the implementation of the multi-mind BCI is not done by a module of the cBCI but by the application operated with the BCI, which receives the outputs of the single-BCIs and decide which one should be used to determine the collective choice. For example, if speed is a requirement of the system, the application could choose the fastest available output, assuming that faster responders are also more accurate. Other options would be to choose the most consistent brain activity or the strongest one [129].

2.5.2 Applications

Collaborative BCIs have been employed in a broad spectrum of applications, from traditional BCI ones (e.g., control and communication [223]) to group tasks (e.g., video games and decision making).

In communication, integrating brain signals from multiple participants allowed Cecotti and Rivet [17] to improve the offline performance of a P300 speller.

These results were then validated online by Kapeller *et al.* [78], who showed that the aggregation of EEG signals of eight participants allowed a cBCI to reach perfect performance on single-trial classification. Both studies recognised that communication is not a realistic application for cBCIs, as multiple users should agree on what to spell beforehand. However, these results showed the potential of cBCIs to enhance single-BCI performance.

Similarly to single-user BCIs, cBCIs have also been used to control external devices. In a simple movement-planning task, the cBCI developed in [212] yielded accuracies of up to 95% when predicting the direction of the movement (left vs. right) up to 250 ms before the actual motor response. Based on these encouraging results, Poli *et al.* [141] used cBCIs to perform complex control tasks. In that study, the neural signals from two participants were used jointly to control a spacecraft simulator through an analogue online cBCI. Other researchers developed SSVEP-based cBCIs that allowed pairs of participants with amyotrophic lateral sclerosis to operate a robot by sending target sequences of commands [94, 95].

The brain activity from multiple users could also be used in a *competitive* manner, especially for developing innovative video games [129]. In this scenario, brain signals are usually fused at the application level and the outputs of different single-user BCIs may be (a) used to control different avatars (e.g., cars) in a game [93, 4, 71], (b) compared to control a unique aspect of the interface according to the intentions of the "winner" [97] or (c) taken into account independently for shared control of a unique interface [91, 167, 11, 94]. While competition is at the basis of the majority of video games, cBCIs have also been applied to arcade games played in a collaborative manner. For example, in the BCI version of the popular video game Space Invaders developed in [86], the two users operating the

cBCI scored extra points if they were able to reduce the number of repetitions needed for successful selection of a target.

One of the most promising applications of multi-mind BCIs is probably decision making, as this is a task in which groups of users generally achieve superior performance than individuals [181]. Researchers have mostly applied cBCIs to target detection tasks, where groups of users have to decide whether a target object/person is present or not in a scene. A first attempt in this direction was made in [213], where participants were performing a detection task consisting in identifying a target stimulus. Users were asked to release a button when they saw the target (Go/NoGo task). The detection accuracy achieved by the cBCI integrating EEG signals from multiple participants was substantially superior than that obtained with single-user BCIs. Furthermore, the multi-mind BCI was able to accelerate the decision with respect to the motor action, as also shown in [212]. A following study [229] validated these results with an online cBCI with groups of six participants performing a discrimination task between faces and cars images following the Go/NoGo approach. In recent years, cBCIs have also been applied to more complex and challenging decision-making tasks, including face recognition [72], detection of visual targets in slow [230] and rapid [180, 112] presentation of images, and target localisation within images [109].

While the previous studies in decision making have shown the advantages of cBCIs with respect to single-user BCIs, one may wonder if such systems would also be more accurate than non-BCI users. Eckstein *et al.* [36] conducted a study in which they asked participants to discriminate between pictures of cars and faces. The performance achieved by individual observers was compared with that obtained using a cBCI merging brain signals at the decision level. While

the cBCI was faster than humans in making decisions, it required at least seven users to achieve the same accuracy of individual observers.

The results obtained in [36] triggered in researchers the idea of combining behavioural responses (which were more accurate) and BCIs (which were faster) to obtain superior group decisions. These hybrid cBCIs were firstly proposed in [142], where the neural signals of each decision maker were used to estimate his/her probability of having made the correct decision, a measure which was called "confidence". Group decisions were then obtained by weighing behavioural responses according to these confidence estimates. The preliminary results showed that this hybrid approach could provide the expected superior performance both in accuracy and speed.

2.6 Visual Search

One of the main decision-making tasks used in this thesis is visual search. It consists in a perceptual process involving visually-scanning the environment in search for an item of interest [222]. We perform visual search tasks on a daily basis, e.g., when looking for a particular item in a drawer containing many different objects or scanning our home for misplaced keys. Visual search, in the form of looking for a suspect or a potential terrorist within a crowd or in surveillance video, is also a key element of policing and counter intelligence. Despite there being clear evolutionary advantages in animals quickly identifying dangerous elements in the environment, humans invariably find visual search tasks slow, taxing and difficult to carry out (although performance varies across different people, contexts and details of the task performed, as well as with the experience and age of the

observer [62]).

Given the important role of visual search, it is not surprising that experimental visual-search paradigms have been extensively used in the study of perception and visual attention for more than 30 years [221, 35, 201]. These studies have shown that attentional mechanisms are vital to succeed in this task, both when single or multiple targets are present [15].

In a typical experiment, observers are asked to look at a display containing a number of different items and establish whether or not a particular object of interest (i.e., "target") is present in the scene. To make the task harder, the items in the scene which are not the target (i.e., "distractors") share some common features with it (e.g., shape, colour).

Visual search experiments usually follow two main approaches [221]. On the one hand, in the *percent correct* approach participants are presented a stimulus for a short period of time and have to decide whether or not the target is present. In this method, the aim of the participant is to *maximise the number of correct answers* in a difficult situation where too little information is available. On the other hand, in the *speed based* approach the stimulus is presented to participants until they provide a response (although, many studies introduce a timeout after which the experiment moves on even without a response from the participant recording an invalid decision). In this approach, the aim of the participant is to *minimise the response time* (RT) to give a correct answer.

The design of visual search experiments generally requires to set various parameters, including (a) the number of targets and distractors in a trial, (b) which features characterise the target (e.g., shape, colour, size, orientation, motion, etc.), (c) how many features the distractors share with the target, (d) the target

ratio (i.e. the probability of a target being present in a trial), (e) the duration of the stimuli, and (f) the timeout for acquiring a response. The choice of these values is strictly related to the difficulty of the task [2] and to the brain patterns that could be detected [103].

Various studies on visual search have shown that when the feature identifying the target is the colour, the differences in both the brain activity [103] and the response times [54] recorded in target and non-target trials are bigger than if the feature is the motion, the size or the orientation. This is because the attention of the participant is more focused on colour than on orientation and motion [54].

2.6.1 Face Recognition

In security and surveillance, a particularly interesting application of visual search is to identify an individual, usually via a process of "face recognition". Humans are generally extremely good and fast in recognising faces [92], even if they have seen the target person only once or in the presence of different facial expressions or lighting conditions. Our brain has a complex network of regions dedicated to process face information, the fusiform face area being its computational hub [52]. Due to the complexity of this task, the human brain splits face recognition in multiple stages, including pre-attentive processing, template fitting, and template evaluation [92].

In the last decades much effort has been spent in the development of algorithms to automatically identify a target person from a digital image or a video stream, achieving very good performance in controlled conditions [236, 207]. Recent results have made automated systems trained on large datasets comparable

or even superior than humans [184, 166]. However, in dynamic environments (e.g., with changes of lighting) [236] or when only a very limited number of training examples of the target face are available [185], the performance of automatic face recognition systems deteriorates significantly.

When we see a face, our brain reacts with specific ERPs, starting with the N170, which peak latency occurs between 130 and 200 ms after the face stimulus onset [121]. The N170 represents the most reliable difference in the brain activity on the scalp between faces and non-face objects [161]. A few milliseconds later, familiar faces elicit the N250 ERP [186]. Thus, EEG activity could be used to reveal how we judge people, for example in political elections [204].

The generation of specific brain patterns in the presence of a target face has made possible to further improve the accuracy of BCIs for control and communication [19]. Moreover, it has allowed the development of specific BCIs to accelerate and augment human performance in face recognition. A combination of different ERPs, including N170 and P300, was used in [234] to achieve an average accuracy of 88% in recognising inverted faces. A BCI to discriminate between familiar and unknown faces was developed in [87]. Other studies adopted the rapid serial visual presentation (RSVP) protocol to develop BCI systems able to recognise target faces amongst images of celebrities and relying on the brain activity of single [14] or multiple [72] users. Shared-control systems based on both computer vision and BCIs have also been proposed to further improve performance [214].

Chapter 3

A Hybrid Framework for Aiding Decision-Making

This chapter describes the architecture of a collaborative BCI (cBCI) for group decision making, from the data acquisition to the validation of the results. This cBCI will be used in the following chapters to make group decisions in multiple environments, from visual matching to speech perception. Most of the material in this chapter has been published in the paper [143].

3.1 Introduction

As presented in the previous chapter, cBCIs have shown the potential to overcome the traditional limitations of single-user BCIs, including low information transfer rate (ITR) and accuracy. The first encouraging results obtained with cBCIs in decision making [36, 230, 229] showed that groups of BCI users can make better decisions than single non-BCI users. Those studies focused on pre-

dicting individual decisions from the neural signals and then aggregating them in a variety of ways to obtain group decisions. The performance of the cBCI was then compared with that of traditional group decisions. In other words, a fully-neural approach (cBCI) was compared with a fully-behavioural one (non-BCI individuals or groups).

The neural and behavioural approaches, however, could be complementary. Humans make decisions as a result of different cognitive processes, including attention, perception, learning, memory, and thinking [40]. The information gathered from our senses (perception) is firstly filtered and integrated with previous knowledge (memory), then we reason to, finally, make a decision. In this pipeline of different cognitive processes our brain discards information considered unreliable or not of interest for the current task. For example, if we see an image for a few milliseconds, our brain will rely mainly on the information gathered from the part of the image under the focus of attention and ignore the rest, therefore remaining subconscious [40]. However, part of this information could actually be very important for the decision-making task, even if we are not aware of it. In fact, the outcome of a decision seems to be encoded in the neural activity much earlier than the user reaches awareness [176], the so called "unconscious mind". Other physiological signals such as involuntary eye movements and heart rate seem also to be correlated with mental workload and decision making [13], although we do not directly control them.

To augment cognition and improve decision making, a BCI could use neural *and* other physiological signals to directly extract relevant information to the decision-making task from the unconscious mind of the user. Part of this information represents the "decision confidence", which is the probability that the

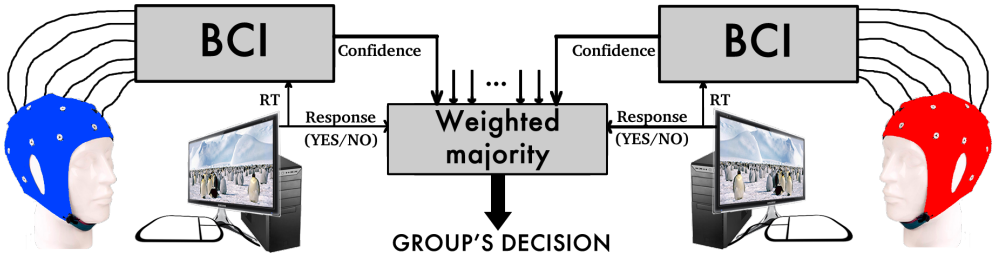

Figure 3.1: Architecture of the proposed collaborative BCI to improve group decisions.

current decision is correct [148]. The cBCI could then obtain group decisions by weighing the individual responses of each group member according to these neural-based confidence estimates [143] – see Figure 3.1. As a result, the integration of conscious and unconscious mind can then lead to superior decisions.

This chapter describes a *hybrid* framework for cBCIs that uses neural and behavioural information to improve group decision making. Section 3.2 presents the main criteria for designing experiments that have been used in this thesis to test the proposed framework. The methods used for recording and preprocessing behavioural, neural and other physiological signals are described in Sections 3.3 and 3.4, respectively. The corresponding methods used to transform this information into decision confidence estimates are then discussed in Sections 3.5, 3.6 and 3.7. Section 3.8 explains how group decisions can be made using these confidence estimates to integrate behavioural responses. Section 3.9 describes how the group performance obtained with different methods are compared. The chapter ends with Section 3.10 which draws some conclusions and makes suggestions for future research and improvements of the proposed framework.

3.2 Experimental Design

Making decisions can be a very challenging task, especially when critical decisions have to be made, for example in health and defence. Neuroimaging techniques such as EEG can reveal important information about the different cognitive processes that lead to a decision. When the decision-making task is related to target detection and recognition (as it is in this thesis), the P300 is usually considered as the most informative ERP for both visual [61, 160] and auditory [144] stimuli.

3.2.1 Notation and Common Features

In the decision-making experiments conducted in this thesis, let N be the total number of trials composing the experiment, each of which includes at least a decision to be made. In order to reduce the effects of drops on performance due to the tiredness/boredom of participants, the trials are split into B blocks (sessions) of $\frac{N}{B}$ trials each. At the end of each block, volunteers are allowed to take a break and rest for a few minutes.

Preliminary results presented in previous research [142] have shown that a hybrid cBCI could augment group performance in a very constrained and simple target detection task. In this thesis, we decided to extend that work to more complex and realistic tasks by proceeding in small steps. For this reason, all decision-making experiments conducted in this thesis share some common features (listed below) which were also used in [142] and that make the recognition of specific ERP components (e.g., P300) easier. For a tutorial on designing ERP experiments, the reader could refer to [101].

Feature 3.2.1 *The decision-making task follows the oddball paradigm.*

In cognitive psychology and BCI, target-detection experiments usually adopt the *oddball paradigm* [146]: users are presented sequences of two different stimuli in a random order, with one (i.e., target) occurring much less frequently than the other (i.e., distractors). Rare stimuli of interest cause a more prominent P300 in the EEG recording, which could easily be detected by the BCI [152]. In many decision-making tasks, users provide an answer more often than another. For example, a driver waiting at an intersection with the red traffic light will decide to keep pushing the brake while the time is passing. When the traffic light turns green (target event), the driver has to decide to release the brake and start accelerating. A similar example involving a more critical decision is when a driver is approaching an intersection with the green light. Most of the times, the car passes with the light remaining green, leading to the decision of accelerating (i.e., standard choice). However, in some cases the traffic light turns yellow while the car is approaching the intersection. In that case, the driver has to decide, in a fraction of a second, whether to pass or brake.

Feature 3.2.2 *The user has to decide between two possible choices.*

In some circumstances the range of possible choices is very large, but most often we deal with binary decisions where we only have two alternatives (i.e., yes/no) [56]. Hence, it is reasonable to focus on *binary decision-making*. However, we should note that the proposed framework could be extended to support multiple-choice decision-making tasks.

Feature 3.2.3 *The task is challenging for a single individual.*

The main aim of the proposed framework is to make better decisions than the average human and group of humans. It is, therefore, obvious that the task

undertaken by the decision makers should not be too simple for them, otherwise the group would not bring any significant advantage.

3.2.2 Protocol

In all experiments, a trial starts by presenting the participant a fixation cross in the middle of the screen for a brief amount of time (i.e., 1 second). This allows the EEG signals to return to the baseline after the previous stimulus and the user to prepare for the next stimulus.

The fixation cross is then followed by the stimulus characterising the experiment, which could be a picture or an audio recording. In order to make the task more difficult (see Feature 3.2.3), several tricks have been adopted. Visual stimuli are presented for a very brief amount of time t_s to preclude the brain the possibility to process all the information gathered from the senses. A mask similar to the ones shown in Figure 3.2 could also be presented after the visual stimulus to clear the iconic memory. When concerning auditory decision-making tasks, the stimuli are spoken sentences affected by various types of noise, making it difficult for the auditory system to understand what is being said in its entirety.

After the stimulus and, possibly, the mask, a display reminding the user to make his/her decision is generally presented.

3.3 Data Recording

The proposed hybrid cBCI for decision making uses a combination of behavioural and physiological measures to improve group decisions. In this section, we describe the methodology used to record each of these measurements.

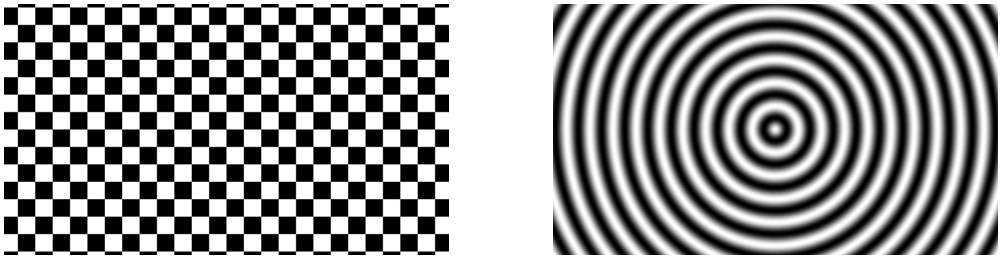

Figure 3.2: Examples of masks to make target detection with visual stimuli more challenging.

The user's decision and response time (RT) are acquired through button clicks of an ordinary USB mouse operated with the right hand. Users are instructed to press the left mouse button to indicate the presence of the target and to press the right mouse button otherwise. While there are typically RT differences when participants use their non-preferred hand over the preferred one, such differences are very small [135] and whether the preferred hand is faster or slower than the non-preferred one depends on the task (e.g., see [81]). Therefore, this constraint is unlikely to affect the individual and group performance.

RTs are measured from the stimulus onset. The USB polling rate is 125 Hz and, therefore, the maximum hardware jitter on the RT measurement is 8 ms. The software presenting the stimuli captures mouse click events every 5 ms and, so, in the worst case scenario the jitter is increased by 5 ms. Furthermore, the EEG status channel of the ActiveTwo device is used to mark the event, which had a further maximum jitter of 1 ms. Summing up, the total maximum jitter on RT measurements is 14 ms, which is far smaller than the average RT in the decision-making tasks considered in this thesis, making the hardware jitter negligible.

Neural data of participants undertaking the decision-making experiments are recorded from 64 electrode sites (according to the 10/20 international system)

using a BioSemi ActiveTwo EEG system. The electrodes are placed on a BioSemi EEG cap worn by the participant with a small amount of conductive gel used to improve conductance and signal quality. We ensure that the impedance of the electrodes is below 20 $k\Omega$. Two additional electrodes are placed on the earlobes for reference.

Eye movements and blinks are recorded by means of a Jazz eye tracker plugged into the Biosemi EEG system and placed on the forehead of the participant on the top of the EEG cap – see Figure 3.3. The eye tracker allows recording of both horizontal and vertical eye movements.

In the experiment conducted in this thesis additional physiological measures have been recorded, including breathing frequency, heart rate and galvanic skin response. These measures have not been used in this thesis but are available for further research, given that they correlate with attention, mental workload and decision confidence [24, 44, 13]. Breathing frequency is recorded by means of a respiration belt worn by the participant on the chest and plugged into the Biosemi EEG system. Heart rate is recorded via two additional electrodes placed on both wrists of the participant. The difference between the two signals is then computed and processed to extract relevant information. Finally, the galvanic skin response is recorded by measuring the impedance of the skin via two passive Nihon Kohden electrodes placed on the index and middle fingers of the left hand of the participant, in order to not obstruct the operation of the mouse.

Neural and physiological signals are sampled at 2,048 Hz.

All experiments conducted in this thesis lasted approximately two hours, including preparation time and task familiarisation. Each participant was paid a base rate of £16 for volunteering and signed an informed, consent form before

Figure 3.3: Position of the different sensors recording the physiological signals. The eye tracker is placed on top of the EEG cap without obstructing the sight of the participant.

taking part in the experiment. The research described in this thesis has received MoD and University of Essex ethical approval in July 2014.

3.4 Data Preprocessing

EEG data from each channel are referenced to the mean of the electrodes placed on each earlobe. Data are then band-pass filtered between 0.15 and 40 Hz with a non-causal 14677-tap FIR filter obtained by convolving a windowed low-pass filter with a windowed high-pass one. The choice of these filters is motivated by the promising results obtained with the visual matching task (see Chapter 4 and [143]). Artefacts caused by eye-blinks and other ocular movements are re-

moved by using a standard subtraction algorithm based on correlations between the average value recorded at electrode sites Fp1 and Fp2 and the average value recorded at F1 and F2 [153].

EEG data are then segmented into two types of epochs: stimulus-locked and response-locked. Stimulus-locked epochs are extracted from 200 ms before the onset of the stimulus and have a duration of 1900 ms, while response-locked epochs also last 1900 ms but start 1200 ms before the user's response – see Figure 3.4. The extracted epochs are then de-trended on a channel-by-channel basis by subtracting the average voltage recorded in the first five samples. Depending on the experiment conducted, it is possible for the response- and stimulus-locked epochs to overlap (albeit to different degrees). However, it should be noted that the stimulus-locked epochs are still very different from the response-locked ones and, therefore, together they carry more information than each type on its own.

Epoch data are then low-pass filtered with an optimal 820-tap FIR filter designed with the Remez exchange algorithm [116] with a pass band of 0–p_b Hz and a stop band of s_b–1024 Hz. The choice of this filter is motivated by the promising results obtained with the visual matching task (see Chapter 4 and [143]). The data are finally down-sampled to s_r Hz to speed up the computation without affecting the detection of meaningful variations (e.g., P300s) in the EEG data. p_b, s_b and s_r have been set to 14, 16 and 32 Hz, respectively, for the face recognition experiment (see Chapter 8) and to 6, 8 and 16 Hz, respectively, for the remaining experiments. Finally, the first and last 200 ms of each epoch were trimmed (see black striped areas in Figure 3.4) to obtain epochs of 1500 ms and avoid transient effects. Therefore, each epoch is represented by a total of 48 and 24 samples per channel for s_r equal to 32 or 16 Hz, respectively.

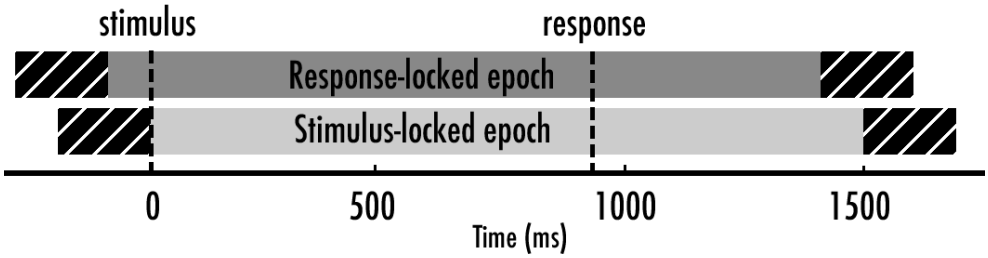

Figure 3.4: Protocol adopted to segment the EEG data into stimulus-locked and response-locked epochs. The black striped areas are trimmed after low-pass filtering and subsampling the extracted epochs.

The vertical component of the eye movements recorded by the eye tracker is also band-pass filtered between 0.15 and 40 Hz with the same filter used for the EEG data. We use the vertical component as this is also influenced by eye blinks, which correlate with the mental workload [13], and because in preliminary tests we found that the horizontal component did not seem to contribute any additional information. The resulting signal is then referenced to the mean value recorded during the presentation of the fixation cross, i.e., one-second interval before the stimulus. The eye data are then segmented into stimulus-locked and response-locked epochs, the former starting at the onset of the stimulus and lasting 500 ms, the latter starting 250 ms before the user's response and also lasting 500 ms – see Figure 3.5. The choice of a different duration for the eye movement epochs compared to the EEG epochs is motivated by the reasonable assumption that, on the one hand, the eyes will move mainly during the presentation of the stimulus (250 ms) and the mask (250 ms), so there is no need to extend the stimulus-locked epochs to more than 500 ms. On the other hand, with the response-locked epochs, we would like to capture the eye activity when the user is about to provide an answer. The promising results obtained in visual search (see Chapter 5 and [199]) validated this choice.

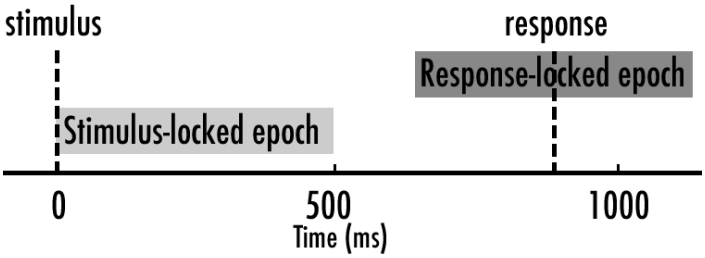

Figure 3.5: Protocol adopted to segment the vertical component of the eye movements into stimulus-locked and response-locked epochs.

We should note that the non-causal filters adopted by our framework could prevent it to be used in online applications. However, it is possible to modify the FIR filters to be causal [218] or to use different filters already used in collaborative BCIs [229]. While this thesis is mainly focused on the off-line validation of the proposed cBCI, we believe it is possible to extend the framework to support also real-time BCI applications.

3.5 Epochs Labelling

The strategy adopted by the proposed hybrid cBCI to improve group decisions is to estimate the confidence of each group's member (i.e., the probability of his/her decision to be correct [122]) via machine learning algorithms. This requires ground-truth information on the actual confidence in an appropriate training set to fit the predictive model. In principle participants could be asked to rate their degree of confidence in the decisions of the training set and the cBCI could use these values to fit its model. However, as we will see in Chapter 6, this measure can be biased and unreliable [96, 132].

Another approach would be to use the *correctness* of individual decisions to fit the cBCI model, as this information is available to the cBCI in the training

set. This will lead to assign high values of confidence when the response provided by the participant is very likely to be correct and low values otherwise, a so-called *well-calibrated* system [122]. This strategy is optimal considering that the proposed cBCI uses these confidence values to weigh individual responses and obtain group decisions. Moreover, it follows the approach of rational observers, who tend to be less confident when they do not have enough information to make an informed choice (and, therefore, are more likely to be incorrect) and more confident when they are likely to be correct.

To train the machine learning component of the cBCI, the trials in the training set in which the decision made by a participant was correct (independently from the presence or absence of the target) have been labelled as "confident" (-1 label), and the trials where the decision was incorrect as "non-confident" ($+1$ label). This means that the cBCI is trained to predict whether a user made a confident (correct) or a non-confident (incorrect) decision, and not, unlike other studies [229, 17], to predict the response of the user.

3.6 Feature Extraction

One of the aims of this thesis was to find the best combination of behavioural and physiological features for estimating the decision confidence. For these reasons, different types of features have been extracted from the preprocessed signals in the various experiments and the performance of cBCIs based on each type (or combinations of them) have been compared. This section provides a brief overview of the methods used across the thesis for extracting neural and eye-movement features.

3.6.1 Neural Features

The available EEG data are characterised by high dimensionality: each epoch is represented by either 24 or 48 samples for each of the 64 available channels. This means that, even in the best scenario, the cBCI has to predict the decision confidence of one trial from a total of 3,072 values. However, the number of trials available in the training set is much smaller (about 300) than the number of features describing a trial. In these conditions, the predictive power of the machine learning algorithms reduces as the dimensionality increases (Hughes phenomenon [70]). Therefore, the proposed cBCI requires a process of feature extraction and selection to reduce the dimensionality of the classification process.

A well-known method used for this purpose in BCI research is Principal Component Analysis (PCA) [33], an orthogonal linear transformation that projects the data into a subspace where the components are ordered by the magnitude of their variance. PCA is based on the idea that it is possible to represent most of the variation in the original dataset with a small set of "principal" components, which are linear combinations of the original variables [158]. These components are obtained by extracting the eigenvalues and eigenvectors of the covariance matrix. Spatial PCA has been used in BCI research to select the most representative channels for the task at hand [189]. However, information related to the decision confidence is also likely to appear in the temporal domain. Thus, in this thesis we adopt a *spatio-temporal* PCA [30], which considers each sample of each channel in an epoch as a separate stochastic variable. For each trial, the epochs recorded in the 64 channels are concatenated and a covariance matrix is computed. The PCA features are extracted by performing the dot product between the first p

eigenvectors of the covariance matrix (i.e., most important principal components) and the voltage values in the concatenated epoch.

Another important method broadly used in BCI, especially for identifying motor imagery, is Common Spatial Pattern (CSP) [155, 235, 159]. This *supervised* spatial filter aims at separating a multivariate signal (e.g., EEG) into additive subcomponents having maximum difference in variance between two classes. It could be thought of as a supervised version of PCA. Let X_1 and X_2 be the sets of trials associated with class 1 and class 2, respectively. CSP aims at finding the component w^T such that the ratio of variance between the two sets is maximised:

$$w = \text{argmax}_w \frac{||wX_1||^2}{||wX_2||^2} \tag{3.1}$$

After computing the CSP matrix on the data of the training set for each type of epochs, the data on the test set are transformed by performing the dot product between the epochs and the CSP matrix. The first and last columns of the resulting matrix are then selected as they represent the most significant patterns, i.e., those with the maximum difference in variance. The variances of these two columns are then used as neural features to represent the decision confidence. In this thesis we only use two neural features for each type of epochs to promote efficiency and generalisation.

It should be noted that while the version of PCA we employ is spatio-temporal, CSP takes into account only spatial information. This could lead to losing important information related to the decision confidence stored in the temporal domain, especially considering that we deal with ERPs. For this reason, we have also used a spatio-temporal version of CSP termed Local Temporal Cor-

relation Common Spatial Pattern (LTCCSP) [233] for extracting features from the EEG data. LTCCSP introduces a weight matrix to impose larger coefficients on patterns that are similar within a local temporal range τ. In this thesis, we have empirically set $\tau = 10$ regardless of the final sampling rate (i.e., either 16 or 32 Hz). Once the LTCCSP matrix is computed and multiplied by this new weight matrix, the process of extracting neural features is similar to the one used for CSP.

3.6.2 Eye-Movement Features

When the decision-making task involves visual stimuli, eye movements could also be related to the decision confidence [208] as well as to the mental workload [7]. For these reasons, we extracted three features from the stimulus-locked epochs of the eye-movement vertical component. The first feature is represented by the total distance covered by the eyes along the vertical axis during stimulus presentation (i.e., first 250 ms of the epoch). This feature aims at describing the number of saccades and the effort made by the eyes in spotting the target in the stimulus. Ideally, if the total distance is high it is likely that the eyes did not spot the target and, therefore, the participant is less confident about the decision. Another feature extracted from these epochs is the standard deviation of the vertical eye movements during stimulus and mask presentation (i.e., whole stimulus-locked epoch). This feature is likely to describe how spread the eye movements are during and after the stimulus presentation. Furthermore, we also compute the mean of the numerical derivative of the vertical eye movements in the same time window, to consider the velocity of the eyes in scanning the picture.

An additional feature is then extracted from the response-locked epochs. The first derivative of the signal recorded in the epoch is computed and its mean is used as a feature representing the velocity of eye movements before and after making the decision.

The promising results obtained in visual search (see Chapter 5 and [199]) validated the choice of these four features.

3.7 Confidence Estimation

Given a feature vector composed by a subset of the features described in the previous section, the cBCI needs to predict the decision confidence of the user in a particular trial. This requires a machine learning algorithm. We chose Least Angle Regression (LARS) [37] for its linearity (i.e., to keep the framework simple) and its intrinsic ability to also perform feature selection, which might be useful in future extensions of the current framework. The positive results obtained in visual matching (Chapter 4 and [143]) validated this choice.

The decision confidence is computed as follows:

$$f = \sum_{j=1}^{C} a_j \cdot x_j + \epsilon \tag{3.2}$$

where a_j and ϵ are constant coefficients (to be identified via a training set when fitting the model) and x_j is the j-th component of the feature vector.

Once a confidence estimate, f_i, is available for a particular decision of participant i, it is transformed to a weight w_i according to the following negative

exponential weighting function:

$$w_i = \exp(2.5 - f_i). \tag{3.3}$$

This weighting function has been chosen in preliminary tests and is motivated by the desire to allow confident users to count substantially more than uncertain users in the group's decision, thanks to the negative exponential. By adding the constant 2.5 to the exponent we ensure there is reasonable variation in weights in the range of values of LARS' outputs, a necessary condition to do better than the majority rule. This weighting function is also desirable as it is always positive, avoiding negative weights which would imply changing "yes" decisions into "no" ones or *vice versa*.

It should be noted that, by mapping incorrect decisions to label $+1$ and correct ones to -1 (see Section 3.5), the raw prediction given by the cBCI is proportional to the probability of the user to be *incorrect*. To transform this into the probability of being correct (i.e., our interpretation of the decision confidence), we use the non-linear weighting function to associate higher values of confidence to high probabilities of being correct and *vice versa*.

3.8 Group Decisions

As shown in Section 2.5, different methods could be used to integrate decisions of multiple participants to obtain a group's decision. Multi-brain fusion at the decision level (see Figure 2.3) seems to give the highest performance amongst feature and signal fusion techniques. Moreover, for the structure of the proposed frame-

work, decision fusion is the most appropriate method for integrating individual responses.

The cBCI obtains group decisions by using a weighted majority rule, described as follows:

$$d_{group} = \texttt{sign} \sum_{i=1}^{m} w_i \cdot d_i \tag{3.4}$$

where `sign` is the sign operator, m is the group's size, $d_i = \{-1, 1\}$ is the decision of participant $i = 1, \ldots, m$ ($d_i = -1$ means a correct decision), and $w_i \in \mathbb{R}^+$ is the weight associated with the confidence of participant i in the current decision computed as described in the previous section. In case of tie (i.e., $d_{group} = 0$), a random decision is made. While ties could easily happen in even-sized groups using standard majority, they are very unlikely to happen when using a weighted majority as the weights are real numbers.

3.9 Results Validation

In order to validate the results obtained by the cBCI and reduce the risk of overfitting, 10-fold cross-validation is used to split the dataset of each experiment in 10 different training and test sets. In each fold 90% of the trials are used for training and the remaining 10% for testing. The same non-overlapping sets are built for each participant.

All the possible $\binom{P}{s}$ groups of size s that could be assembled with the P participants are then built, for $s = 2, 3, \ldots, P$. The average cross-validation error rate obtained by each group using the proposed cBCI with different sets of features is compared with that achieved by traditional groups using the standard majority

(i.e., a weighted majority where $w_i = 1, \forall i{=}1, \ldots, m$). To test if the observed differences in error rates using different methods are statistically significant, we compare the error distributions within each group size by using the one-tailed Wilcoxon signed-rank test with the Bonferroni correction. We choose this paired-data test since all decision methods are applied to the same groups and as it relies on fewer assumptions than parametric tests (i.e., it does not assume that the data are Gaussian distributed).

3.10 Conclusions

This chapter has described the architecture of the proposed collaborative BCI for improving group decision making, as well as listing the shared features of the experiments that have been used to test the cBCI (see following chapters).

While most of the data recording and preprocessing procedures and techniques are used in all experiments, the feature extraction step changes between different tasks. The reason behind this is dual: (a) traditional methods used in previous studies did not work well with new, realistic experiments, and (b) during our research we identified certain methods for feature extractions (e.g., CSP) performing much better than others (e.g., PCA) used previously – see Chapter 5.

The feature extraction step is probably the most important design choice of the cBCI. For this reason, the classification method used to transform the features into decision confidence (i.e., LARS) has been reused in all experiments. More advanced and traditional machine learning algorithms such as logistic regression and support vector machines could be used and their performance could be compared in future research. Moreover, it would be interesting to study the

performance of cBCI-assisted groups in experiments including multi-choice decision tasks (i.e., not satisfying Feature 3.2.2).

Chapter 4

Improving Group Performance in Visual Matching

This chapter describes the first results obtained with the proposed framework described in Chapter 3 when applied to a simple visual matching task. Most of the material in this chapter has been published in [143].

4.1 Introduction

Decades of research in artificial intelligence have been spent trying to build computer systems that could outperform the human visual system. Despite recent advances in computer vision, the human brain remains superior in processing and interpreting the information coming from the senses for most of the applications. This is because of its ability of processing visual information using features and learning processes, which are critical for recognition but not used in computer vision algorithms [193]. However, our visual system is not perfect. When the

perceptual load is high (e.g., when processing complex and crowded scenes), the time available is not sufficient for completing the processing, or the attention is divided amongst multiple tasks, our brain can make mistakes. Phenomena like attentional blink and repetition blindness, which have been studied for years, can show the limitations of our perception and cognition[133, 21, 106, 31] which result in observers being able to perceive only a subset of the features of a complex scene.

These limitations can lead to suboptimal performance in tasks that require visual perception, for example decision making [68]. When critical decisions have to be made, wrong perception could have serious consequences, for example in identifying a threat in a scene. To partly overcome these limitations, two or more individuals could be involved in the decision-making process. Groups generally have augmented perception, especially when the information is not shared among their members [190], and error correction capabilities, which could produce better decisions than an individual. Although two heads are not necessarily better than one [5], technology such as BCIs could further enhance group perception. Collaborative BCIs have already been successfully used for enhancing the detection of a visual stimulus [230, 229] or the discrimination between images of faces and images of cars [36].

This chapter examines the possibility of using the hybrid cBCI presented in Chapter 3 to augment group performance in a visual matching task characterised by high perceptual load and high speed of stimulus presentation. As mentioned before, in these conditions human perception may not only be incomplete but also incorrect, leading to erroneous decisions. The cBCI could tap into the unconscious and conscious processes and extract relevant information to improve

the evaluation of the images. This research has been described in [143].

The chapter is organised as follows. Section 4.2 describes the visual matching experiment and details the methodology used by the cBCI to make group decisions introduced in Chapter 3. Section 4.3 presents and discusses the results obtained in the experiment. Finally, Section 4.4 draws some conclusions.

4.2 Methodology

This section describes the protocol used in the visual matching experiment and briefly recalls the methods employed by the cBCI to obtain group decisions.

It should be noted that the design of the experiment and the data recording were performed in a previous research project and, therefore, were not part of this PhD thesis. Here, we analysed the data and used them to assess the performance of the cBCI described in Chapter 3. This study received ethical approval on the 30th of May 2012 by the Research Director of the School of Computer Science and Electronic Engineering of the University of Essex on behalf of the university's Faculty Ethics Committee.

4.2.1 Participants

Data were gathered from 11 healthy participants with normal or corrected-to-normal vision (average age 30.6 ± 9.5 years, 6 females, 8 right handed). The preliminary analysis of individual performance of the participants revealed that one observer gave responses that were hardly distinguishable from random. For this reason, the data recorded from that participant were discarded and, therefore, the analysis were conducted on the remaining 10 participants.

4.2.2 Stimuli and Tasks

Participants underwent a sequence of 8 blocks of 28 trials each, for a total of 224 trials. Each trial (see Figure 4.1) started with the presentation of a fixation cross in the middle of the screen for 1 second, followed by a black screen for another second. Then observers were presented with a sequence of two displays, each showing a set of shapes. The two displays were showed for 83 ms (5 frames of a 60 Hz screen) and 100 ms (6 frames), respectively. The first display was immediately followed by a mask for 250 ms and a black background for 100 ms. The mask was a vertical sinusoidal grating with a period of 1 degree subtending approximately 8 degrees. Following this sequence of displays, observers had to decide, as quickly as possible, whether or not the two sets of shapes were identical. Responses were given with the two mouse buttons (left for "identical", right for "different"), controlled with the right hand, and response times (RTs), measured from the onset of Set 2, were recorded. Each stimulus display consisted of three shapes (subtending approximately 1.5 degrees and being approximately 1.8 degrees apart), which could be any combination of a triangle, square and pentagon (see Sets 1 and 2 in Figure 4.1). The same shape was allowed to be present multiple times within a set. Each shape was coloured either in pure white (corresponding to normalised RGB (1,1,1)) or light grey (RGB (0.65,0.65,0.65)). Shapes were presented on a black background.

With two colours and three possible shapes we can obtain six elements: white triangle, light grey triangle, white square, light grey square, white pentagon, light square pentagon. Each set of shapes contained three elements and, therefore, there were a total of $6^3 = 216$ different stimuli, leading to a $216^2 = 46,656$

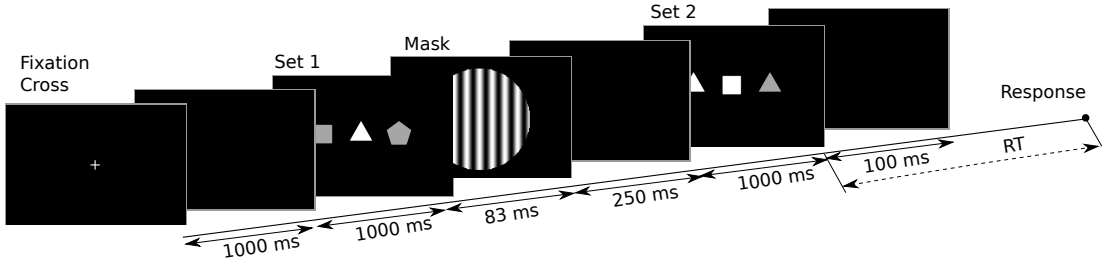

Figure 4.1: Stimulus sequence used in the visual matching experiment.

possible set combinations. Each pair of displays was classified by counting the number of matching features (i.e., colour and shape) of their ordered shapes, a number that we called *degree of match* (DoM). If all three stimuli of Set 1 differ in both shape and grey level from the three stimuli in Set 2, we have a DoM of 0; if the first two elements shares one feature (e.g., the same colour) and the third element shares both features, that is a DoM of 4; etc. So, DoM ranges from 0 to 6, with 6 corresponding to a perfect match between Set 1 and Set 2.

It should be noted that the DoM influences the difficulty of the task at hand. If the two sets of shapes are very similar (high DoM) but different, the processing of the displays would require more time than that available, making the observers more erroneous. The same happens when the two sets of shapes match (DoM = 6) as participants have to make sure they did not miss any mismatching feature before deciding that the two displays contained the same shapes. On the other hand, if the two sets do not share any feature, it should be quite straightforward for the user to make the correct decision. In order to better control the difficulty of the task, the experiment was designed to have an equal proportion of each DoM category in each block. Therefore, there were four trials for each value of DoM $\in \{0, 1, 2, 3, 4, 5, 6\}$.

The order of the trials was randomly shuffled and identical sequences were

used for all participants. This ensures that all participants underwent exactly the same experiment, which should increase repeatability and reproducibility, while allowing groups to be formed offline to test the performance obtained with the proposed cBCI without requiring to acquire data from all participants simultaneously.

The experimental blocks were preceded by a session of practice to allow observers to familiarise with the task and the stimuli. Participants were seated comfortably at about 80 cm from an LCD screen. Briefing, preparation of participants (including checking and correcting the impedances of the electrodes used for EEG recording) and task familiarisation took approximately 30 minutes, while the experiment took about 20 minutes.

4.2.3 Data Acquisition and Transformation

Participants undertook the visual matching experiment in conditions of complete absence of communication or any other form of social influence.

Neural data were acquired and preprocessed as explained in Sections 3.3 and 3.4.

As a first test of our cBCI, as a method for extracting neural features we adopted space-time PCA. We selected the 24 principal components of each epoch as neural features. This corresponds to a 1 to 64 reduction from the original 1,536 features (i.e., 24 samples for each of the 64 channels available). Due to the simplicity of the classifier used to transform the features into decision confidence and to reduce the risk of overfitting, we decided to only use response-locked epochs. We did analyse response- and stimulus-locked epochs but found that

the former contained more information related to the decision confidence than the latter – see Section 4.3.5. Therefore, we extracted the 24 neural features from the response-locked epochs starting 1000 ms before the response and lasting 1500 ms.[1]

4.2.4 Decision Confidence Estimation

The 24 PCA neural features extracted from each epoch were transformed in confidence estimates by using LARS, as described in Section 3.7. The presence of this machine learning component required splitting the available data into a training set (used to fit the model) and a test set (to evaluate the model on unseen data). We then used the correctness in the decision as ground-truth information for the confidence estimation (see Section 3.5). As it is customary for small dataset, such as the ones used in BCI research, and as described in Section 3.9 we adopted a *k-fold cross-validation* approach. In order to ensure all folds had the same number of samples, as the number of trials (224) is divisible by 7 and by powers of 2 up to 2^5 but not by 10 (as indicated in Section 3.9), in this experiment we used k = 2, 4, 7, 8, 14, 16, 28, 32, 56, 112 and 224 (leave-one-out strategy). Since the performance varied very little with k [143], we will only report results for $k = 16$.

The data on the training set have been used to fit the LARS model on a participant-by-participant basis. Then, the neural features in the test set have been transformed into confidence correlates using the fitted model. For the rest of the chapter, we will call these neural confidence correlates *nf*, to indicate their

[1]For simplicity, in this experiment we did not use physiological measures other than the brain signals to predict the decision confidence. Eye movements and the other measures described in Section 3.3 have been recorded and we plan to use them in future research.

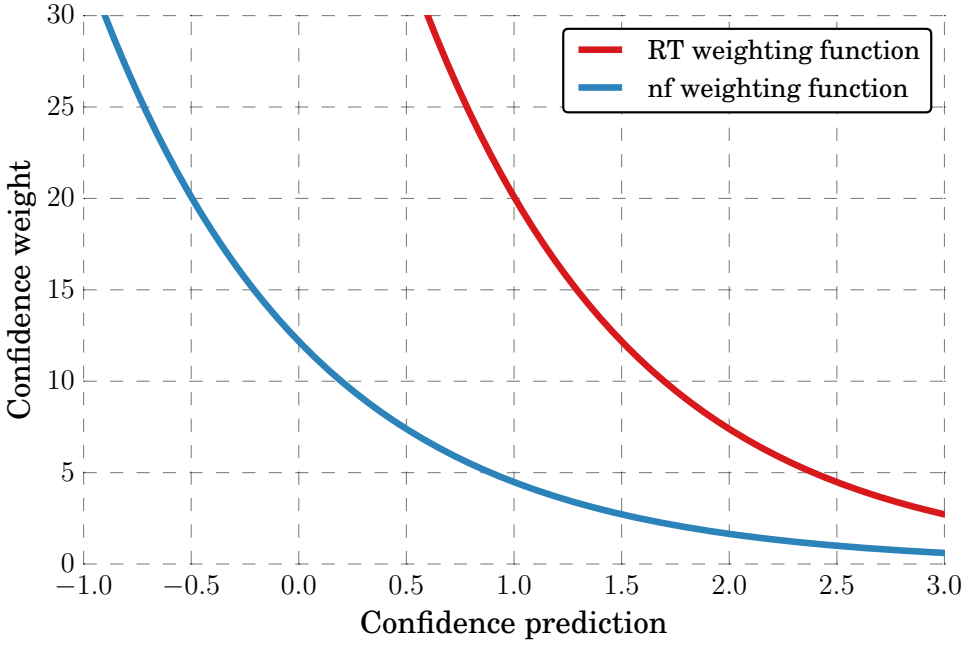

Figure 4.2: Plots of the negative exponential weighting functions adopted in our studies to transform neural (blue) and behavioural (red) correlates of confidence into weights. The shape of these functions allows confident decisions to count more than uncertain ones.

dependency on the neural features only. These confidence values were then transformed to weights using the negative exponential weighting function described in Section 3.7 and plotted in blue in Figure 4.2.

Response times have also been used as an alternative "behavioural method" to measure the decision confidence, as they are predictors of correctness [100]. As described earlier, slower RTs are generally associated with uncertainty in the decision and, therefore, a high likelihood to err. To obtain confidence weights from the raw RTs, we used a weighting function similar to the one used for the neural features given by

$$w_{RT,i} = \exp(4 - RT_i), \tag{4.1}$$

where RT_i is the response time for observer i in a particular decision. Figure 4.2 shows a plot of this weighting function in red.

Furthermore, we have also used an optimal combination of the behavioural and neural features as confidence weights. For the rest of the chapter, we will call these "neuro-behavioural" confidence correlates *RTnf*. Given an observer i, the decision confidence estimated using the 24 neural features and the response time in a particular decision, the neuro-behavioural confidence weights are computed as:

$$w_{RTnf,i} = 0.75 \cdot w_{RT,i} + 0.25 \cdot w_i, \tag{4.2}$$

where $w_{RT,i}$ and w_i are the weighting functions described in Equations (4.1) and (3.3), respectively.

The choice of the coefficients 0.75 and 0.25 was simply guided by our experience. BCIs tend to be relatively unreliable in single-trial classification tasks. Since our system requires trial-by-trial decisions, by giving more influence to the confidence weight inferred from *RT* we attempted to compensate for the higher noise expected in *nf*. By combining these two methods we hoped to obtain a more robust confidence measurement which would then result in better decisions.

4.2.5 Making Group Decisions

The simplest method to obtain group decisions from a set of individual responses is by using the standard majority rule (i.e., traditional non-BCI groups). In this case, all observers' decisions (either a "yes" or a "no") count the same. The final decision is based on straight majority for teams with an odd number of members and majority followed by the flipping of an unbiased coin in the case of ties for

teams with an even number of members.

In this experiment, group decisions made with the majority rule were compared with those obtained by using a weighted majority, where the decision made by each observer was weighed according to the confidence weights computed using either the behavioural, neural or neuro-behavioural methods described in Section 4.2.4. Given the confidence weights c_i of participant i computed according to either the RT-based ($w_{RT,i}$), nf-based (w_i) or $RTnf$-based ($w_{RTnf,i}$) methods explained earlier for all group's members, the group decision is made as:

$$\text{decision}_{\text{group}} = \begin{cases} \text{yes} & \text{if } \sum_{i \in \mathcal{Y}} c_i > \sum_{j \in \mathcal{N}} c_j \\ \text{no} & \text{otherwise,} \end{cases} \tag{4.3}$$

where $\mathcal{Y}$ and $\mathcal{N}$ represent the sets of all observers in the group who decided "yes" and "no", respectively.

Since response times are influenced by, and thus can reveal, the confidence in a decision [100] and that more confident responders are more likely to be correct, we could assume that, typically, faster responders are correct more often than slower ones. For this reason, we decided to also investigate the group performance of a behavioural decision-making system where only the fastest responders in a group were allowed to influence the group decision, as will be described in detail in Section 4.3.4.

4.3 Results

This section presents and discusses the results obtained with the 10 participants and the cBCI described in the previous section.

4.3.1 Individual Decisions

We start our analysis by looking at the differences in performance shown by the 10 participants when performing the task in isolation and without any manipulation of their decisions.

The individual performance of the participants in the visual matching task used in our experiment was quite variable, with error rates ranging from just below 5% to over 20% – see Figure 4.3. The average error rate across all participants was 12.5%, showing that the task was quite challenging for individuals. Interestingly, if we look at the subset of trials where matching pairs of stimuli were presented, we see that participants gave incorrect decisions in only 0 or 1 out of the 28 matching pairs, thereby showing a very high sensitivity to identical sets. The bulk of the errors, instead, were due to participants that indicated as "matching" stimuli that were actually not containing the same shapes.

4.3.2 Metacognitive Accuracy of Confidence Estimates

Let us now turn our attention to the neural and behavioural correlates of decision confidence.

To investigate the relationship between correct/incorrect responses and the confidence with which decisions were taken (i.e., metacognitive accuracy [122]), we studied the distributions of the RT, nf and $RTnf$ confidence weights obtained

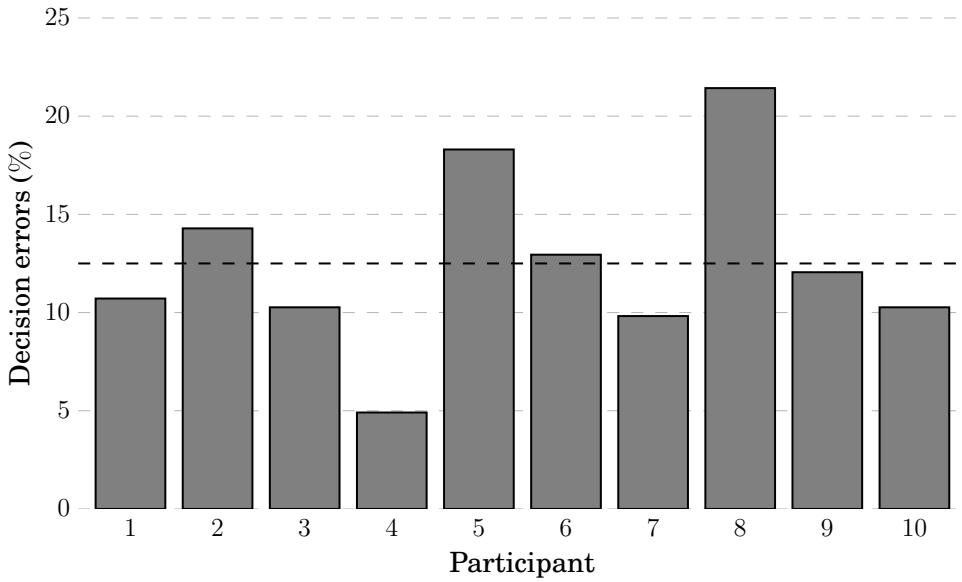

Figure 4.3: Percentage of erroneous decisions made by each participant in the 224 trials of our experiment. The average error rate across participant is indicated by the black dashed horizontal line.

as indicated in Equations (4.1), (3.3) and (4.2), respectively.

We started by binning the data (obtained via cross-validation) on the basis of whether a decision made in a trial by an observer was correct or incorrect. Table 4.1 reports the medians of the confidence weights associated to the behavioural feature RT and the neural features nf, and the neuro-behavioural mixing of the two, $RTnf$, for correct and incorrect trials. The corresponding box plots and density functions (obtained via a kernel-based estimator) are shown in Figure 4.4. As one can see from these, the medians of the confidence weights are significantly lower for the incorrect decisions than for the correct ones for all the features used.

We used two non-parametric tests to assess whether these differences were statistically significant: the one-way Kruskal-Wallis test and the Wilcoxon rank-sum

Table 4.1: Medians (across all participants) of the confidence weights associated to behavioural, neural and neuro-behavioural methods as a function of whether the user's response was correct or incorrect.

Decision	*RT*	*nf*	*RTnf*
Correct	27.514	26.967	27.543
Incorrect	22.721	21.943	22.412

test.[1] Sample sizes were 1,960 for the "correct" class and 280 for the "incorrect" class. The use of non-parametric tests was required as the distributions of confidence weights (see Figure 4.4(right)) are clearly non-Gaussian. In all comparisons and for both tests, $p<10^{-17}$ with statistics $H>77.7$ and $W>151,740$ in all cases. These tests indicate that trials where the confidence weights were characterised by lower values were also those where decisions were more difficult (and were, therefore, taken with a high level of uncertainty) than those characterised by higher weights. Behavioural, neural and neuro-behavioural estimates of the decision confidence seem therefore to provide a good metacognitive accuracy across participants.

We also repeated the analysis on a participant-by-participant basis to further validate these results. Table 4.2 reports the p-values of the one-way Kruskal-Wallis and the Wilcoxon rank-sum tests. As can be seen, the weights associated to the "correct" trials are significantly different than those related to "incorrect" trials for most methods and participants. In particular, it should be noted that the *RTnf* method seems to be the best out of the three analysed, as the distributions of the confidence weights are significantly different for all participants.

We then binned the data on the basis of the degree of match of the stimuli

[1] Unlike what we reported in [143], here we used the approximated Wilcoxon rank-sum test as implemented in `R`. However, this did not significantly affect the results.

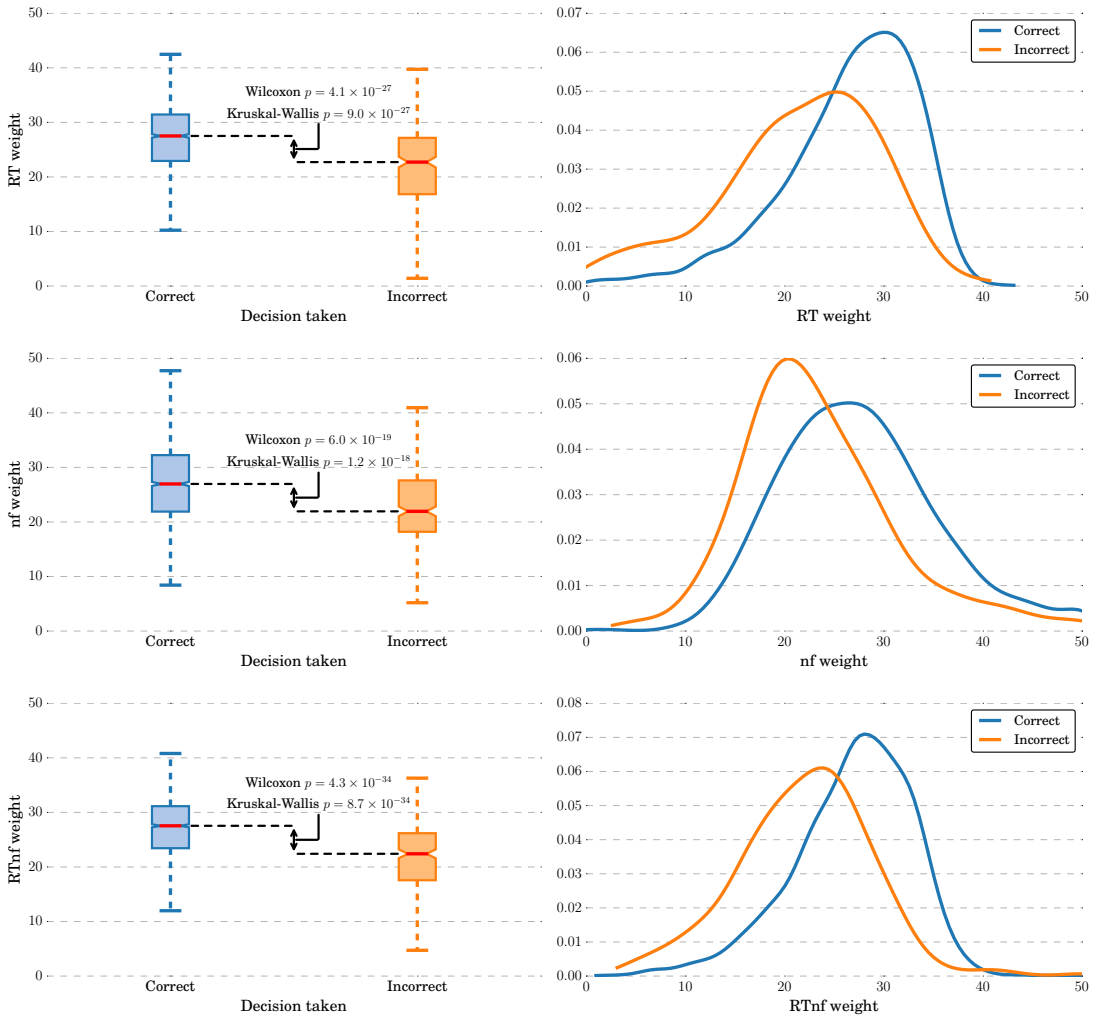

Figure 4.4: Box plots representing the distributions of the weights across participants for different features and decisions (left) and corresponding probability density functions (right). The plots on the left also report the p-values of the Kruskal-Wallis and Wilcoxon rank-sum tests comparing the two distributions.

Table 4.2: p-values of the one-way Kruskal-Wallis test (KW) and Wilcoxon rank-sum test comparing the distributions of the weights for "correct" and "incorrect" trials for different features for each participant. p-values below the significance level 0.05 are reported in bold.

	RT		*nf*		*RTnf*	
User	KW	Wilcoxon	KW	Wilcoxon	KW	Wilcoxon
1	$\mathbf{2.7 \times 10^{-4}}$	$\mathbf{1.4 \times 10^{-4}}$	$\mathbf{4.4 \times 10^{-4}}$	$\mathbf{2.2 \times 10^{-4}}$	$\mathbf{3.7 \times 10^{-5}}$	$\mathbf{1.9 \times 10^{-5}}$
2	9.3×10^{-2}	$\mathbf{4.7 \times 10^{-2}}$	1.1×10^{-1}	5.7×10^{-2}	$\mathbf{2.8 \times 10^{-2}}$	$\mathbf{1.4 \times 10^{-2}}$
3	$\mathbf{1.1 \times 10^{-2}}$	$\mathbf{5.6 \times 10^{-3}}$	$\mathbf{2.8 \times 10^{-2}}$	$\mathbf{1.4 \times 10^{-2}}$	$\mathbf{6.7 \times 10^{-3}}$	$\mathbf{3.3 \times 10^{-3}}$
4	7.2×10^{-2}	$\mathbf{3.6 \times 10^{-2}}$	$\mathbf{1.3 \times 10^{-2}}$	$\mathbf{6.4 \times 10^{-3}}$	$\mathbf{1.7 \times 10^{-2}}$	$\mathbf{8.4 \times 10^{-3}}$
5	$\mathbf{7.6 \times 10^{-7}}$	$\mathbf{3.7 \times 10^{-7}}$	$\mathbf{1.1 \times 10^{-4}}$	$\mathbf{5.7 \times 10^{-5}}$	$\mathbf{9.2 \times 10^{-8}}$	$\mathbf{4.7 \times 10^{-8}}$
6	$\mathbf{1.4 \times 10^{-3}}$	$\mathbf{6.7 \times 10^{-4}}$	2.3×10^{-1}	1.2×10^{-1}	$\mathbf{1.7 \times 10^{-3}}$	$\mathbf{8.3 \times 10^{-4}}$
7	$\mathbf{8.6 \times 10^{-5}}$	$\mathbf{4.7 \times 10^{-5}}$	$\mathbf{1.2 \times 10^{-3}}$	$\mathbf{6.2 \times 10^{-4}}$	$\mathbf{4.5 \times 10^{-6}}$	$\mathbf{2.2 \times 10^{-6}}$
8	$\mathbf{9.6 \times 10^{-3}}$	$\mathbf{5.0 \times 10^{-3}}$	$\mathbf{4.6 \times 10^{-4}}$	$\mathbf{2.3 \times 10^{-4}}$	$\mathbf{2.9 \times 10^{-4}}$	$\mathbf{1.5 \times 10^{-4}}$
9	$\mathbf{2.3 \times 10^{-5}}$	$\mathbf{1.1 \times 10^{-5}}$	$\mathbf{1.7 \times 10^{-3}}$	$\mathbf{8.8 \times 10^{-4}}$	$\mathbf{9.0 \times 10^{-7}}$	$\mathbf{4.6 \times 10^{-7}}$
10	$\mathbf{8.1 \times 10^{-6}}$	$\mathbf{4.1 \times 10^{-6}}$	$\mathbf{5.2 \times 10^{-3}}$	$\mathbf{2.6 \times 10^{-3}}$	$\mathbf{6.2 \times 10^{-6}}$	$\mathbf{3.1 \times 10^{-6}}$

presented in each trial, as the DoM is an indicator of the objective difficulty of the task of discriminating them. Table 4.3 reports the medians (across all participants) of the confidence weights associated to different features as a function of the DoM of the stimuli used in a trial. The corresponding box plots are shown in Figure 4.5.

Overall, as we hypothesised, stimuli configurations characterised by higher DoM, which are thus objectively harder to decide upon and more likely to end up with incorrect decisions, are associated with lower confidence weights. This suggests that the neural and behavioural features do indeed capture the decision confidence.[1]

[1]We should note that here we are comparing confidence values in trials of different DoM *regardless of the correctness of the decision*. Hence, although for DoM=6 (matching stimuli) we have a median confidence lower than for other stimuli, there were still significant differences between trials where the user made the correct choice (confidence higher than the median) and

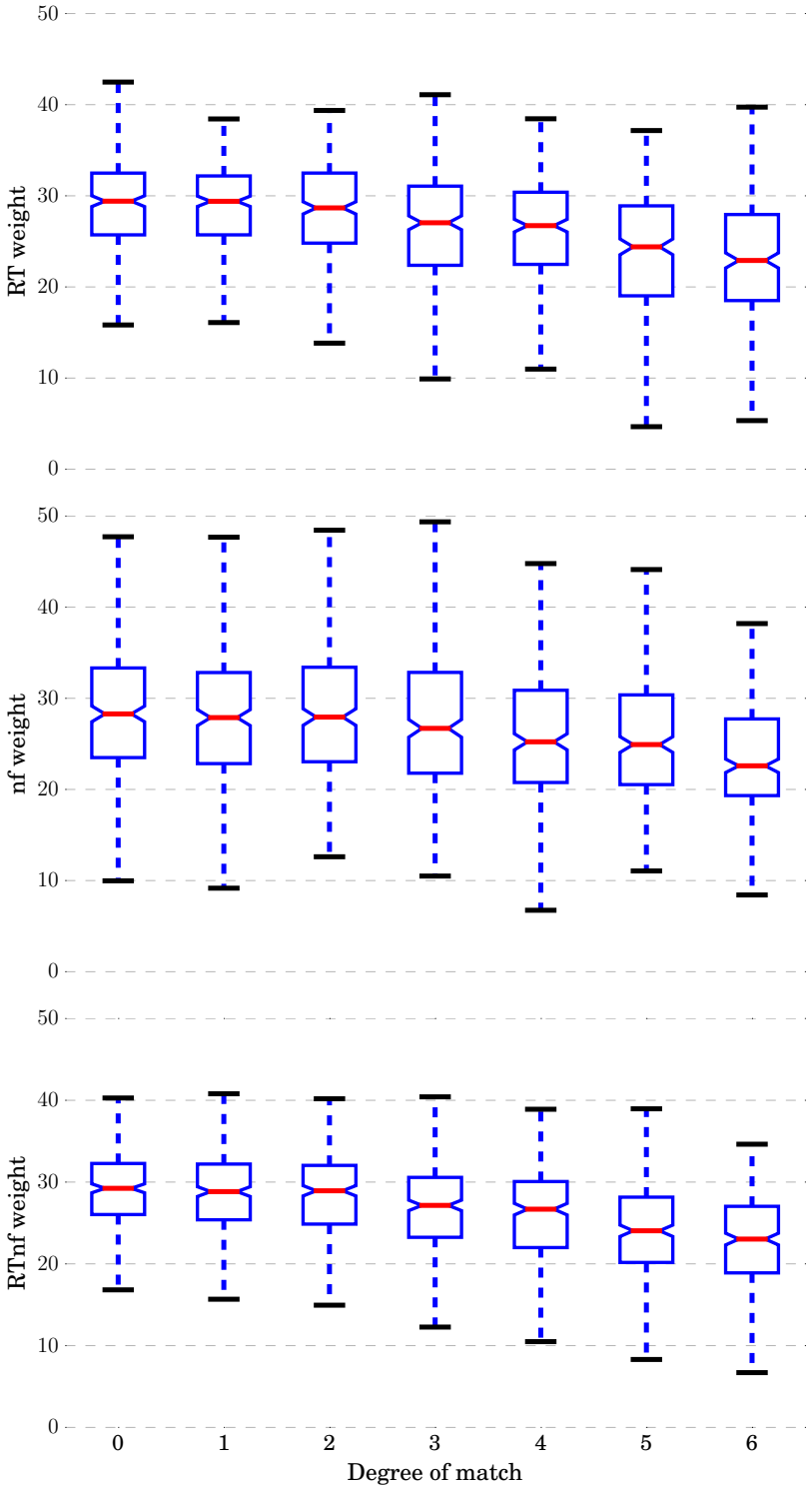

Figure 4.5: Box plots representing the distributions of the confidence weights for different DoM when the decision confidence is estimated using only RTs (top), only the neural features (middle) or the combination of them (bottom).

Table 4.3: Medians (across all participants) of the confidence weights associated to behavioural, neural and neuro-behavioural methods, as a function of the degree of match (DoM), of the pair of stimuli used in a trial.

DoM	*RT*	*nf*	*RTnf*
0	29.410	28.286	29.236
1	29.396	27.881	28.836
2	28.673	27.939	28.936
3	27.041	26.701	27.151
4	26.726	25.224	26.686
5	24.399	24.923	24.045
6	22.904	22.591	23.030

4.3.3 Group Decisions

To test the performance obtained using our cBCI framework, as described in 3.9, we compared the performance of single observer decisions (presented in Section 4.3.1) with group decisions made by groups of increasing size for all possible memberships of the groups. With our 10 participants, we had $\binom{10}{m}$ groups of size m.

For each group size we computed the average error rates for when the majority rule was applied, and the error rates of the three confidence-based methods described before (i.e., *RT*, *nf* and *RTnf*). The results are shown in Figure 4.6. The data are also reported in numerical form in Table 4.4. As one can see, in all methods studied except when using majority rule for groups of size 2, group decisions were superior to the decisions of single observers (the statistical significance is studied later), suggesting that integration of perceptual information across non-communicating observers is possible and beneficial.

trials where the observer was wrong (confidence lower than the median).

Table 4.4: Average error rates (%) *vs* group size for the four methods used to obtain group decisions. The minimum error rate for each group size is shown in bold face.

Group Size	*Majority*	*RT*	*nf*	*RTnf*
1	**12.50**	**12.50**	**12.50**	**12.50**
2	12.50	10.27	10.41	**9.74**
3	7.23	**7.16**	7.36	7.18
4	7.23	6.18	6.32	**5.96**
5	5.28	**5.10**	5.20	5.12
6	5.28	4.67	4.69	**4.57**
7	4.31	4.25	**4.13**	4.18
8	4.31	3.92	**3.67**	3.95
9	3.79	3.92	**3.52**	3.79
10	3.79	3.12	**2.67**	3.12

These results also show that the straight majority is generally outperformed by the other three methods. This is particularly evident with groups having an even number of members where the coin-tossing required by majority rule in the presence of ties implies that performance is the same as that of groups with one fewer member. The data also show that of the three other methods, the *RTnf*-based method appears to be the most consistent, being best or second best in 9 out of 10 cases. Furthermore, the performance of groups of large sizes (from 7 upward) starts saturating, possibly to a worse asymptote than the performance of the methods based on confidence correlates.

It is also interesting to note that while performance of the *nf*-based method appears to be inferior to *RT*-based and *RTnf*-based methods for groups of sizes 2 to 6, it is the best method for groups of 7, 8, 9 and 10 members. This suggests that our choice of coefficients in Equation (4.2), while making *RTnf* a generally good all-rounder, may have been suboptimal for the larger groups. This issue should be explored in future research.

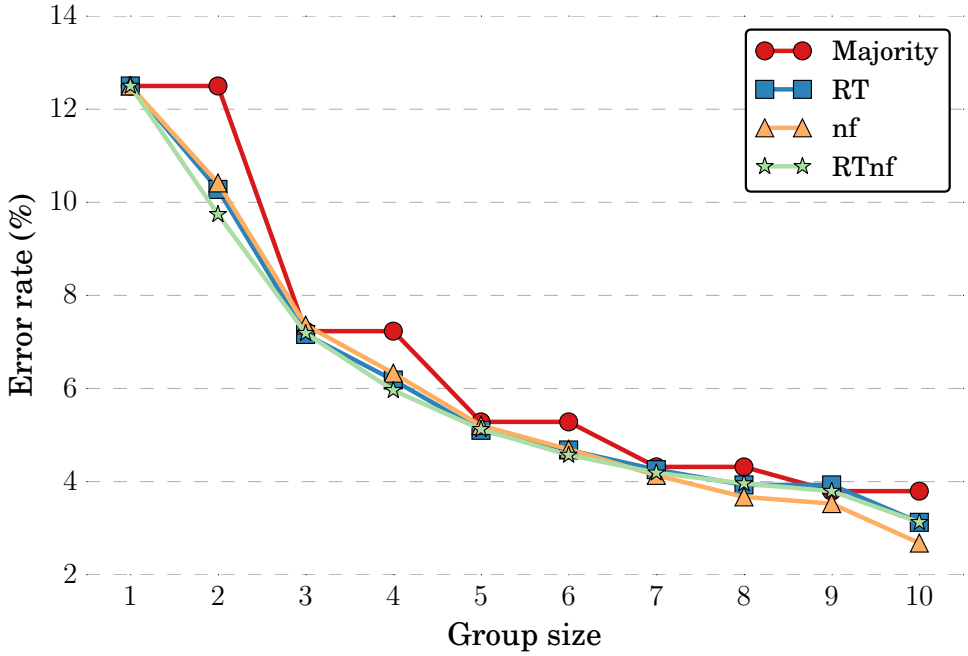

Figure 4.6: Average percentage of errors for different group sizes for the four methods for group decisions tested in this study.

To test if the observed differences in error rates in Figure 4.6 and Table 4.4 are statistically significant, we compared the *distributions* of errors made. We started by comparing the error distributions of single observers with those of groups of increasing size (for the four methods of group decision tested) using the Kruskal-Wallis statistical test. Table 4.5 reports the p-values and statistics returned by the test. This shows that for groups of size 2, the $RTnf$-based method is very close to be statistically significantly better than single observers, while for the RT- and nf-based methods the overlap of the distributions and sample sizes are such that statistical significance is not achieved despite the performance of all methods being on average 2 to 3% better than the single observers' case (as shown in Figure 4.6). On the contrary, for groups of size from 3 to 9 group decisions are always significantly superior to single observers. Finally, we should

Table 4.5: p-values and corresponding H statistics (in brackets) returned by the Kruskal-Wallis test when comparing the performance of single observers against the performance of groups of increasing sizes and adopting different decision methods. Sample sizes are reported in the second column. p-values below 0.05 are in bold face.

Group size	*Samples*	*Majority*	*RT*	*nf*	*RTnf*
2	45	0.751561 (0.1)	0.088386 (2.9)	0.274314 (1.1)	0.050447 (3.8)
3	120	**0.000094** (15.2)	**0.000080** (15.5)	**0.000077** (15.6)	**0.000070** (15.8)
4	210	**0.000065** (15.9)	**0.000009** (19.7)	**0.000011** (19.3)	**0.000006** (20.5)
5	252	**0.000002** (22.4)	**0.000002** (23.0)	**0.000002** (22.6)	**0.000002** (22.9)
6	210	**0.000003** (21.7)	**0.000001** (24.1)	**0.000001** (24.2)	**0.000001** (24.5)
7	120	**0.000001** (24.9)	**0.000001** (24.9)	**0.000000** (25.6)	**0.000000** (25.5)
8	45	**0.000002** (22.4)	**0.000002** (23.0)	**0.000001** (23.3)	**0.000002** (23.0)
9	10	**0.000174** (14.0)	**0.000172** (14.1)	**0.000146** (14.4)	**0.000146** (14.4)
10	1	0.113024 (2.5)	0.113024 (2.5)	0.113024 (2.5)	0.113024 (2.5)

note that our group of size 10 is, unsurprisingly, not significantly superior to single observers, even though its performance is superior to *all* the single observers ones (see Figure 4.3), due to it being a sample of just one data point.

We then compared the error distributions across the group-decision methods *within* each group size. Since errors are paired in each comparison (by the fact that the two methods being compared were applied to exactly the same groups), here we used the one-tailed Wilcoxon signed-rank test. The corresponding p-values and statistics are reported in Table 4.6.

As expected, we found that several of the small differences shown in Figure 4.6 and Table 4.4 are not significant. To make it easier to see which differences were significant, we summarise the p-values obtained in our tests using the statistical-significance preference-relation diagram shown in Figure 4.7. Groups of size 1 (all methods performing the same) and 10 (where we only have one such group) are not reported as no difference is statistically significant. For other groups

Table 4.6: p-values and corresponding W statistics (in brackets) returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes adopting the four decision methods considered in the study. Samples sizes are indicated in the last row of the table. p-values below 0.05 are in bold face.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is *RT* better than Majority?	**0.0000** (83)	0.1518 (1102)	**0.0000** (1086)	**0.0000** (5441)	**0.0000** (2068)	0.1790 (1324)	**0.0002** (147)	0.7813 (14)
Is *nf* better than Majority?	**0.0000** (60)	0.9923 (1913)	**0.0000** (1966)	**0.0443** (7240)	**0.0000** (2006)	**0.0001** (902)	**0.0000** (35)	0.0625 (0)
Is *nf* better than *RT*?	0.7981 (519)	0.9911 (2735)	0.9678 (9132)	0.9707 (11577)	0.4837 (8298)	0.0506 (1647)	**0.0184** (244)	0.0625 (5)
Is *RTnf* better than Majority?	**0.0000** (7)	0.1634 (631)	**0.0000** (444)	**0.0000** (2826)	**0.0000** (1441)	**0.0039** (920)	**0.0000** (133)	0.6875 (5)
Is *RTnf* better than *RT*?	**0.0133** (207)	0.8182 (1470)	**0.0000** (3721)	0.8166 (5882)	**0.0014** (2594)	0.0538 (765)	0.6754 (204)	0.2813 (5)
Is *RTnf* better than *nf*?	**0.0283** (316)	**0.0014** (1081)	**0.0000** (4647)	**0.0412** (7033)	**0.0362** (6284)	0.8537 (1790)	0.9978 (431)	1.0000 (10)
Sample size	45	120	210	252	210	120	45	10

sizes, while at one end of the spectrum we see that majority is almost always the worst method of the four, at the other end we see that the *RTnf*-based method is statistically superior to majority in 6 out of 8 group sizes. Moreover, *RTnf* is superior to the *RT*-based method in 3 out of 8 group sizes and is superior to the *nf*-based method in 5 out of 8 cases. Both the *nf*-based and *RT*-based methods are also competitive against majority. In particular, *nf* is superior to majority 6 times and almost statistically superior one further time (being inferior to it only for groups of size 3).

Nonetheless, one would probably choose the *RT*-based method if group sizes were small or if there was not a need for the slightly better performance afforded by *nf* for larger groups. This is because, of course, using *RT* on its own to measure the confidence does not require the use of a BCI, with its associated

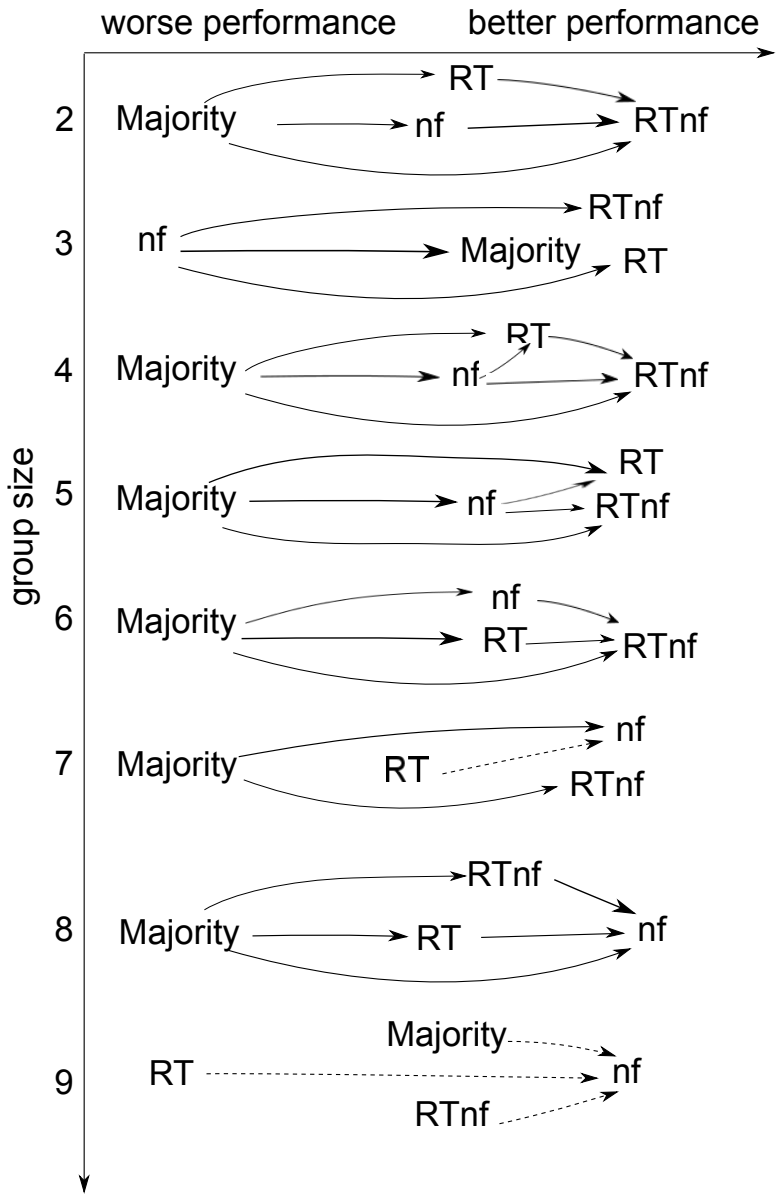

Figure 4.7: Statistical preference-relation diagram representing the results reported in Table 4.6 graphically. For each group size, a one-tailed Wilcoxon signed-rank test was executed, comparing the performance obtained with different decision methods. Solid arrows indicate that the method at the arrow-head is statistically superior to the method at the other end of the arrow (p-value lower than 0.01) while dashed arrows indicate near statistical significance ($0.01 \leq p < 0.05$).

and obvious drawbacks in terms of practicality and setup time. However, if top performance is required, the *RTnf*-based method seems to be the overall leader, although had we been able to test larger groups it is likely that the *nf*-based method would have potentially resulted top.

We should note that the results obtained by using *nf* and *RTnf* to measure the decision confidence are influenced very little by the number of folds chosen for cross-validation (while, of course, the results of majority and the *RT*-based method are exactly the same for any choice of folds as no learning process takes place in such methods). To illustrate this, in Figure 4.8 we report the error rates for the *RTnf*-based method as a function of group size and number of folds. A statistical comparison of the performance obtained with different numbers of folds using the Wilcoxon exact test with Bonferroni correction showed that in only 13.8% of the 550 comparisons required for a full analysis[1] differences were statistically significant. Also, for most group sizes the differences are very small. This suggests that the case of 16 folds on which we focused in most of the chapter is reasonably representative.

Let us now focus on decision times. In Figure 4.9 we report the average time required by groups of each size to make a decision after the presentation of the second stimulus set. Since all groups members must have made their decision before the group can make a choice, a group's response time is the maximum RT across group members. Unsurprisingly, the higher accuracy shown by bigger groups in Figure 4.6 comes at the cost of an increased group response time. In most cases it is unlikely that waiting an extra few hundreds milliseconds would be a problem, but in some circumstances (e.g., in critical decision making)

[1] With 11 numbers of folds and 10 group sizes, there are $10 \times \binom{11}{2} = 550$ pairwise comparisons.

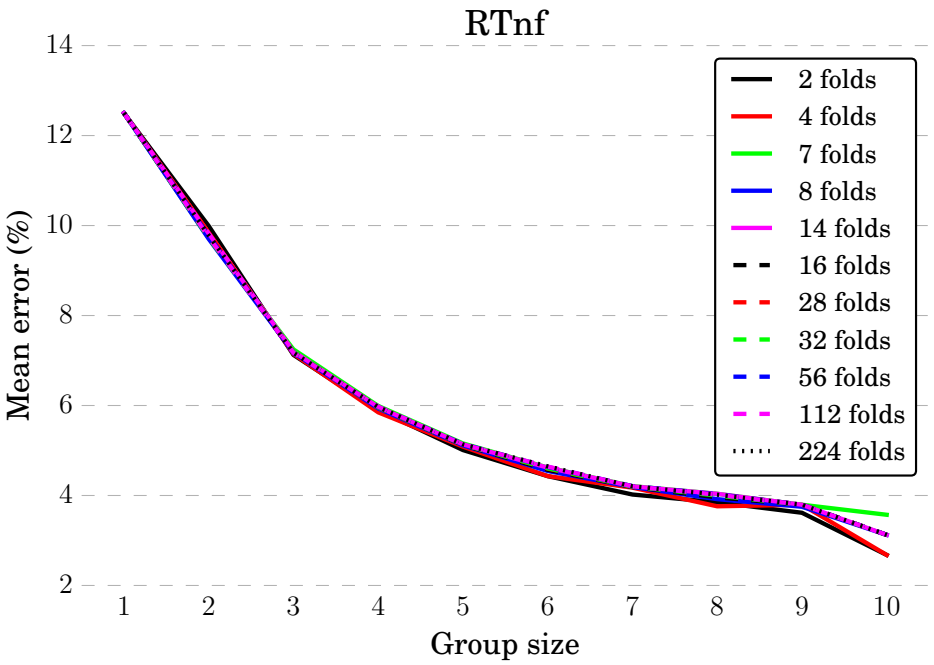

Figure 4.8: Average percentage of errors *vs* group size and number of cross-validation folds for group decisions made with the *RTnf*-based method.

minimising the decision time could be vital. Hence, in the next section we show a strategy to reduce the group response time with a minor impact on accuracy.

The improvement in performance seen in groups of increasing size in Figure 4.6 might simply be due to the increased likelihood of inclusion of the top-performing participants in the larger groups. For instance, our top performer, participant 4, will only be included in 20% of the groups of size 2, in 50% the groups of size 5 and 90% of the groups of size 9. It is possible that the presence of that participant in a group would be sufficient to drive the error rate of the groups downward significantly. In principle, it might be the case that groups do not perform better than their best member. Of course, we know that this is not the case, at least for groups of size 6 or above, simply because the group error rates are *below* the error rate of our top participant. However, to investigate this issue more thoroughly,

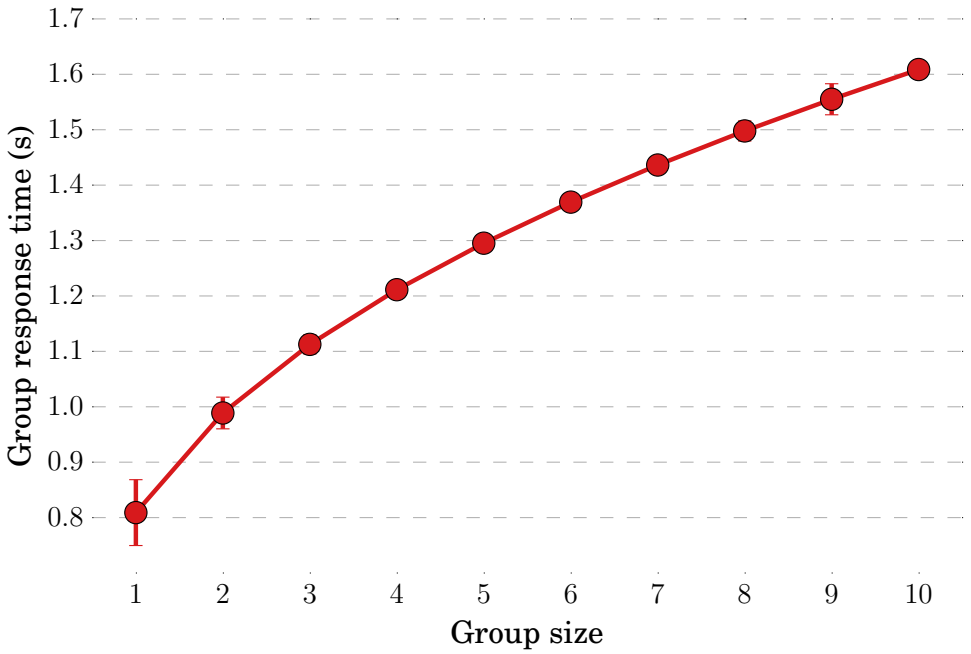

Figure 4.9: Average time required for groups of each size to make a decision. The error bars show the standard error of the mean.

for each group of a given size, we have compared the performance of the group obtained by our *RTnf*-based method to that of its best individual performer. Figure 4.10 reports the median difference in error rates between the two, for each group size. The figure makes it quite clear that group decisions are to a significant extent the result of a process of integration of confidence across participants, and not only the result of top performers driving group errors down.

4.3.4 Performance of Fastest Responders

Let us further investigate the relationship between performance and response times. As expected from the literature [100], also in our experiment there is a relationship between the relative speed with which observers give their response

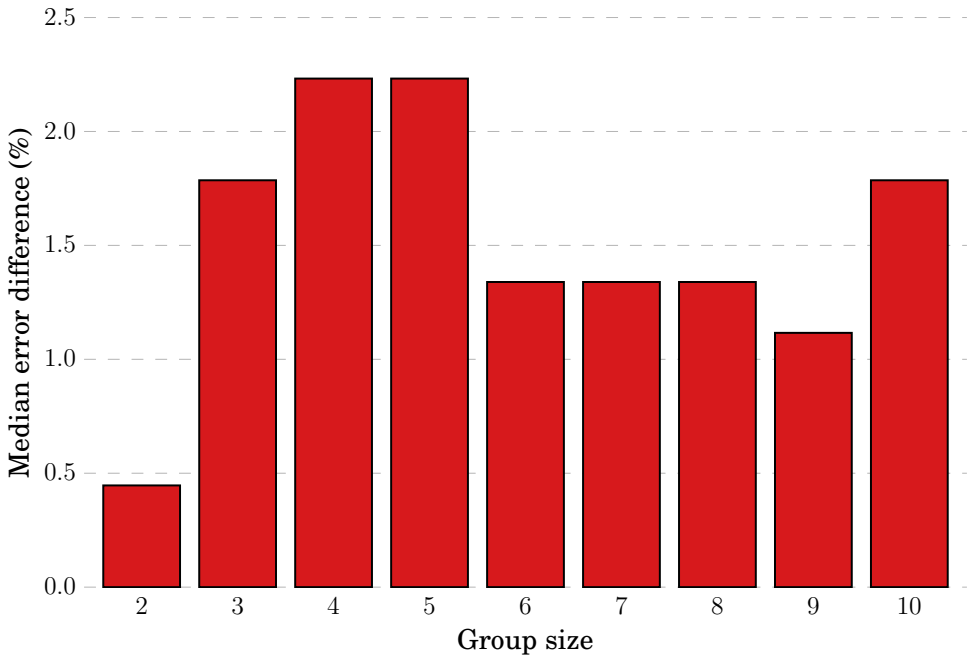

Figure 4.10: Medians of the differences in error rates between group decisions made with *RTnf* and decisions taken by the best performer in each group. Positive values indicate the extent to which groups were *better* than their best performers.

and the correctness of the decisions, with faster respondents being on average correct more often than slower ones (see Figure 4.4(top right)). Also, as we have seen in Figure 4.9 the larger a group the longer the delay in getting the group's response. So, *we wondered whether we could improve group decision times with relatively little impact on group accuracy if we allowed only the faster responders in a group to influence the group's decision*, as described in Section 4.2.5. In particular, we considered groups of all sizes and for each size we looked at what level of performance could be achieved by making decisions based on the fastest respondent, the two fastest respondents, and so on, in each trial.

Figure 4.11 compares the accuracies obtained with different groups sizes (and different sub-group sizes) with the corresponding response times for a group. More specifically, Figure 4.11(top) shows a plot of the mean group response time *vs* the mean group error rate for each group size when using the majority method. In the plot, circles of different diameters represent different numbers of fastest responders ("# voters" in the figure) from each group which were allowed to vote. That is, with the exception of the largest circle on each line (which represents the error vs RT trade-off for groups where everyone votes), only the decision of the fastest subgroup were used to determine group decisions. Figure 4.11(bottom) reports the corresponding results for the $RTnf$-based method. Let us analyse these data.

Firstly, results confirm that the fastest respondents ("# voters=1") tend to be the most accurate. On average a single observer has an error rate of 12.5% (see data point for the "group size=1" case) while selecting the response of the fastest performer in each trial produces an error rate of less than 8% for groups of size 5 or above (irrespective of decision method). Of course, the larger the group

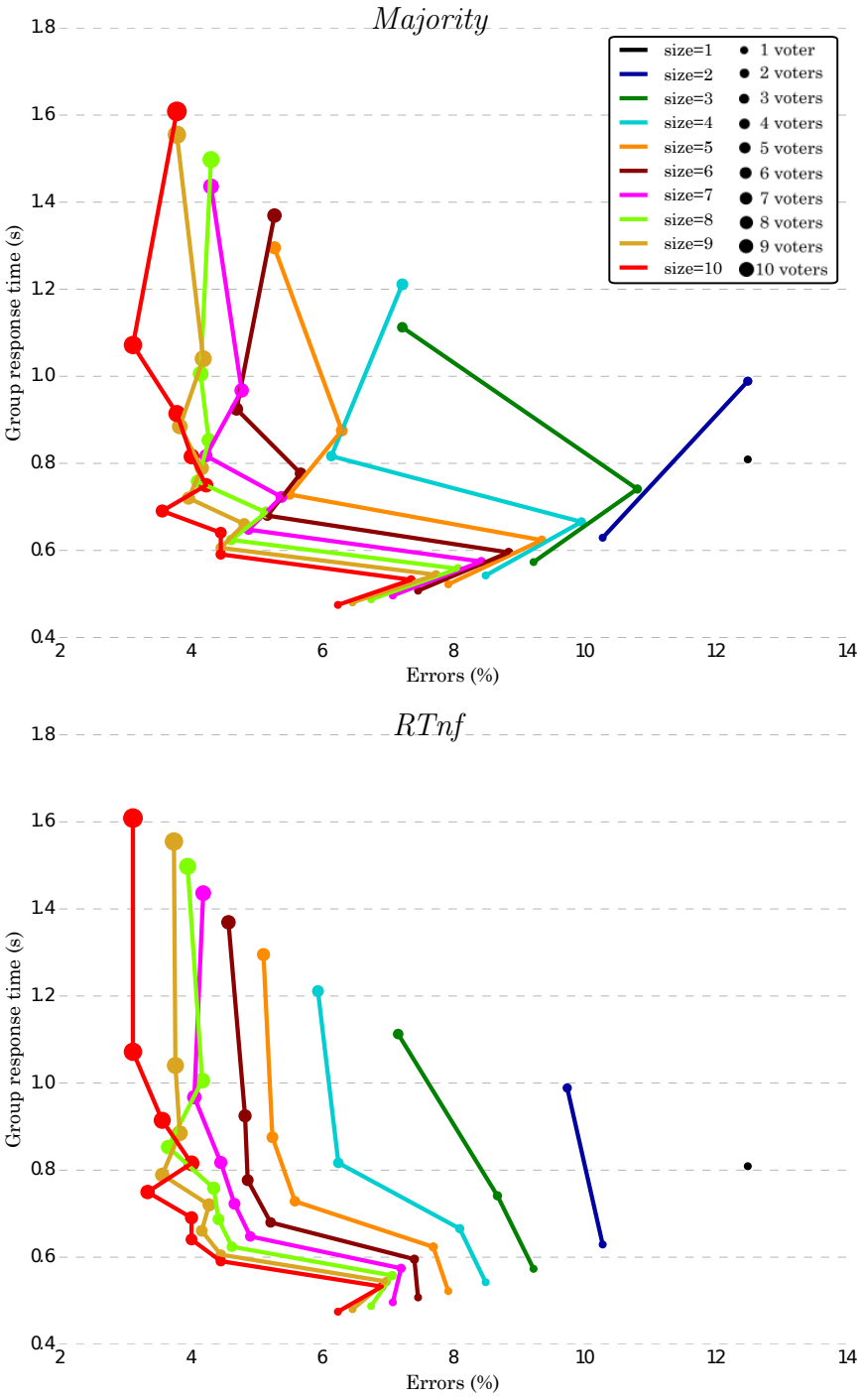

Figure 4.11: Comparison of the accuracies obtained with different groups' sizes and different numbers of voters from within a group against the corresponding response times for the group when using the majority (top) and *RTnf* (bottom) group-decision rules. Each line colour represents a group size. Circles of different diameters represent different numbers of fastest responders ("# voters") from each group which were allowed to vote.

considered the shorter the response time of the fastest respondent. So, fastest respondents for groups of sizes 9 take 480 ms on average to make a decision, while the full group takes approximately three times longer (1550 ms).

Secondly, we see that for the majority method there is no gain in using fastest-pair ("# voters=2") decisions over fastest-respondent decisions ("# voters=1"), as the former are both slower and more error-prone than the latter. On the contrary, for the *RTnf*-based method, we see that fastest pairs are almost always more accurate (but slower) than single fastest respondents. For instance, for groups of size 3, single fastest respondents make decisions in 560 ms while pairs take 730 ms. However, while the error rate for fastest respondents is the same (9.2%) for majority and *RTnf*, the error rate for the fastest pair is 10.8% for majority but only 8.6% for the *RTnf*-based method.

Thirdly, we see that when only the fastest triplet of observers ("# voters=3") is allowed to make a decision, there is a very marked improvement in accuracy over pairs or single fastest respondents for both majority and the *RTnf*-based method for all group sizes. The benefits of such a scheme are particularly clear for larger groups where the fastest triplet's response is much faster compared with the full group response, while the accuracy is significantly better than for pairs or single fastest respondents. For instance, for groups of size 9, the fastest triplet has an error rate of 4.4% and a response time of 610 ms for both majority and the *RTnf*-based method.

Fourthly, for fastest subgroups of four observers ("# voters=4") we see a similar situation to that of the fastest pairs. That is, one never gains from using the fastest four observers to make a decision with majority rule, as accuracy is worse than for the three fastest observers and speed is slower. However, with the

RTnf-based method we see that, for groups of size 4, 5, 6 and 7, the four fastest observers are more accurate (but obviously slower) than any smaller subgroup. This behaviour seems to be present also at larger subgroup sizes.

Finally, this approach of considering only the fastest respondents for computing the group decision could also compensate the disadvantage in speed of using response-locked epochs, which require to collect neural data even *after* the response is provided. This is particularly useful in online systems, where real-time constraints apply.

4.3.5 ERP Analysis

We used two statistical tests to analyse our ERP data sets. To get an indication of the differences in the statistical distributions of ERPs for correct and incorrect responses, we grouped all ERPs (irrespective of the participant they pertained to) into two corresponding sets. We then applied the Kruskal-Wallis test to compare the voltages measured in each channel at each time step in the two data sets.

We also performed a two-tailed Wilcoxon signed-rank test for paired samples to compare the mean ERPs obtained on an individual basis. It should be noted that, for the central-limit theorem, means tend to be distributed according to a normal distribution. So, in principle one could also use a paired-sample t-test to perform this comparison. We performed both this test and the Wilcoxon test (which does not assume normal distribution) on our data. Differences in p-values were minimal. Here we prefer to report only the results of the statistically-weaker Wilcoxon test as this relies on fewer assumptions.

Figure 4.12 shows the stimulus-locked grand averages (averages of individ-

ual averages) of the ERPs recorded in our experiment for correct and incorrect responses for channels Fz, Cz, Pz, Oz, C3, C4, P5 and P6 and the p-values of the statistical tests comparing the signals for correct and incorrect trials in the period immediately following the onset of stimulus Set 2. Figure 4.13 shows corresponding response-locked grand averages.

If we look at the grand averages in Figure 4.12, we see that generally there are seemingly small differences between the ERPs for correct and incorrect trials. Differences do exist, however, particularly in the region where the P300 wave peaks (approximately 500 ms after the presentation of Set 2) and for central and posterior electrodes in the right hemisphere, i.e., Cz, Pz, C4 and P4. Similar differences are present in many other channels in the same regions, as shown in Figure 4.14(left) which shows a snapshot of the scalp potentials recorded 500 ms after the presentation of the stimulus (in a stimulus-locked reference system).

If we look at the response-locked grand averages in Figure 4.13, however, we see much larger differences between the correct and incorrect responses in all 8 channels shown, either in the period preceding the response or during it or in both, with most of these differences being highly statistically significant. Similar differences are present in most other channels, as shown in Figure 4.14 which shows snapshots of the scalp potentials recorded 500 ms before the response (centre) and at the response (right).

We should note that a response-locked reference system amplifies the differences in the duration of the memory-retrieval and decision phases following the presentation of the stimulus for the two conditions. More specifically, P300s start approximately 600 ms before the response for incorrect decisions and approximately 400 ms before the response for correct decisions (as the corresponding

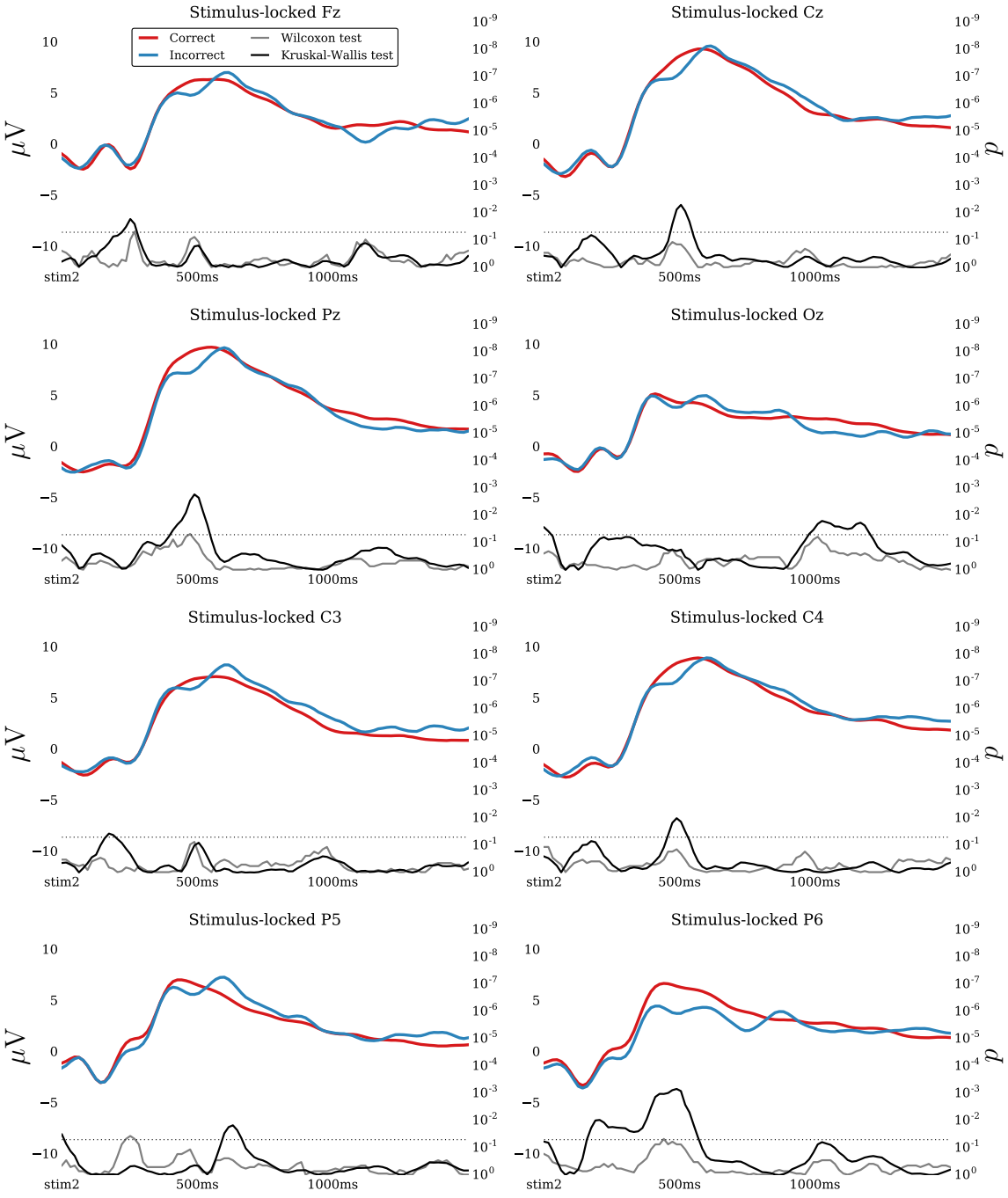

Figure 4.12: Stimulus-locked grand averages of the EEG activity (in μV) for channels Fz, Cz, Pz, Oz, C3, C4, P5 and P6 and corresponding temporal profile of the p-values of the Wilcoxon signed rank test comparing participant-by-participant averages (grey) and of the Kruskal-Wallis test for all ERPs recorded, irrespective of participant (black), in each error class. The dotted lines represent the 5% confidence level. The corresponding axes are oriented so that values above that line indicate statistical significance and *vice versa*.

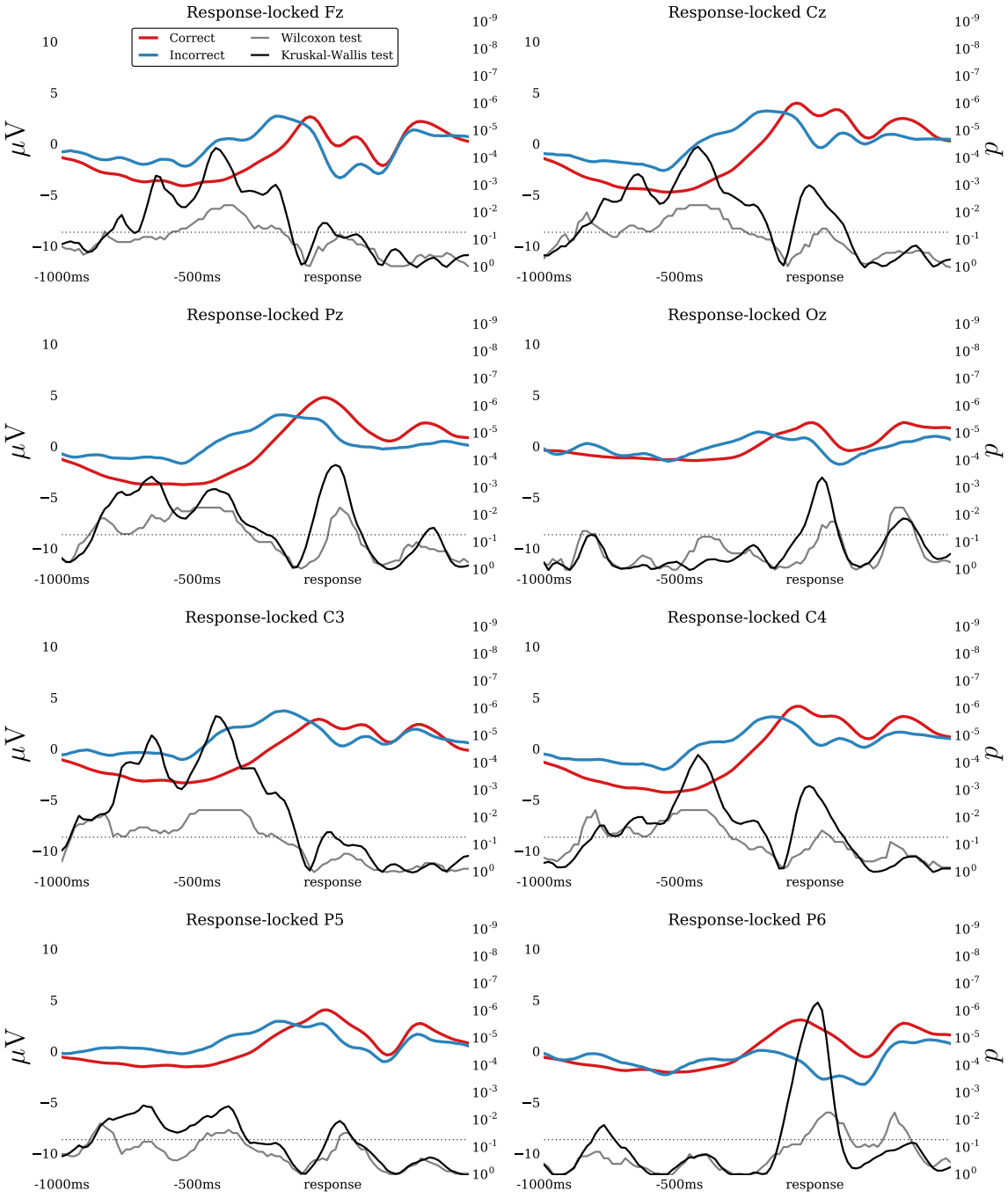

Figure 4.13: Response-locked grand averages of the EEG activity (in μV) for channels Fz, Cz, Pz, Oz, C3, C4, P5 and P6 and corresponding temporal profile of the p-values of the Wilcoxon signed rank test comparing participant-by-participant averages (grey) and of the Kruskal-Wallis test for all ERPs recorded, irrespective of participant (black), in each error class. The dotted lines represent the 5% confidence level. The corresponding axes are oriented so that values above that line indicate statistical significance and *vice versa*.

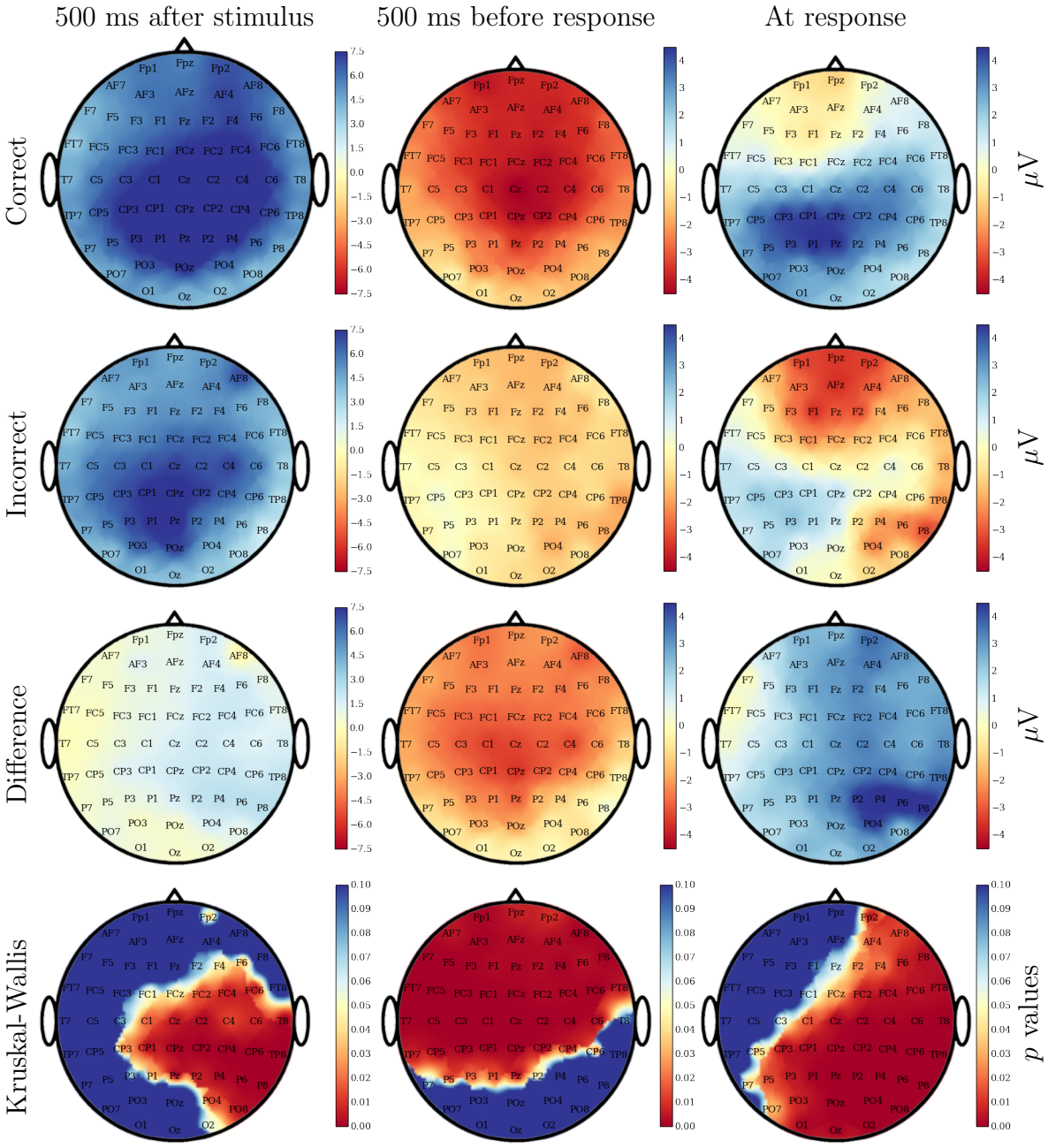

Figure 4.14: Scalp maps representing the grand averages of the EEG activity (in μV) recorded 500 ms after the presentation of the second stimulus, as represented by the stimulus-locked epochs (left), and 500 ms before the response (centre) and at the response (right), as represented by the response-locked epochs. Rows represent the activity for correct and incorrect trials (first two rows), the difference between them (third row) and the corresponding p-values of the Kruskal-Wallis test used to compare the two sets.

median response times are approximately 880 ms and 690 ms, respectively). They peak at approximately 400 ms and 200 ms before the response, respectively. This temporal shift and the small differences in P300 amplitude seen in the stimulus-locked grand averages for the two conditions cause the large statistically significant differences observed in a response-locked reference system up to 150 ms before the response (see Figure 4.14(centre)).

4.4 Conclusions

This chapter has described the results obtained by groups of observers undertaking a visual matching task making decisions using the cBCI framework described in Chapter 3. To test our ideas in a suitably constrained environment, we used a particularly simple set of visual stimuli, which, however, were presented very briefly thereby making the matching task arduous. We compared group decisions with those made by single non-BCI users and identically-sized groups of non-BCI users. The approach we have taken is unusual in relation to previous studies on collaborative BCI as here we have exploited not only neural data but also behavioural measures of confidence to weigh group members' decisions on a decision-by-decision basis.

Experimental evidence gathered with 10 participants conclusively indicates that group decisions (whether BCI-assisted or not) are nearly always statistically significantly superior to single user decisions. Also, BCI-assisted group decisions obtained by weighting observers' responses via our nf-based and $RTnf$-based methods were almost always statistically better than those obtained by equally-sized (non-BCI) groups adopting the majority rule. These methods are

particularly beneficial to groups of an even size, where the standard majority rule is unable to reach a decision more accurate than random in the presence of ties.

We also analysed the relationship between performance and response times. As predicted, we found that faster individual RTs are associated with increased accuracy. We also found that the larger a group, the longer it takes to gather all the single decisions and give a group response, so that the advantage obtained by groups over a single observer in terms of accuracy is associated with a disadvantageous response time. Based on these observations, we considered a scheme where only the fastest respondents of each group influence the group's decision and found that this improves significantly the group's response time with very little or no cost in terms of accuracy, making groups not only more accurate but also faster than single observers.

Although there are many advantages of group decision making, difficulties in communication and interaction, strong leadership and group judgement biases can sometimes be obstacles, particularly when accurate and fast decisions have to be taken. Here we demonstrated that, for a simple visual matching task, the proposed cBCI framework achieves some of the benefits of groups decisions, namely error correction and knowledge/certainty integration, without requiring intra-group communication and, thereby, avoiding some of the potential weaknesses of group decision-making.

One of the aims of this thesis was to develop a method based on neural features to estimate the decision confidence of multiple participants and improve group performance. Several ERP components may be possibly used to predict the accuracy or confidence of one's response. We chose to include in our neural feature the ERPs in the proximity of the response (before and after it) by providing the

system with a 1500 ms response-locked window of EEG starting 1 s before the response. We found that this gives reliable information on decision confidence, but in the following chapters and in future research we will also explore other possibilities.

This chapter has illustrated a very first application of the proposed cBCI framework which has, inevitably, some limitations. For example, here observers performed a relatively simple visual matching task, which is nowhere as complex as those carried out in realistic decision-making situations. The following chapters of this thesis will investigate more demanding real-world scenarios, with different perceptual modalities (e.g., audio signals) and with more complex decisions. Furthermore, we will also investigate whether it is possible to extend our approach to decisions where the team members are exposed to different sources of information (unlike here, where they were exposed to exactly the same information) – see Chapter 8.

Chapter 5

Augmenting Group Performance in Visual Search

This chapter explores the possibility of using the proposed cBCI framework to augment visual search performance of groups of users. It describes the results obtained with two visual search experiments: (a) one using simple shapes (i.e., oriented and coloured rectangles) where the task consisted in spotting an irregular item, and (b) one using realistic stimuli (i.e., pictures of arctic environments) where participants had to spot the presence of a photorealistically-added polar bear. Most of the material in this chapter has been published in [198, 199].

5.1 Introduction

One of the most important tasks of the visual system is to perform visual search, namely to scan the environment in search for an item of interest. We perform this task multiple times per day but, despite evolution, humans still find it taxing

and difficult – see Section 2.6.

The promising results obtained with simple visual matching task described in Chapter 4 encouraged us in exploring the possibility of using our cBCI to make group visual search more accurate, as well as applying that system to more complex visual tasks. This would allow us to validate those results in a different context, while other cBCIs have generally been validated with only one task – see Section 2.5. Visual search is a task that is perceptually and cognitively different from the visual matching task previously tested. The high perceptual load (due to the large number of non-targets presented in each display), the difficulty of discriminating between targets and non-targets (due to the shared features between the target and the non-targets) and the fast presentation of each display render decisions very hard in this task. This chapter describes the results obtained in this investigation via two main studies.

In the first study, we designed a traditional visual search experiment (Experiment 1) in which participants were presented a display containing a set of vertical and horizontal, red and green rectangles (bars) for 250 ms and had to decide whether or not a vertical red bar (i.e., target) was present. This experiment used simple stimuli similar to the ones used in [103]. However, to make the task even more difficult for a single user, we reduced the pop-out effect by using a combination of features to identify a target (i.e., colour and orientation) instead of a single feature. Hence, it was not sufficient for a participant to search for red bars or vertical ones to identify the target in the display, but he/she had to focus on the two features together. We then used the framework described in Chapter 3 to obtain group decisions based on the confidence estimated from a combination of physiological and behavioural signals.

The second study aimed at moving towards real-life applications of the proposed cBCI. We designed a new experiment (Experiment 2) where the stimuli were realistic images representing arctic environments. In each display, a variable number of penguins (i.e., distractors) were present and, in target images, a polar bear was also present in a random (but realistic) location. Participants had to report whether or not they had seen a polar bear (i.e., target) in the display, which was presented for 250 ms. These displays aimed at simulating environments in which the target can camouflage, as this makes the visual search task more difficult and realistic (e.g., for threat detection).

Furthermore, this chapter describes the advances made in identifying the best set of behavioural and physiological correlates of the decision confidence. We used a combination of (a) neural features extracted from both stimulus- and response-locked EEG epochs via spatio-temporal common spatial patterns, (b) eye movement features extracted from both stimulus- and response-locked epochs recorded via an eye tracker, and (c) RTs, to estimate the decision confidence of the user and obtain superior group decisions. The choice of using both stimulus- and response-locked epochs was guided by previous studies showing that both are informative of the decision-making process [229, 143].

Eye movements have been studied for years because they seem to reveal many hidden information, such as mental workload [114] or personal emotions. Eye blinks are the rapid closing and reopening of the eyelid that a human performs several times every minute. Eye blink rate and duration are two of the most common used indicators for fatigue and workload. Researchers have used eye blinks to detect workload in many situations, including heavy professions like drivers [88, 156, 7] or air traffic controllers [13]. Therefore, eye movement sig-

nals could represent an additional source of information related to the decision confidence for our cBCI.

5.2 Methods

5.2.1 Participants

Ten healthy volunteers (average age 28.5 ± 6.0 years, 4 females) took part in both experiments. The order of the experiments was counterbalanced, so that five observers undertook Experiment 1 first and then Experiment 2, and the remaining observers undertook the experiments in the opposite order. All participants had normal or corrected-to-normal vision.

5.2.2 Stimuli and Tasks

We designed both experiments by using the *percent-correct* approach described in Section 2.6, so that the difficulty of the task was due to the lack of time available for scanning the whole image.

Each experiment consisted of 8 blocks of 40 trials each, for a total of 320 trials. Figure 5.1 shows the sequence of displays presented in a trial for Experiment 1 (top) and Experiment 2 (bottom), which they only differed in the image used as a stimulus and both followed the protocol described in Section 3.2. In both experiments, the stimulus was followed by a mask consisting in a black and white 24×14 checkerboard presented for 250 ms.

In Experiment 1, the stimulus consisted in a display containing a set of 40 bars, either green (RGB (0,1,0)) or red (RGB (1,0,0)), vertical or horizontal, on a

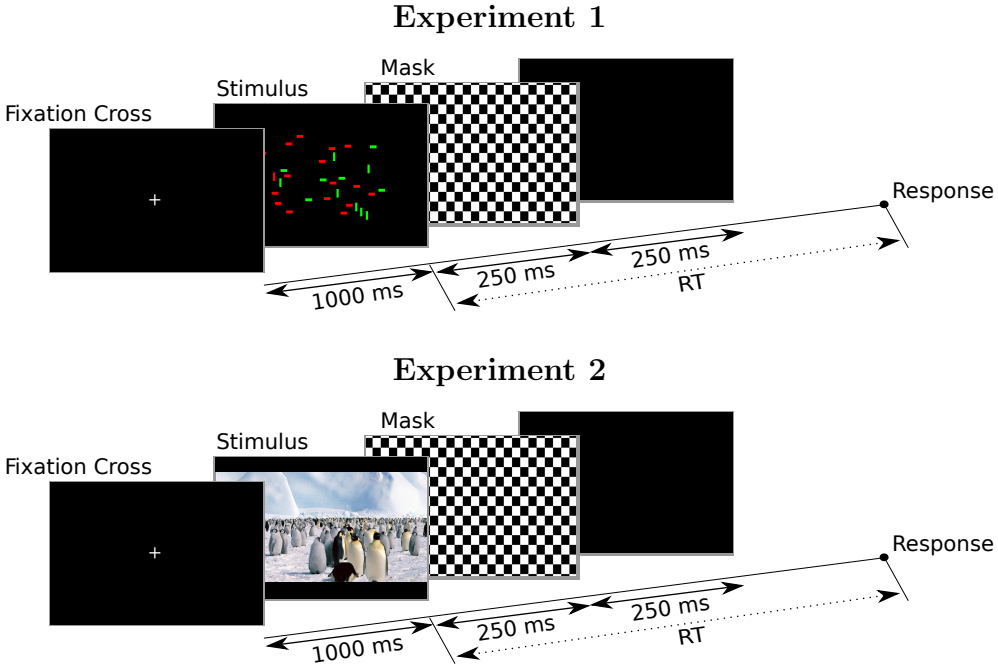

Figure 5.1: Sequence of displays presented in the Experiments 1 (top) and 2 (bottom).

black background, which was presented for 250 ms. Participants had to decide, as quickly as possible, whether or not there was a vertical red bar, the *target*, among the vertical green, horizontal green and horizontal red bars, the *distractors*.

The position of the bars was randomly selected (without allowing overlaps between bars) within a rectangular screen region subtending approximately 17.7 degrees horizontally and 11.9 degrees vertically. Bars subtended approximately 1.09 degrees in their longer dimensions and 0.36 degrees in their shorter dimension. The number of distractors of each type was also randomly selected, but ensuring that at least one instance of each type was present in the display. Sample displays with and without the target are shown in Figure 5.2(top).

For Experiment 2, we used a set of manually-created realistic images representing an arctic environment containing a variable number of penguins (distractors)

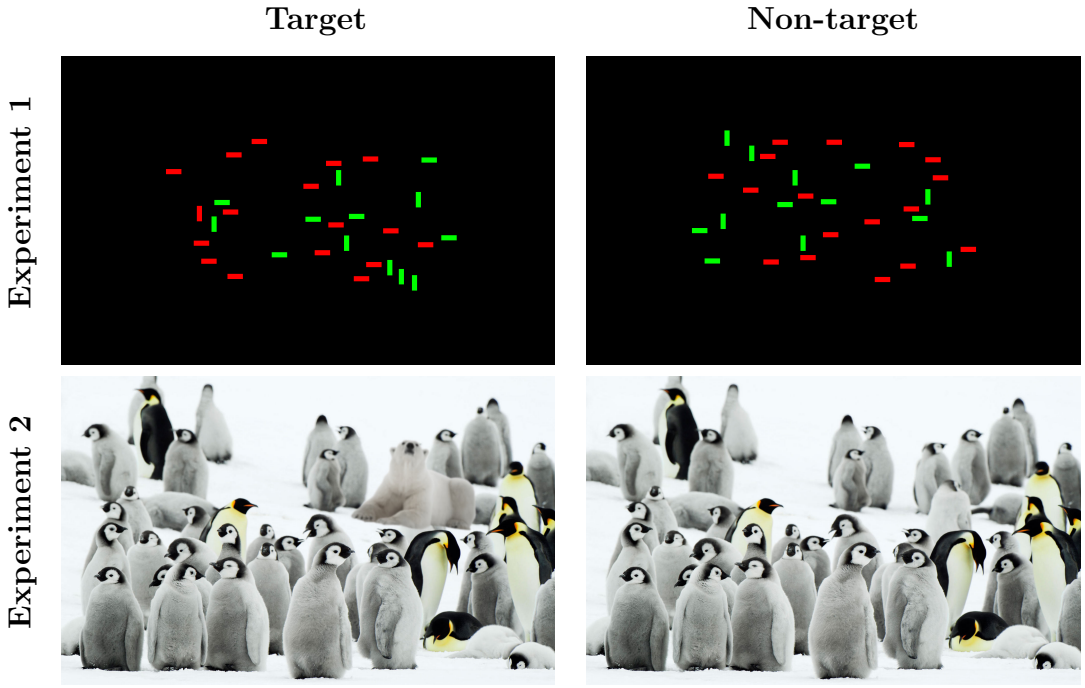

Figure 5.2: Examples of displays with and without the target (left and right, respectively) used in experiments 1 (top) and 2 (bottom).

and, possibly, a polar bear (target). We used five different arctic environments (backgrounds) as non-target stimuli. We then added two different bear pictures in four possible positions to each background to obtain 40 different images containing the target.

An example of a background (and non-target stimulus) and an example of a correspondent display containing the target are shown in Figure 5.2(bottom). Each image was displayed in full screen mode and subtended approximately 30.29 degrees horizontally and 19.22 degrees vertically.

The sequences of stimuli used in each experiment were randomly generated, stored and reused with all participants. This made it possible to test offline the benefits of combining the decisions of different participants to form group deci-

sions using the proposed cBCI without requiring to collect data in parallel. The stimuli containing the target were presented in 25% of trials of each experiment.

Participants were comfortably seated at about 80 cm from an LCD screen. Briefing, preparation of participants and task practice of both experiments (2 blocks of 10 trials each) took approximately 45 minutes, while the actual experiments lasted approximately 25 minutes each. The two experiments were undertaken on the same day with a break of a few minutes in between.

5.2.3 Data Acquisition and Transformation

Participants undertook experiments in conditions of complete absence of communication or any other form of social influence.

Data were acquired and preprocessed as explained in Chapter 3. As already mentioned in Chapter 3, for these experiments we set $p_b = 6$ Hz, $s_b = 8$ Hz and the final sampling rate $s_r = 16$ Hz. We have also verified that it is possible to slightly improve the classification performance of the cBCI by using $p_b = 14$ Hz, $s_b = 16$ Hz and $s_r = 32$ Hz. However, this has the significant disadvantage of tripling the feature extraction time.

For each participant, we have applied LTCCSP to the response- and stimulus-locked epochs of the training set separately to obtain two projection matrices, W_{Rlckd} and W_{Slckd}, respectively. The original epochs were then transformed using these matrices to obtain two new feature spaces where data are organised in such a way that the first and the last columns of each have the maximum and the minimum difference in terms of variance, respectively. The variances of the first and last columns of the response-locked and the stimulus-locked transformed

epochs (four features in total) were then used as neural features.

It should be noted that, as we show in Section 5.3, LTCCSP allows to achieve very good performance with only two features, while with PCA (used in the visual matching experiment described in Chapter 4) we needed 24 features to obtain good performance. This significant reduction in the number of features allowed us to use also the information extracted from the stimulus-locked epochs to better capture the brain activity correlated with the decision confidence.

As done in the visual matching experiment, we have also added the RT to the feature vector used by the cBCI to estimate the decision confidence.

To complement the neural and behavioural features, in these experiments we have also extracted four features from the vertical component of the eye movements recorded by the Jazz eye tracker associated to both stimulus- and response-locked epochs (see Chapter 3). These four features contain information about the occurrence of eye blinks and the activity of the eyes during the experiment.

5.2.4 Confidence Estimation and Group Decisions

The decision confidence of each participant was estimated by using all the neural, behavioural and physiological features described in the previous section. As discussed in Chapter 3, we split the available data into a training set, which was used to compute the LTCCSP matrices and fit the model used to predict the decision confidence (LARS), and a test set, which was used to evaluate the performance of the cBCI. In these experiments we used 10-fold cross-validation to reduce the risk of overfitting, as this choice of k guarantees that all the folds have the same number of samples (i.e., 288 in the training set and 32 in the test

set). The confidence estimates obtained from the data available in the test set of each participant i were then transformed into confidence weights w_i by using the negative exponential weighting function described in Equation (3.3).

For comparison, in Experiment 1 we have also estimated the decision confidence by using different subsets of the available types of features, namely RTs (as done in Chapter 4), RTs and eye features, and LTCCSP neural features and RTs. Moreover, to assess whether or not LTCCSP better identifies neural correlates of the decision confidence than PCA used previously, we have concatenated the stimulus- and response-locked epochs and extracted PCA features from the resulting epochs.[1] We then compared the performance obtained by a cBCI using these PCA features and the RTs with the performance of a cBCI using LTCCSP features (extracted separately from response- and stimulus-locked epochs) and RTs.

Considering the results obtained with Experiment 1 (reported in Section 5.3), for Experiment 2 we have only considered a cBCI estimating the decision confidence with LTCCSP neural features, RTs and eye movements, and a cBCI using only LTCCSP neural features and RTs.

It should be noted that in the cBCI used in Chapter 4, the confidence estimated by using both PCA neural features and RTs was obtained by training two different classifiers (one for each type of feature), the outputs of which were then combined to obtain a confidence estimator. However, we found that this added

[1] As discussed in Chapter 4, the high number of features required by PCA to achieve good performance (namely, 24) increased the risk of overfitting of the linear classifier used (LARS). For this reason, we previously decided to extract neural features from the response-locked epochs only. However, here, for a fair comparison, we decided to include both types of epoch in the analysis by concatenating them, so that the total number of features remains 24 but PCA and LTCCSP have the same information available to identify neural correlates of the decision confidence.

complexity was not necessary. Hence, here we trained a single linear model with all the features used by the cBCI, which further reduced the free parameters in our system.

Group decisions were then made as described in Section 3.8 by using the various confidence estimates analysed in this chapter and we compared the cBCI decisions with choices made by non-BCI groups using the standard majority rule.

5.3 Results

5.3.1 Individual Performance

Since the main aim of this thesis is to develop a cBCI to improve human performance, we start by looking at the errors of each participant in the two visual search tasks considered in this chapter. Figure 5.3 shows the error rates of each participant for Experiment 1 (left) and 2 (right). Observers had very different individual levels of performance. Moreover, the average error rate in both visual search experiments was higher than the average error rate achieved by participants of the visual matching task described in Chapter 4, confirming that these visual search tasks are much more taxing and difficult for individuals.

5.3.2 Group Performance in Experiment 1

Figure 5.4 shows the mean error rate for groups of different sizes in Experiment 1 making their decisions using the majority rule as well as the confidence-based methods analysed in this study. Table 5.1 provides a numerical representation of the same information.

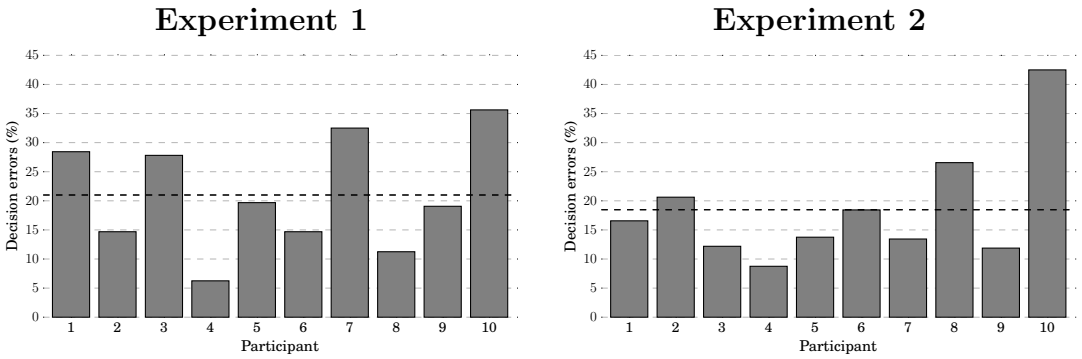

Figure 5.3: Error rates of participants of Experiment 1 (left) and 2 (right). The average error rate across the participants of each experiment is shown by the dashed black lines.

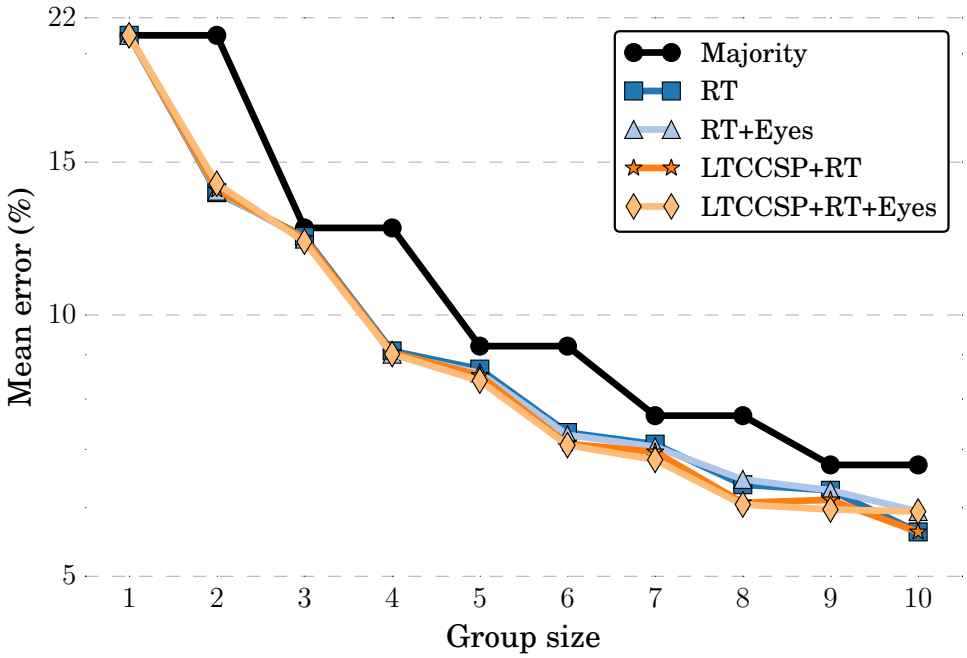

Figure 5.4: Average percentage of errors vs group size in Experiment 1 for group decisions made by: (1) the majority rule, (2) a RT-based decision system, (3) a RT- and eye-based decision system, (4) a cBCI using LTCCSP neural features and RTs, and (5) a cBCI using LTCCSP neural features, RTs and eye movements features. The y axis uses a logarithmic scale.

As we found in Chapter 4, also for a visual search task a reason why confidence-based decision-making rules outperform simple majority is that they meaningfully break ties (which are otherwise resolved with a random decision) in even-sized groups. Indeed, as we can see both in Table 5.1 and Figure 5.4, the difference in performance for such groups is usually much greater than for odd-sized groups. However, all our confidence-based systems, but particularly the cBCI based on LTCCSP, RTs and eye movements features, appear to augment human decision-making performance also with odd-sized groups.

We have also seen that, as found for the visual matching task (Chapter 4), the performance of the cBCI system using only behavioural features appears to be worse than when using a combination of neural and behavioural features for most group sizes (i.e., compare "RT" and "LTCCSP+RT" columns in Table 5.1).

The p-values of the Wilcoxon signed-rank tests performed to compare the error distributions across different methods are reported in Table 5.2. Sample sizes are indicated in the last row of the table. It is clear that for all group sizes our new LTCCSP-based cBCI yields group decisions that are significantly better than traditional (majority-based) group decisions. Also, for many group sizes such decisions are significantly better than those made by groups assisted by cBCIs using only a subset of the types of features available.

When analysing group decision times we found similar results to those obtained with the visual matching task (Chapter 4): groups increase decision times by up to 70% compared to individuals. However, as we did in Chapter 4, we verified that group RTs can be shortened by allowing only the fastest respondents to contribute in the group's decision (data not reported). With this technique, again there are many choices that allow cBCI-assisted groups to be both faster

Table 5.1: Tabular representation of the results in Figure 5.4. The best results for each group size are shown in boldface while the worst are in italics.

Group size	Majority	RT	RT+Eyes	LTCCSP+RT	LTCCSP+RT+Eyes
1	**21.00**	**21.00**	**21.00**	**21.00**	**21.00**
2	*21.00*	**13.83**	13.89	13.94	14.17
3	*12.60*	12.31	12.23	12.26	**12.15**
4	*12.60*	9.09	**9.01**	9.05	9.02
5	*9.21*	8.66	8.58	8.52	**8.40**
6	*9.21*	7.32	7.28	7.11	**7.08**
7	*7.66*	7.10	7.05	6.96	**6.81**
8	*7.66*	6.38	6.47	6.08	**6.05**
9	*6.72*	6.28	6.28	6.13	**5.97**
10	*6.72*	**5.62**	5.94	**5.62**	5.94

Table 5.2: Statistical comparison of methods for group decisions for different group sizes in Experiment 1. The table reports the p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes adopting the different decision methods analysed in this chapter. The number of groups of each size that could be assembled with 10 participants is indicated in the last row of the table. p-values below the statistical significance level 0.05 are in bold face.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is RT better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0063**
Is RT+Eyes better than RT?	0.6378	**0.0003**	**0.0128**	**0.0002**	0.0535	0.0544	0.9341	0.5562
Is LTCCSP+RT+Eyes better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0026**
Is LTCCSP+RT+Eyes better than RT?	0.9773	**0.0000**	**0.0288**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0264**
Is LTCCSP+RT+Eyes better than RT+Eyes?	0.9811	**0.0001**	0.4176	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0116**
Is LTCCSP+RT+Eyes better than LTCCSP+RT?	0.9548	**0.0001**	0.3212	**0.0000**	0.2416	**0.0000**	0.2810	0.0599
Sample size	45	120	210	252	210	120	45	10

Table 5.3: Tabular representation of the results in Figure 5.5. The best results for each group size are shown in boldface while the worst are in italics.

Group size	Majority	LTCCSP+RT	LTCCSP+RT+Eyes
1	**18.47**	**18.47**	**18.47**
2	*18.47*	13.49	**13.30**
3	*12.04*	12.00	**11.97**
4	*12.04*	9.94	**9.81**
5	*9.98*	9.90	**9.76**
6	*9.98*	8.69	**8.63**
7	*8.91*	8.76	**8.58**
8	*8.91*	8.22	**8.08**
9	*8.12*	7.94	**7.81**
10	*8.12*	7.81	**7.50**

and more accurate than single individuals. For instance, by allowing only the fastest 2 respondents in groups of 5 to decide in our LTCCSP-based cBCI, error rates are halved while RTs are approximately 200 ms shorter than for an average individual.

5.3.3 Group Performance in Experiment 2

Let us now analyse the performance of groups in Experiment 2, where the visual search task uses realistic stimuli.

Figure 5.5 shows the mean error rate of groups of different sizes in Experiment 2 making their decisions using the majority rule, a cBCI based on LTCCSP neural features and RTs, and a cBCI based on LTCCSP neural features, RTs and eye movements features. Table 5.3 provides a numerical representation of the same information.

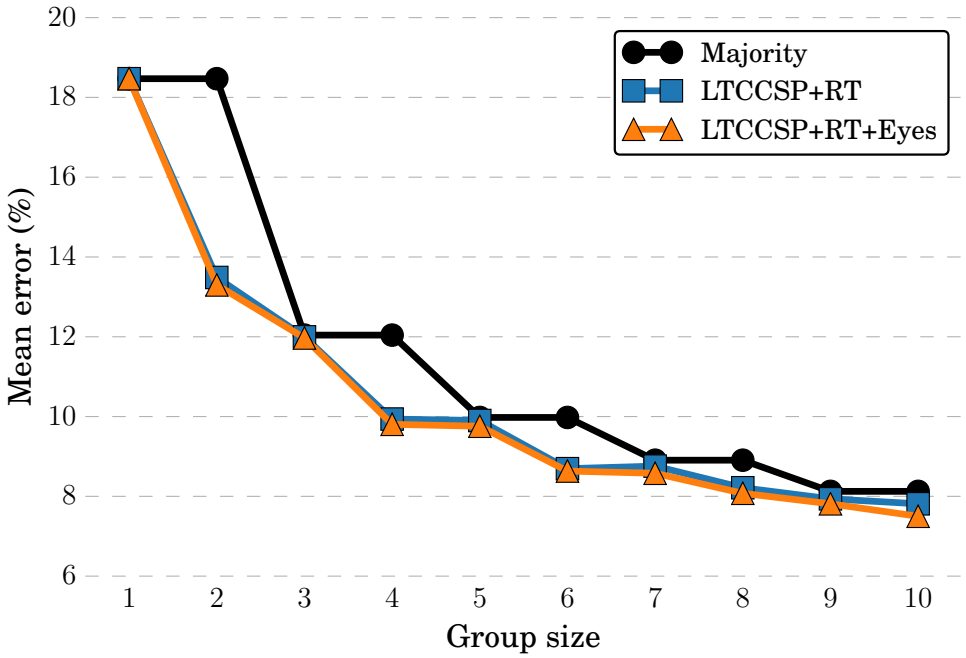

Figure 5.5: Error rates of groups of different size in Experiment 2 making decisions using (1) the majority rule, (2) a cBCI using LTCCSP neural features and RTs, and (3) a cBCI using LTCCSP neural features, RTs and eye movements features.

Table 5.4: Statistical comparison of methods for group decisions for different group sizes in Experiment 2. The table reports the p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes adopting the majority rule, a cBCI based on LTCCSP neural features and RTs, and a cBCI based on LTTCSP neural features, RTs and eye movements features. The number of groups of each size that could be assembled with 10 participants is indicated in the last row of the table. p-values below the statistical significance level 0.05 are in bold face.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is LTCCSP+RT better than Majority?	**0.0000**	**0.0118**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0354**
Is LTCCSP+RT+Eyes better than Majority?	**0.0000**	**0.0003**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0096**
Is LTCCSP+RT+Eyes better than LTCCSP+RT?	0.1332	**0.0195**	**0.0004**	**0.0000**	**0.0245**	**0.0000**	**0.0119**	**0.0359**
Sample size	45	120	210	252	210	120	45	10

These results confirm that the cBCI boosts group performance over traditional majority even in realistic visual search, making another important step towards bringing this cBCI out of the lab. Most of the reduction in error rates happens for even-sized groups, where the cBCI is able to break ties better than coin flipping (used by majority in case of ties). Moreover, the addition of eye movements features seems to slightly improve the performance of cBCI-assisted groups even more, especially for large groups.

To validate these differences statistically, we used the one-tailed Wilcoxon signed-rank test as done previously – see Section 5.3.2. The p-values are reported in Table 5.4. While both confidence-based methods are significantly superior than standard majority, the cBCI based also on eye movements features is significantly better than the cBCI based only on neural features and RTs for all group sizes 3–9. The two cBCIs perform on par for groups of size 2.

5.3.4 Group Performance Across Tasks

To gather some preliminary evidence on the degree of performance improvement that our cBCI can deliver across tasks, in Figure 5.6 we compare the results obtained with the less challenging visual matching task described in Chapter 4 and the results obtained by groups performing the visual search task studied in Experiment 1. In either case we report the results obtained with Majority (solid lines) and a cBCI using 24 PCA neural features extracted from the response-locked epochs and the RTs (dashed lines). We have plotted these data using a logarithmic scale as this makes it possible to compare the *relative* improvements across systems (equal distances along the ordinates correspond to equal improvement percentages). For reference, we also report the results obtained in the visual search task by our best method: the cBCI based on LTCCSP response- and stimulus-locked neural features, RTs and eye movements (black dotted line).

The most apparent feature in the figure is that the lines representing the visual matching task (blue) and those representing the visual search task (red) run almost parallel, indicating that both Majority and the PCA-based cBCI provide the same relative benefits as the group size varies. Of course the cBCI lines are below the Majority lines (as we have already discussed). However, the distances between the solid and the dashed lines of each colour follow a very similar profile. This indicates that the relative benefits obtained by the cBCI over Majority at each group size are comparable across the two tasks. Indeed, the average increase in performance across group sizes brought by the PCA-based cBCI is 8.6% for visual matching and 8.7% for visual search.

These results corroborate the hypothesis that the approach used by the cBCI

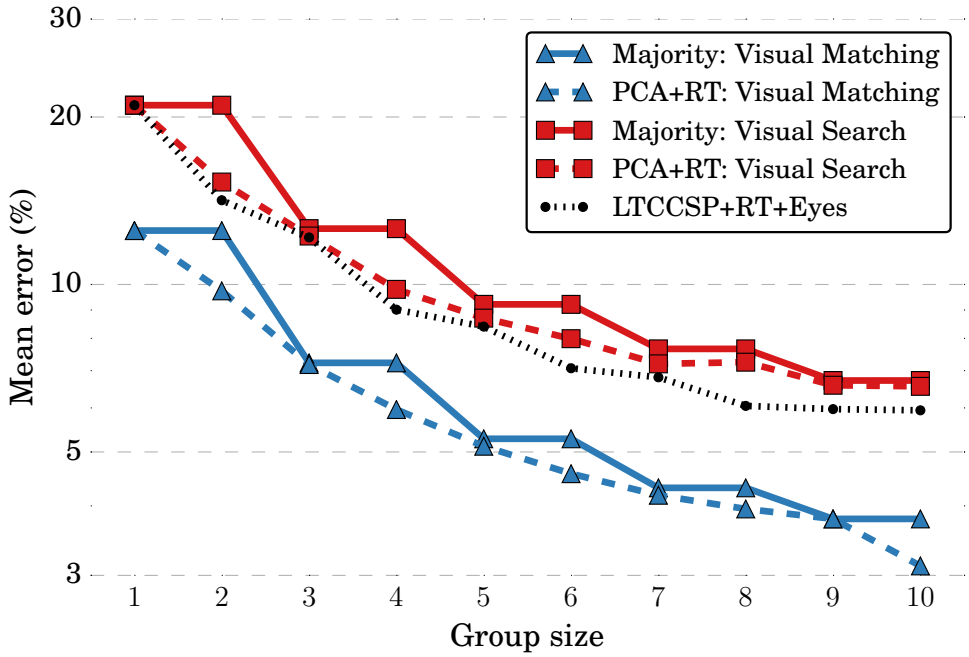

Figure 5.6: Comparison of the results obtained in Chapter 4 with a visual matching task and the results of the present work with the visual search task of Experiment 1 obtained with Majority and PCA-based cBCIs. The black dotted line represents the results of the cBCI based on LTCCSP, RT and eye features in the visual search task. The ordinate axis uses a logarithmic scale.

to obtain and exploit correlates of decision confidence generalises well to tasks of different nature and difficulty.

5.3.5 LTCCSP vs PCA Neural Features

One of the main contributions of this chapter was to replace PCA with LTCCSP as the method to extract the neural features. To further investigate the advantage provided by this choice, we compared the performance obtained by groups in Experiment 1 using a cBCI based on PCA neural features and RTs with the performance obtained by a cBCI using LTCCSP neural features and RTs. Figure 5.7 shows the error rates of groups of different sizes using the two methods

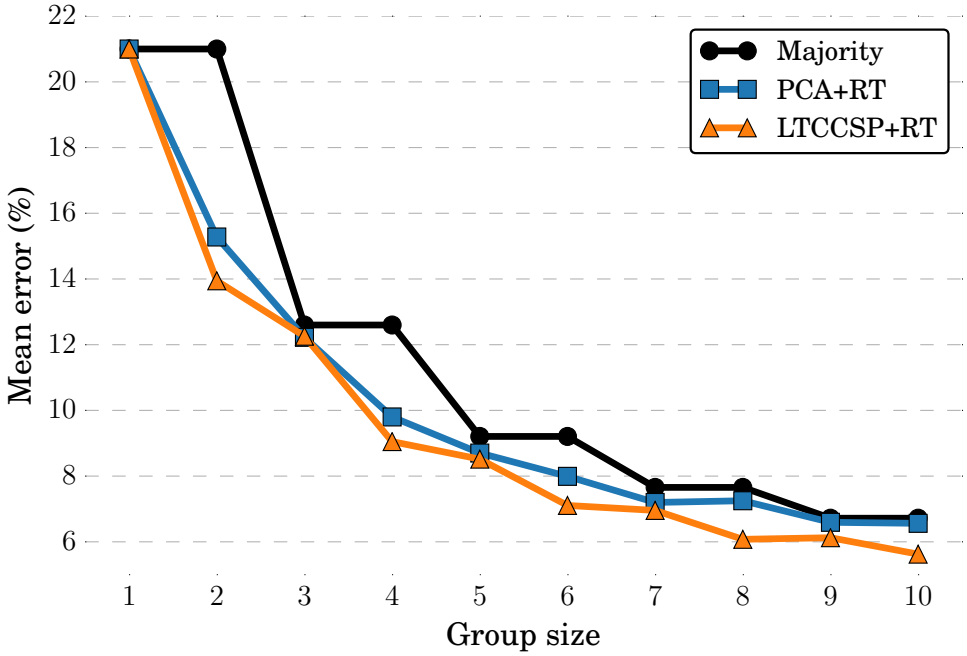

Figure 5.7: Average percentage of errors vs group size for group decisions made by: (1) the majority rule, (2) a cBCI using PCA neural features and RTs, and (3) a cBCI using LTCCSP neural features and RTs.

and the error rates of traditional non-BCI groups using the majority rule. The p-values of the Wilcoxon signed-rank test used to compare the three methods are shown in Table 5.5.

These results show that the cBCI based on LTCCSP neural features is statistically significantly better than the cBCI based on PCA features for all group sizes except for groups of three observers, where the two methods are on par. Moreover, as expected, both cBCIs are significantly better than traditional non-BCI groups using the majority rule for all group sizes, although the PCA-based cBCI performs on par with non-BCI groups for groups of size 9.

Taken together, these results suggest that LTCCSP should be preferred to PCA for extracting neural features.

Table 5.5: p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes adopting the majority rule, the PCA-based cBCI and the LTCCSP-based cBCI. The number of groups of each size that could be assembled with 10 participants is indicated in the last row of the table. p-values below the statistical significance level 0.05 are in bold face.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is PCA+RT better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.2284
Is LTCCSP+RT better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0042**
Is LTCCSP+RT better than PCA+RT?	**0.0000**	0.5562	**0.0000**	**0.0000**	**0.0000**	**0.0001**	**0.0000**	**0.0241**
Sample size	45	120	210	252	210	120	45	10

5.3.6 ERP Analysis

To provide more evidence on why the cBCI achieves superior performance when using also neural features, we analysed the differences in the statistical distributions of ERPs for correct (confident) and incorrect (non-confident) responses in each experiment. Since our cBCI uses both stimulus-locked and response-locked epochs, we show results in both representations. For better visualisation, we down-sampled the epochs data to 64 Hz instead of 16 Hz (as used by the cBCI).

Figure 5.8 shows the stimulus-locked grand averages of a representative subset of the 64 electrode sites used for EEG recording (i.e., Fz, Cz, C3 and C4) for Experiments 1 (left) and 2 (right). As done in Chapter 4, we have used the Kruskal-Wallis test to compare the voltages measured in each channel at each time step for the correct and incorrect trials, and the two-tailed Wilcoxon signed-rank test for paired samples to compare the mean ERPs obtained on an individual basis. The p-values of the statistical tests are also shown in Figure 5.8. Figure 5.9

shows corresponding response-locked grand averages.

These results show that the ERPs of the correct and incorrect classes are significantly different at many time steps in both stimulus- and response-locked representations for both experiments. This suggests that our original decision of discarding the stimulus-locked epochs in the cBCI used with the visual matching task (see Chapter 4) could have led to lose important information about the decision confidence.

The stimulus-locked ERP representation (Figure 5.8) allows the cBCI to see in full resolution [140] and, thus, exploit differences in exogenous and endogenous ERPs associated with the processing and evaluation of the stimulus. In this representation, major differences between correct and incorrect trials occur at approximately 600 ms after stimulus onset, where a slow positive wave has a statistically significantly greater amplitude for the correct than the incorrect decisions. This is likely to be due to the fact that when a trial is particularly hard and, hence, users being unsure of their decision, the amplitude of the P300 is reduced [147], reflecting a more elaborate decision process.

Significant differences between the ERPs elicited in correct and incorrect trials are also present in the response-locked analysis (Figure 5.9). Here the traditional stimulus-locked ERPs associated with early visual processing (such as the P1, N1, P2, and N2) are almost completely absent due to the blurring effect associated with wide RT distributions (see [140] for details) and the preprocessing taking place in the system (in particular the de-trending of the epochs). However, it is apparent that the final phases of the decision-making process (i.e., a few hundred milliseconds before the response) are associated with different amplitudes for correct and incorrect trials, particularly for posterior and occipital channels.

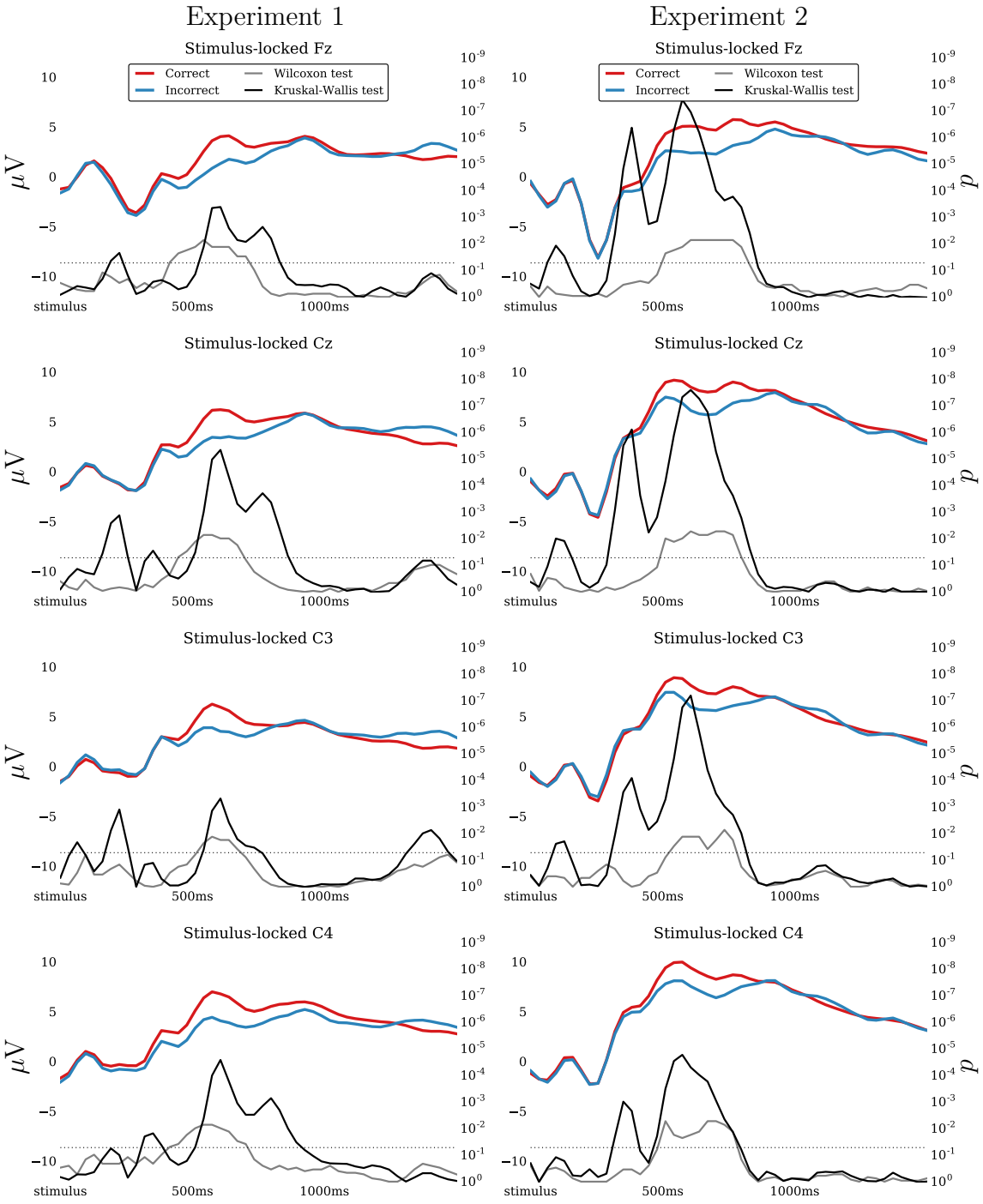

Figure 5.8: Stimulus-locked grand averages for channels Fz, Cz, C3, C4 and corresponding temporal profile of the p-values of the Wilcoxon signed-rank test comparing participant-by-participant averages (grey) and of the Kruskal-Wallis test for all ERPs recorded, irrespective of participant (black), in each error class. The dotted lines represent the 5% confidence level. The corresponding axes are oriented so that values above that line indicate statistical significance.

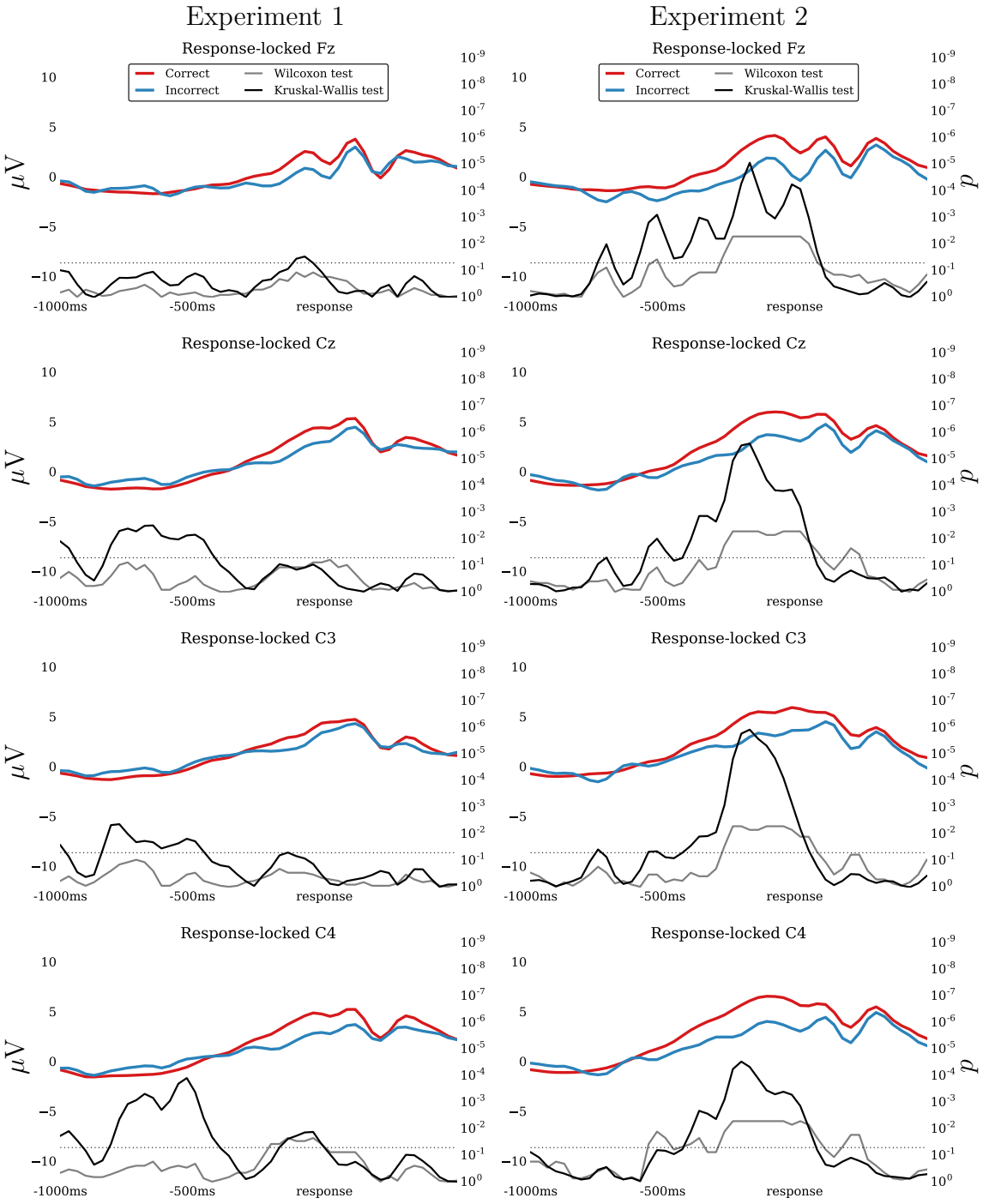

Figure 5.9: Response-locked grand averages for channels Fz, Cz, C3, C4 and corresponding temporal profile of the p-values of the Wilcoxon signed-rank test comparing participant-by-participant averages (grey) and of the Kruskal-Wallis test for all ERPs recorded, irrespective of participant (black), in each error class. The dotted lines represent the 5% confidence level. The corresponding axes are oriented so that values above that line indicate statistical significance.

When comparing the stimulus- and response-locked epochs between Experiments 1 and 2, we can see that the visual search experiment using realistic stimuli generates ERPs which are more significantly different between the correct and incorrect classes than the ERPs of Experiment 1. The grand averages of the correct class show P300 waves which last longer in Experiment 2 than in Experiment 1. Participants might be more engaged with the task due to its reality, which can therefore affect the P300 wave [145]. These results suggest that the choice of using realistic stimuli makes the brain signals more informative for the cBCI, as well as making another step towards real applications of such a system.

To provide an overview of the differences in ERPs between the correct and incorrect trials across the whole scalp, Figure 5.10 shows a snapshot of the scalp potentials recorded 600 ms after the presentation of the stimulus for Experiments 1 and 2, while Figure 5.11 shows another snapshot taken 250 ms before the user's response. We chose these time steps because the differences between the two classes were bigger (e.g., the P300s have their peak between 400 and 700 ms after the stimulus onset [102]). The first three rows of these figures report the scalp maps representing the grand averages for the correct and incorrect trials and their differences, while the last row shows the scalp maps of the p-value of the Kruskal-Wallis test used to compare the voltages recorded at each channel in the two classes.

These scalp maps clearly show how the information provided by stimulus- and response-locked epochs is complementary. Most of the differences of the stimulus-locked representation of the EEG signals are located in the frontal and parietal lobes. Response-locked epochs capture evidence of the decision confidence from the occipital lobe in Experiment 1 and all around the scalp for Experiment 2,

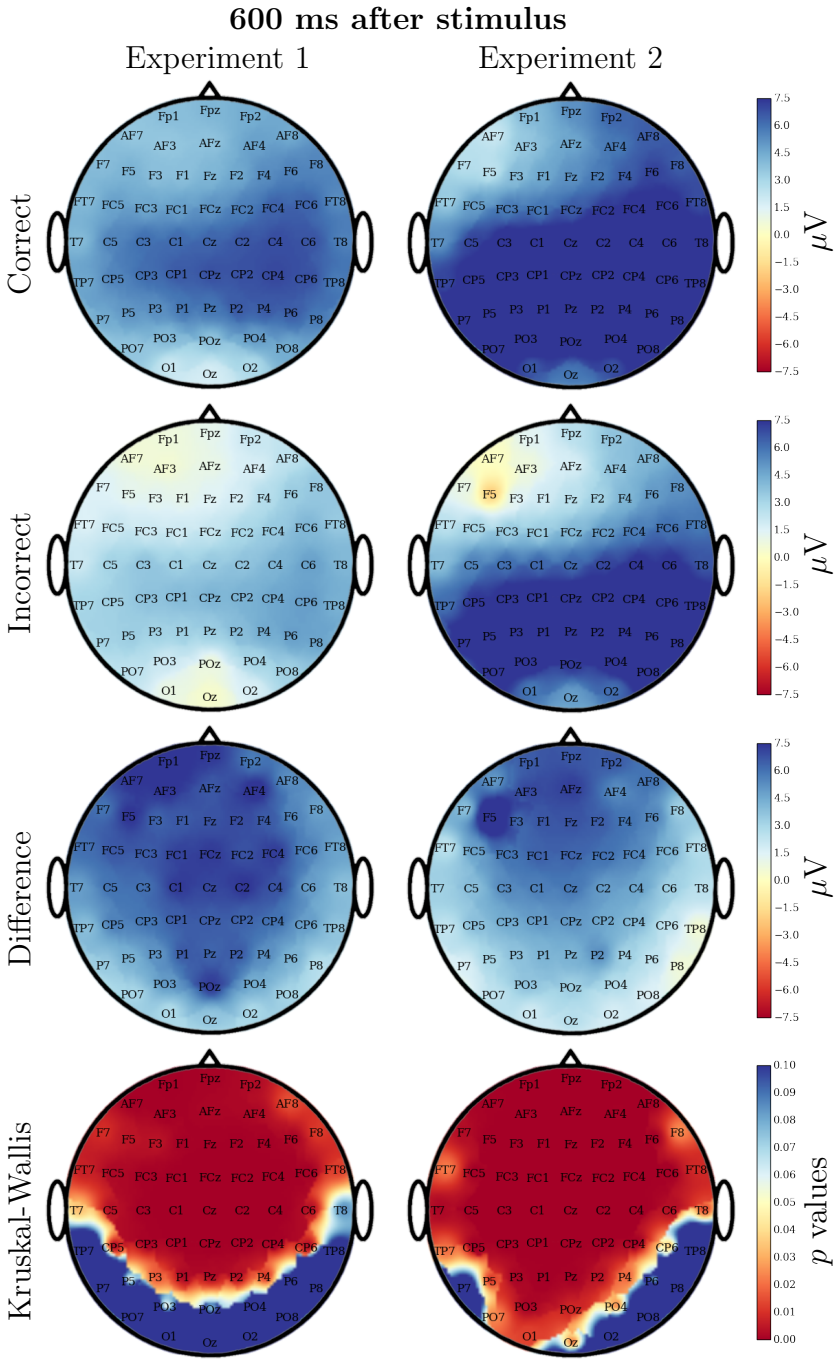

Figure 5.10: Scalp maps of the grand averages of the EEG activity recorded 600 ms after stimulus onset for Experiments 1 (first column) and 2 (second column). Rows represent the activity for correct and incorrect trials (first two rows), the difference between them (third row) and the corresponding p-values of the Kruskal-Wallis test used to compare the two sets (last row).

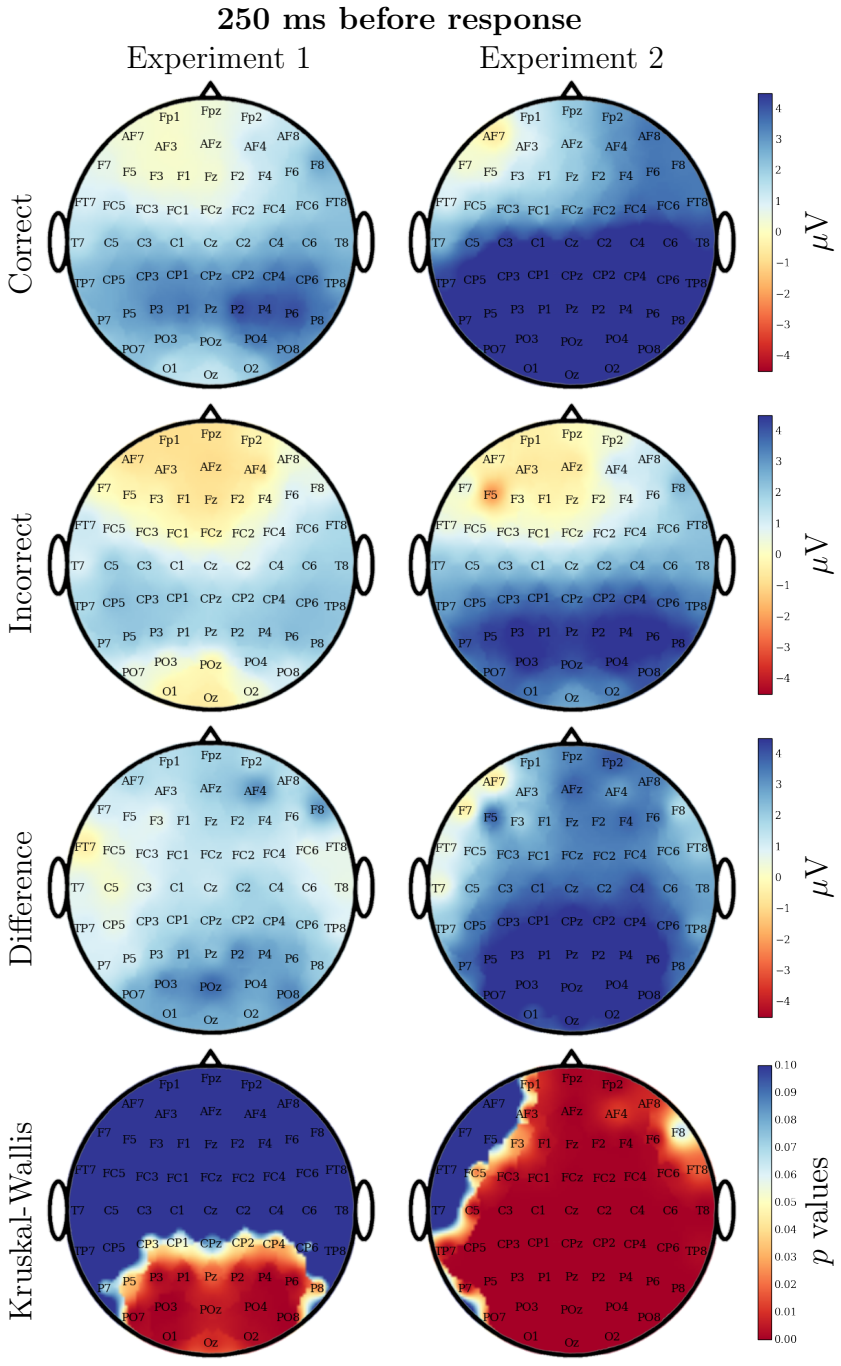

Figure 5.11: Scalp maps of the grand averages of the EEG activity recorded 250 ms before the response of the user for Experiments 1 (first column) and 2 (second column). Rows represent the activity for correct and incorrect trials (first two rows), the difference between them (third row) and the corresponding p-values of the Kruskal-Wallis test used to compare the two sets (last row).

further confirming our previous findings that ERPs are more informative in the visual search task with realistic stimuli than in the one based on bars. These results corroborate our assumption that both representations are useful to estimate decision confidence and should, therefore, be available to the cBCI.

5.4 Conclusions

This chapter has explored the possibility of using the cBCI framework described in Chapter 3 to augment group performance in visual search. We started our investigations with a traditional experiment in which observers had to identify a vertical red bar in a display containing tens of coloured horizontal/vertical bars presented for 250 ms. We then extended our analysis to a more realistic visual search task where participants had to identify a polar bear in an arctic environment containing many penguins.

With both experiments, we have found that cBCI-assisted groups of different size were more accurate than equally-sized non-BCI groups using the simple majority. Most of this group augmentation occurred for even-sized groups, where the cBCI was able to break ties towards correct decisions.

We have also compared the performance of various cBCIs based on different types of confidence correlates, to investigate which combination of features achieves the best group decisions. We have extracted neural features with spatio-temporal common spatial patterns, a technique generally used in motor-imagery BCI [233], from both response- and stimulus-locked epochs. We showed that this approach provided more information to the cBCI to assess the decision confidence. The performance obtained by a cBCI using such features and RTs was compared

with the performance of our previous cBCI based on PCA neural features and RTs. We found that LTCCSP performs significantly better than PCA.

Moreover, this chapter has started investigating the possibility of combining different types of features (i.e., behavioural, neural and physiological) to estimate the decision confidence of a user. We have compared the group performance obtained by a decision-making system based on (1) RTs only, (2) RTs and eye movements features, (3) RTs and LTCCSP neural features, and (4) RTs, LTCCSP and eye movements features. Results indicate that all three types of features provide unique information about the decision confidence and, therefore, the cBCI based on all of them achieves the best performance for most group sizes.

Furthermore, we have verified that our cBCI generalises across tasks. More specifically, we used the cBCI based on 24 PCA neural features and RTs described in Chapter 4 to estimate the decision confidence of the participants undertaking the visual search task with simple stimuli. The results obtained show that both traditional groups based on majority and the cBCI provide the same relative benefits as the group size increases.

When analysing the neural correlates of the decision confidence, we found that the use of realistic stimuli makes correct (confident) decisions easier to be distinguished from incorrect (not confident) ones. This confidence fingerprint could be exploited further with even more realistic tasks (e.g., video-games).

The promising results described in this chapter were obtained with participants performing the experiments in isolation. Group decisions were then *simulated* offline. A drawback of this approach is that it does not consider the impact that collaboration and, in general, being in a group can have on an individual's behaviour and cognitive processing, and, ultimately, on neural activity.

The interaction in a real environment would most likely change the neural signals thereby affecting the performance of a cBCI. In the next chapter, we will investigate the impact of a constrained form of communication on individual and group performance.

Chapter 6

Impact of Group Communication on Visual Search Performance

This chapter studies the impact that a constrained form of communication between pairs of users has on the performance of individuals, traditional groups and cBCI-assisted groups. It also compares the confidence estimated by the cBCI with the confidence estimated by the participants after each decision.

6.1 Introduction

Typically, group decisions are mediated by communication and feedback, whereby members of a group share information and get to know other members' opinions [190]. This often leads to groups having augmented capabilities and intelligence over single individuals. However, communication and feedback do not always provide advantages.

Groups are effective when four conditions apply [181]: (1) individual opinions

are not correlated (diversity), (2) decisions of one individual are not influenced by others (independence), (3) each group member is able to specialise (decentralisation), and (4) it is possible to merge individual opinions into a group decision (aggregation). When some of these conditions are not met, the interactions between group members can have a negative impact on decisions [181, 75]. Moreover, if there are time constraints or if leadership prevails, the process of combining information from freely-communicating individuals can be an obstacle to optimal decision-making [5].

The previous chapters have described a hybrid cBCI which was able to obtain the advantages of groups *without* member interactions. Given that group communication is a double-edged sword [181, 83], one may wonder if allowing communication between the group members assisted by our cBCI would provide further improvement in performance or would be disadvantageous for groups.

The very encouraging results obtained with our hybrid cBCI in visual matching (Chapter 4) and visual search (Chapter 5) were mainly due to the use of the decision confidence estimated from neural, behavioural and physiological signals to weigh individual decisions. In principle, one could more easily and, perhaps, more accurately ask participants themselves to report their decision confidence. This may lead to more accurate group decisions without the need of acquiring the physiological signals required by the cBCI to work, including the noisy and, sometimes, unreliable EEG signals. However, reported confidence is not always accurate. Research has shown that sometimes humans do not report high values of confidence where their decisions are more likely to be correct and *vice versa* [122], which was the assumption our cBCI was based on (see Section 3.5). For example, overconfident people may report high values of confidence when

they are likely to be wrong [96, 132].

This chapter describes the investigation of these two possibilities (i.e., allowing communication and asking users to self-estimate decision confidence) via two experiments.

In the first study, we modified the visual search experiment with realistic stimuli used in Chapter 5 to also ask participants to report their confidence after each decision (Experiment 1). We then compared group decisions obtained using such estimates to weigh individual responses with group decisions made by our cBCI.

In the second investigation, we analysed the impact that a constrained form of communication had on individual and group performance. We designed an experiment where participants were paired while undertaking the visual search task with realistic stimuli described in Chapter 5 and were allowed to exchange information (Experiment 2). The performance of these communicating groups was then compared with the performance of groups of isolated users. Moreover, we also investigated the impact that communication had on the reported confidence estimates.

The chapter is organised as follows. Section 6.2 describes the experiments used in this chapter and the different methods adopted to obtain group decisions. Section 6.3 presents and discusses the results obtained with the participants of our experiments, with a particular focus on comparing the confidence estimates (i.e., reported confidence and BCI confidence) and the group performance with and without user interaction. The chapter ends with Section 6.4 summarising the findings.

6.2 Methodology

6.2.1 Participants

Ten healthy volunteers (average age = 27.4 ± 5.5 years, 5 females) took part in Experiment 1 on different days. Sixteen healthy participants (average age = 28.1 ± 7.2 years, 7 females) were randomly paired and took part in Experiment 2, where they were allowed to exchange information. All volunteers had normal or corrected-to-normal vision.

6.2.2 Experiments

Each experiment consisted of 8 blocks of 40 trials, for a total of 320 trials. Figure 6.1 shows the sequence of stimuli presented in a trial for Experiment 1 (top) and 2 (bottom). In the first four displays, both experiments followed the protocol described in Section 5.2.2, presenting participants the fixation cross, the stimulus, the mask and then asking to indicate their choice with the mouse button. After making a decision (1^{st} response), volunteers were asked to indicate, within 4 seconds, the degree of confidence of their decision ($0 - 100\%$) using the mouse wheel (which varied confidence in 10% steps). Moreover, in Experiment 2 pair members were then shown a display containing the decisions and the degrees of confidence indicated by each of them for 2 seconds. Finally, each pair member was asked again to indicate whether or not the target was present (2^{nd} response). To synchronise Experiment 2, a display containing the text "Please wait" was shown to the fastest member of the pair after indicating his/her confidence, until the other member had also indicated his/her confidence. Response times (RTs)

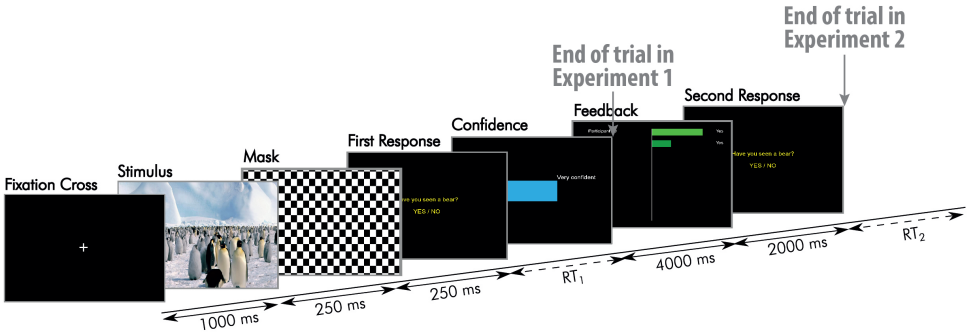

Figure 6.1: Sequence of stimuli presented in the Experiments 1 and 2. The last two displays were only presented in Experiment 2.

were recorded.

The displays used as stimuli for both experiments were obtained from the dataset generated previously (see Section 5.2.2) where (a) six displays where the average error rate across participants in Experiment 1 was below 10% (too easy) or above 90% (too difficult) were discarded, and (b) the number of stimuli was increased by including horizontally-flipped versions of the displays. Thus, the resulting dataset contained 68 stimuli with the target and 10 without it.

The same sequence of displays (randomly generated) was used in both experiments and for all participants. Target occurred in 25% of the trials of each block.

Before an experiment, participants were briefed and familiarised with the task by doing 2 training blocks of 10 trials each. Preparation and practice took roughly 45 minutes. Then, Experiments 1 and 2 lasted about 30 and 40 minutes, respectively. Participants controlled the mouse with the preferred hand and were comfortably seated at about 80 cm from an LCD screen. In Experiment 2, participants were randomly paired and tested in different rooms to avoid direct communication (i.e., the interaction was mediated by the computer as described

above).

6.2.3 Making Group Decisions

Data were acquired and preprocessed as explained in Chapter 3.

We used 10-fold cross-validation to split the dataset into a training set of 288 trials and a test set of 32 trials. Neural features were then extracted as described in Section 5.2.3 by computing the LTCCSP matrices on the data in the training set and using these matrices to transform the data in the test set. Hence, the cBCI used in the two experiments considered in this chapter estimated the decision confidence from 5 features: 2 LTCCSP neural features extracted from each type of epochs (i.e., stimulus-locked and response-locked) and the RT of the 1^{st} response.

Once the features were extracted, we fit the LARS [37] model to predict the decision confidence (as done in Chapter 5) using the data in the training set. Then, the data of each participant in each trial of the test set were transformed into confidence weights w by using the negative exponential weighting function described in Equation (3.3).

To address one of the aims of this study, we have also used the raw confidence reported by the user in a trial as weight w to compute the group decision. In this case, the weights were discrete, i.e., $w = \{0, 10, 20, \ldots, 100\}$.

Group decisions were then made as described in Section 3.8 by using the sign of the weighted sum of members' decisions, where the weights were either the confidence reported by the participants or the confidence weights computed by the cBCI.

Due to the limited number of identical EEG acquisition devices available in our lab, in Experiment 2 we could only test the effects of concurrent communication on pairs. However, to gain some insight on the performance achievable by larger groups of interacting observers, we combined (offline) the 8 pairs in all possible ways to form groups of size 4, 6, etc. We chose this way of proceeding instead of the method described in Chapter 3 and used in Experiments 1 (i.e., combining individual participants in all possible groups of increasing size) to avoid splitting communicating pairs, thereby retaining some of the dynamics observed in such groups. Hence, we had 28 groups of size 4, 56 groups of size 6, and so on.

6.3 Results

This section presents the results obtained with the two experiments.

6.3.1 Communication Worsens Individual Performance

We start our analysis by looking at individual performances in the two experiments. It should be noted that, when considering the 1^{st} responses in Experiment 2 (i.e., those given by the observers before any exchange of information related to the task at hand), in principle the performance of the participants should be similar in the two experiments as they are exposed to the same information.

Figure 6.2 shows the individual error rates in the two experiments. While the participants of Experiment 1 made, on average, 22.6% erroneous decisions, those of Experiment 2, surprisingly, were 50% worse in the same task (i.e., when considering the 1^{st} response), with an average error rate of 33.1%. A Kruskal-Wallis

test confirmed that the error distributions of individual decisions in Experiment 2 before the communication occurred were significantly different from the distributions of the isolated observers in Experiment 1 ($p = 0.0017$).

Even more surprising was the individual performance obtained when using, in Experiment 2, the 2^{nd} response provided by participants after our constrained form of communication (dark grey bars in Figure 6.2(right)). We expected these decisions to be more accurate than the 1^{st} ones as they integrated the information shared within the pair [60]. However, the average error rate across participants was not statistically significantly different from that obtained using the 1^{st} responses (two-sided Wilcoxon signed-rank $p = 0.875$). This suggests that the exchange of information between participants had no effect on their individual performance.

It is known that in certain tasks, such as estimating the number of sweets in a jar [84] or answering factual questions with a numerical answer [99], interactions between participants can negatively affect individual performance. However, we found it surprising that such an effect could occur in the perceptual decision task used in our experiments (*cf.* individual performance in Experiments 1 and 2), especially when considering the first responses provided by participants in Experiment 2 where no interaction happened between the pair's members. This suggests that the *context* in which participants were immersed (i.e., isolated or paired) was sufficient to cause a change in participants' performance.

In Figure 6.2(right) we can see an additional effect of the interaction: in most of the pairs, the performance of the two participants are very similar to each other, especially when considering the 2^{nd} responses. This suggests that interaction seems to have an effect on individual performance, although leading

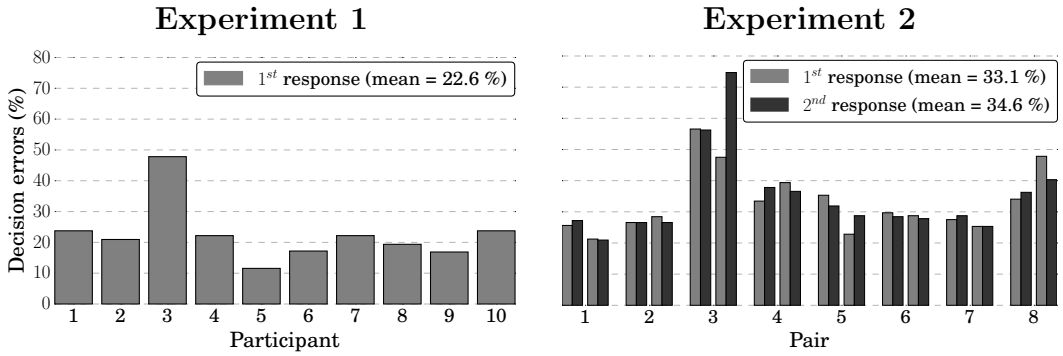

Figure 6.2: Error rates of participants for the two experiments. In Experiment 2 (right) the individual decision errors are based on either the 1^{st} (light grey) or the 2^{nd} (dark grey) responses, given by observers *before* or *after* seeing the decision and the confidence reported by the other group member, respectively.

to higher error rates instead of lower ones.

6.3.2 cBCI Groups Achieve the Best Performance

Figure 6.3 shows, for each experiment, the mean error rate of groups of increasing size making their decisions using the 1^{st} responses provided by participants and adopting (a) the majority rule, (b) a weighted majority where weights are given by the confidence reported by each participant, and (c) a weighted majority using the confidence weights estimated by our cBCI. For Experiment 2 we also report the performance of the majority rule when using the individual decisions provided by participants after exchanging information (2^{nd} responses) – see green line in Figure 6.3(right).

Let us first analyse the results of Experiment 1, where participants undertook the visual search task without any interaction with each other. As done previously, we have used the Wilcoxon signed-rank test to compare the performance of groups of various sizes making decisions using the three methods analysed in

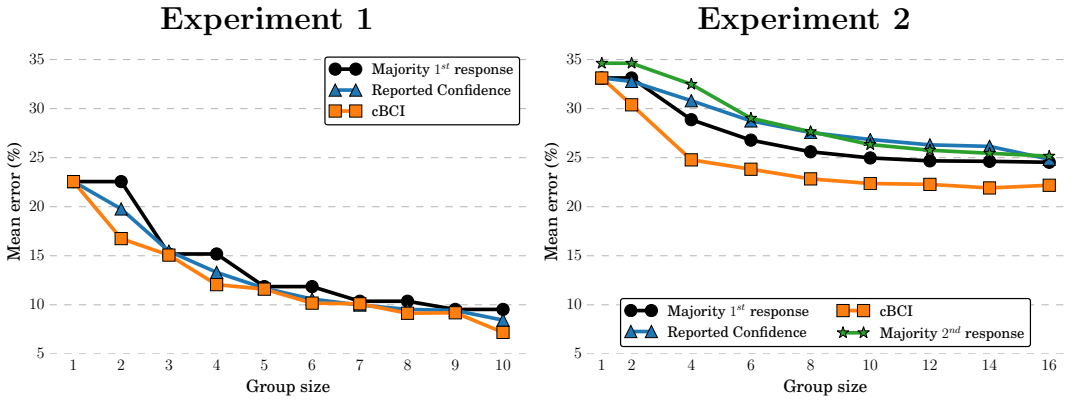

Figure 6.3: Error rates of groups of increasing size in the two visual search experiments conducted in the study. Group decisions are built using: (a) a majority rule based on individual responses (black line), (b) a weighted majority based on the reported confidence (blue), (c) a weighted majority based on the confidence estimated by the cBCI (orange), and (d) only for Experiment 2, a majority rule based on individual responses after the feedback (green).

this chapter. The p-values of these comparisons are shown in Table 6.1.

For group sizes 2–8 the performance of cBCI-assisted groups was significantly better than that of traditional groups using the majority rule, confirming our previous findings described in Chapter 5. The two methods perform on par for groups of size 9. Groups making decisions using the confidence reported by the observers are also superior to majority-based groups for group sizes 2, 4, 5, 6, 7, 8, while the two methods perform on par for group sizes 3 and 9. This suggests that the reported confidence is a good alternative to the cBCI to improve on standard majority when participants are not communicating. However, when comparing the two confidence-based methods, we found that cBCI-assisted groups made significantly better decisions than groups based on the confidence reported by the participants for all *even* group sizes, while the two methods performed on par for the odd group sizes. This suggests that the confidence reported by participants

Table 6.1: Statistical comparison of methods for group decisions for different group sizes in Experiment 1. The table reports the p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes adopting (a) the majority rule, (b) a weighted majority using the confidence reported by participants as weights (ConfidenceMajority), and (c) a weighted majority using the confidence weights estimated by the cBCI. The number of groups of each size that could be assembled with 10 participants is indicated in the last row of the table. p-values below the statistical significance level 0.05 are in bold face.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is ConfidenceMajority better than Majority?	**0.0000**	0.2005	**0.0000**	**0.0019**	**0.0000**	**0.0005**	**0.0000**	0.4719
Is cBCI better than Majority?	**0.0000**	**0.0004**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.1173
Is cBCI better than ConfidenceMajority?	**0.0060**	0.6469	**0.0000**	0.8643	**0.0003**	0.8913	**0.0102**	0.2491
Sample size	45	120	210	252	210	120	45	10

was never superior to the one estimated by the cBCI while the cBCI was able to significantly enhance even-sized group performance.

We now analyse the results obtained in Experiment 2 (Figure 6.3(right)), where a constrained form of communication was allowed between pairs of users. Table 6.2 reports the p-values of the Wilcoxon signed-rank test comparing the performance of groups of different sizes making decisions using either the three methods compared for Experiment 1 employing the 1^{st} responses or a majority rule based on the 2^{nd} responses provided by the participants after exchanging information with the other group's member in Experiment 2.

As seen in the previous section, the individual performance in Experiment 2 was much worse than the performance obtained by isolated participants. Therefore, it is not surprising seeing that, overall, groups of various sizes are generally

Table 6.2: Statistical comparison of methods for group decisions for different group sizes in Experiment 2. The table reports the p-values of the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes adopting (a) the majority rule based on the 1^{st} responses, (b) a weighted majority using the confidence reported by users as weights (ConfidenceMajority), (c) a weighted majority using the confidence weights estimated by the cBCI, and (d) the majority rule based on the 2^{nd} responses. The number of groups of each size that could be assembled with 8 pairs is indicated in the last row of the table. p-values below the statistical significance level 0.05 are in bold face.

	Group size						
Comparison	2	4	6	8	10	12	14
Is Majority better than Majority2?	0.4167	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0002**	**0.0294**
Is Majority better than ConfidenceMajority?	0.6880	**0.0018**	**0.0000**	**0.0000**	**0.0000**	**0.0001**	**0.0124**
Is cBCI better than Majority?	**0.0111**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0071**
Is cBCI better than ConfidenceMajority?	**0.0210**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0071**
Is cBCI better than Majority2?	**0.0104**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0071**
Sample size	8	28	56	70	56	28	8

less accurate than equally-sized groups of Experiment 1, since their decisions are obtained by combining individual responses.

The performance of pairs making decisions using non-BCI methods were on par in Experiment 2, although the method based on the 2^{nd} responses slightly increased the error rates of the pairs. However, cBCI-assisted pair decisions were significantly superior than those made using all non-BCI methods.

When simulating larger groups by aggregating pairs, we found that the cBCI was always superior to the three other methods. Moreover, decisions made by larger groups using the confidence values reported by the participants or their

2^{nd} responses were significantly worse than those made by majority-based groups using the individual responses provided before any interaction.

On the basis of these results we can make three main conclusions. Firstly, *the cBCI provides the best group performance* over the other methods analysed in this chapter regardless of the presence or absence of communication within the pairs. Secondly, the confidence reported by the participants is a valid alternative to the neuro-behavioural confidence estimates provided by the cBCI *only with isolated users*. When observers are communicating, the performance of groups where decisions are based on the reported confidence is never superior and generally significantly worse than the performance of majority-based groups. Thirdly, giving participants the opportunity to change their decision after exchanging information (i.e., 2^{nd} responses) *significantly reduces group performance* for groups of size 4–14 and does not provide any advantage over the 1^{st} responses for pairs. Sections 6.3.3 and 6.3.8 provide more evidence to support the last two considerations.

6.3.3 Paired Context Worsens Metacognitive Accuracy

To investigate further the reasons behind the poor performance obtained by groups of communicating observers when using the reported confidence, we compared the confidence values indicated by participants in correct decisions (D_c set) with those indicated in incorrect decisions (D_i set). As described in Chapter 3, to obtain good group performance with our decision-making system the confidence should correlate with the correctness of the decision (i.e., metacognitive accuracy).

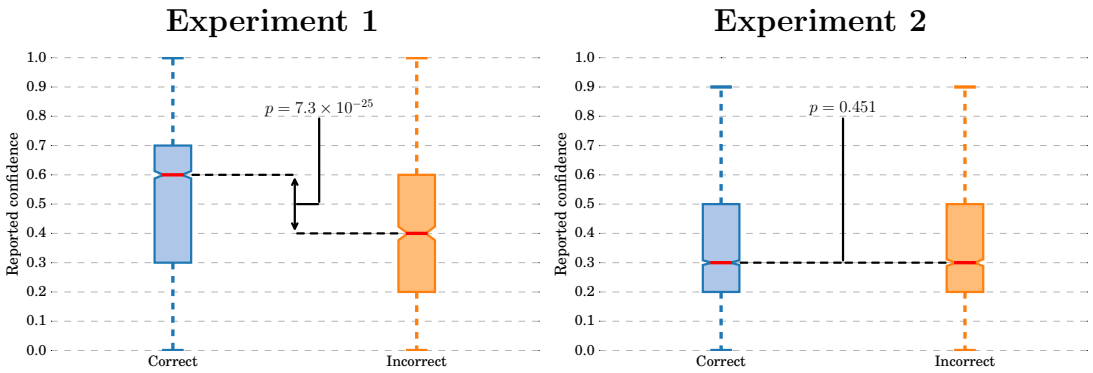

Figure 6.4: Confidence values indicated by participants for correct and incorrect decisions in Experiment 1 (left) and 2 (right) and corresponding p-value of the Kruskal-Wallis test used to compare the two distributions.

Figure 6.4 shows the distribution of the reported confidence values in the D_c and D_i sets. The p-value of the Kruskal-Wallis test used to compare the two distributions is also shown. When participants were not allowed to communicate (i.e., in Experiment 1), these confidence values were good predictors of the correctness of the decision as the two distributions D_c and D_i were significantly different – see Figure 6.4(left). However, observers of Experiment 2 (who were allowed to exchange information) reported confidence values which were totally unrelated with the correctness in the decisions.

These results show that reported confidence is significantly affected by the context in which participants are immerse (i.e., isolated or paired) and, therefore, it is an unreliable predictor of correctness. While reported confidence allows to improve group performance over majority when participants are deciding in isolation, it does not provide any advantage when participants communicate indirectly through their being made aware of each others' 1^{st} responses and reported confidence levels. This leads to group decisions that are often even more erroneous than those made with simple majority.

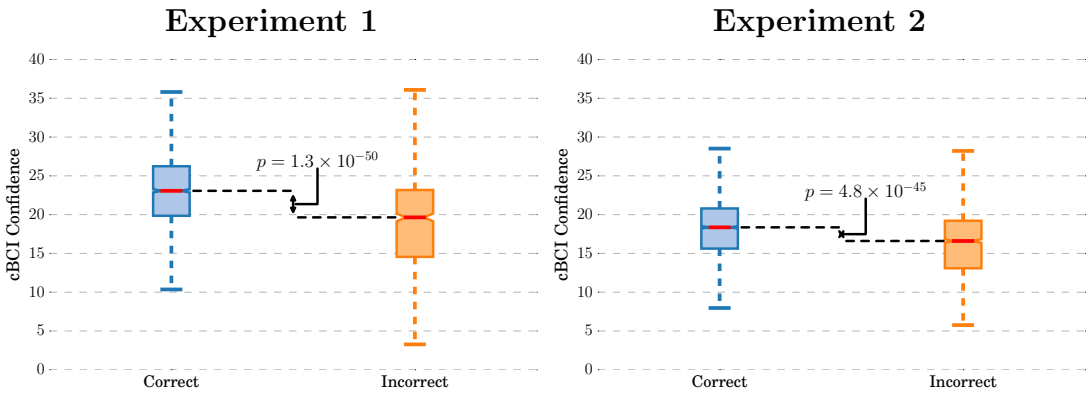

Figure 6.5: Confidence weights estimated by the cBCI for correct and incorrect decisions in Experiment 1 (left) and 2 (right) and corresponding p-value of the Kruskal-Wallis test used to compare the two distributions.

6.3.4 BCI Confidence is not Affected by Context

Similarly to the analysis conducted in the previous section, we have also compared the distributions of the confidence weights estimated by our cBCI for the correct (D_c) and incorrect (D_i) sets of trials. Figure 6.5 shows the results of this comparison.

In both experiments, the cBCI is able to provide confidence weights that are significantly different for the D_c and D_i sets. This makes the cBCI a robust predictor of the correctness of the decision regardless of the context, which explains the superior performance achieved by cBCI groups in both experiments – see Section 6.3.2.

6.3.5 Response Times Correlate with Correctness

In this and the following section, we will examine more in detail the sources of the robustness of the cBCI in estimating a decision confidence that correlates with the accuracy. As described in Section 6.2, our cBCI uses a combination of

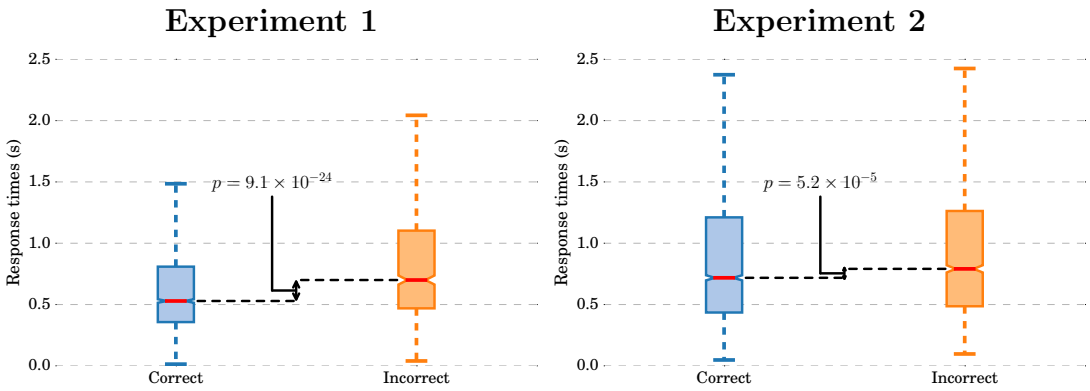

Figure 6.6: Response times for correct and incorrect decisions in Experiment 1 (left) and 2 (right) and corresponding p-value of the Kruskal-Wallis test used to compare the two distributions.

RTs and neural signals to estimate the decision confidence. In this section we examine whether or not the information brought by the RTs is affected by the communication, while in the next section we will perform a similar analysis on the neural signals.

Similarly to what we have done with confidence estimates, we used the Kruskal-Wallis test to compare the distributions of RTs between the correct (D_c) and incorrect (D_i) sets of trials. For Experiment 2, we have considered the RTs of the 1^{st} responses. As shown in Figure 6.6, we found that the RTs distributions were significantly different between D_c and D_i for both experiments. However, it should be noted that in Experiment 2 the two distributions become more similar, suggesting that also RTs are influenced by the context. Nevertheless, they still carry information related to the probability of the decision being correct and, therefore, it is reasonable to use them as a feature for the cBCI to obtain confidence estimates.

6.3.6 Context Changes Neural Correlates of Confidence

We also investigated the impact of the context on the decision-making processes in the brain, as these are the sources of the neural features that our cBCI uses to build the confidence estimations.

We divided the stimulus- and response-locked epochs in the D_c and D_i sets, the former containing the ERPs recorded in trials where the user made a correct decision and the latter with ERPs associated to incorrect responses. We then used the Kruskal-Wallis test to compare the voltages measured at each time step at each electrode site for the two sets. Moreover, we have used the Wilcoxon signed-rank test to compare participant-by-participant averages.

Results from representative electrode sites Fz, Pz, C3 and C4 of Experiments 1 and 2 are shown in Figure 6.7 for stimulus-locked epochs and in Figure 6.8 for response-locked ones.

These results confirm that the neural signals still differ significantly between correct and incorrect trials. However, in both ERP representations the information about decision confidence was less evident in Experiment 2, where participants were paired while performing the visual search task. In this experiment, the grand averages of the ERPs for the two classes look very similar to each other, but still present some statistical differences.

Nevertheless, the relative separation of the confidence values provided by the cBCI for correct (confident) and incorrect (not confident) trials shown in Figure 6.5 indicated that our system is able to provide robust correctness predictors even with the fainter evidence available in Experiment 2, leading to significantly reducing the percentage of erroneous group decisions.

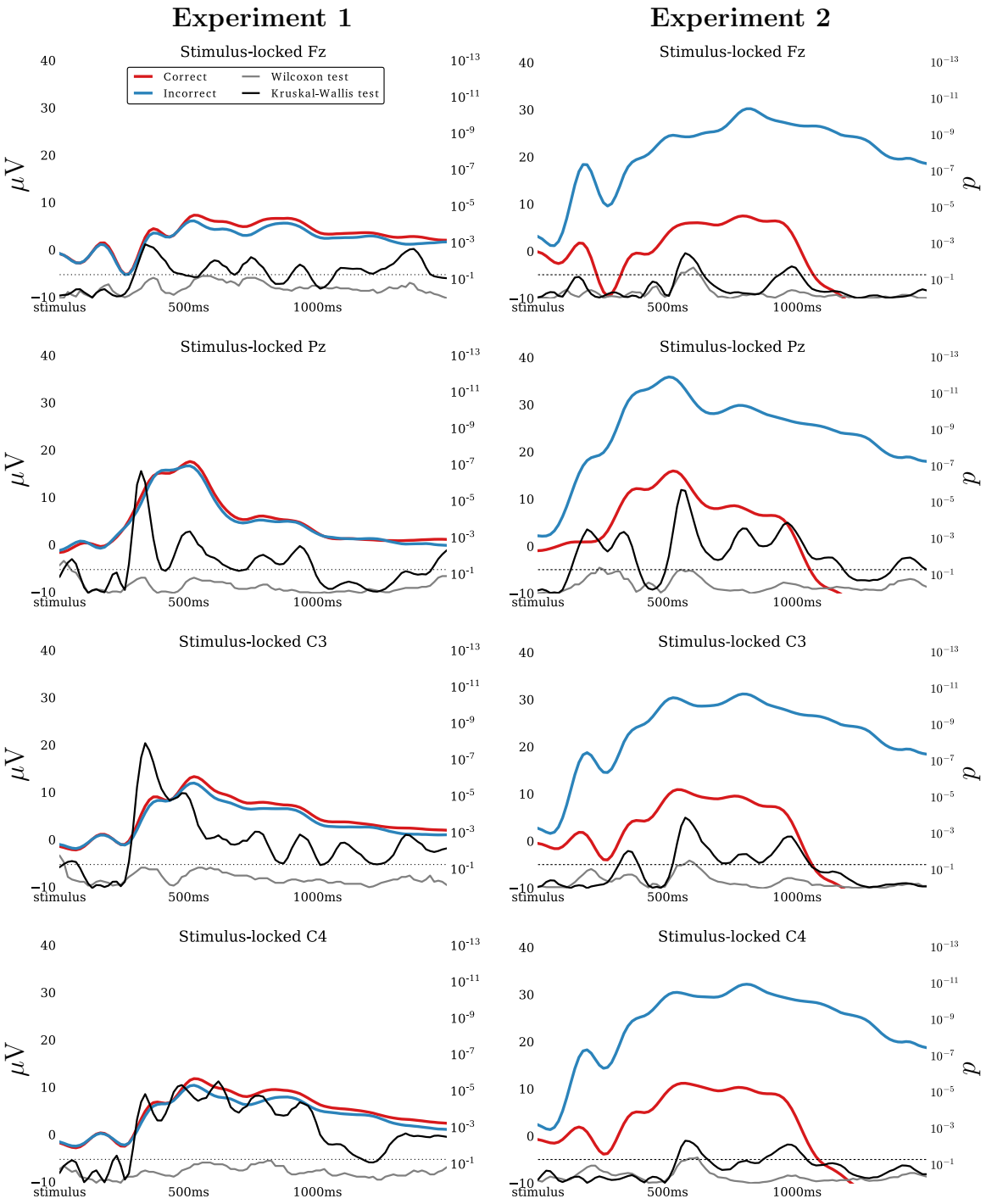

Figure 6.7: Averages of stimulus-locked epochs computed across participants on the correct (red) and incorrect (blue) ERP sets and corresponding temporal profile of the p-values of the Wilcoxon signed-rank test comparing participant-by-participant averages (grey) and of the Kruskal-Wallis test comparing all ERPs recorded in each error class (black) for representative channels Fz, Pz, C3 and C4 for Experiments 1 and 2. p-values above the horizontal dotted line (representing the 5% confidence level) indicate statistical significance.

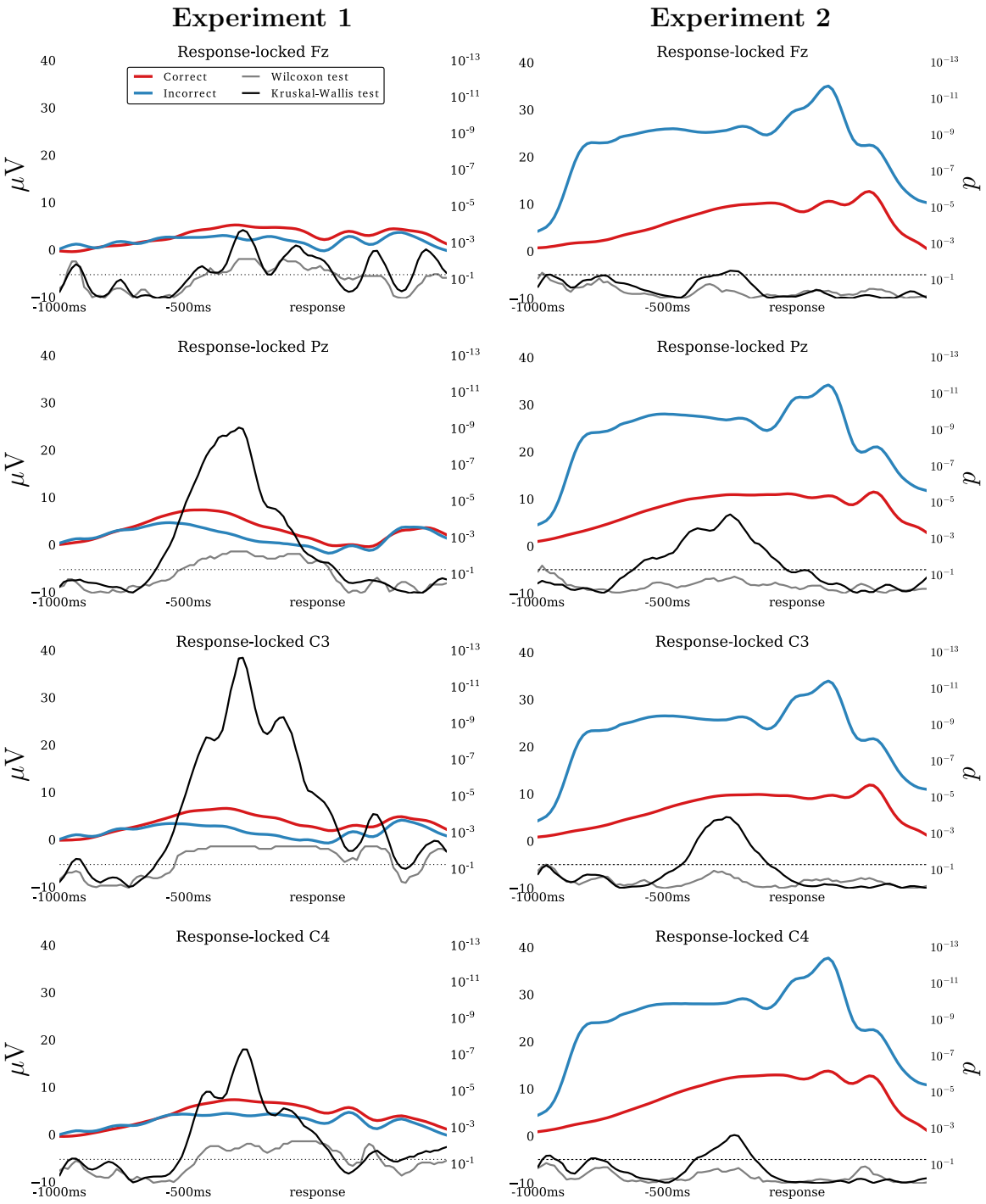

Figure 6.8: Averages of response-locked epochs computed across participants on the correct (red) and incorrect (blue) ERP sets and corresponding temporal profile of the p-values of the Wilcoxon signed-rank test comparing participant-by-participant averages (grey) and of the Kruskal-Wallis test comparing all ERPs recorded in each error class (black) for representative channels Fz, Pz, C3 and C4 for Experiments 1 and 2. p-values above the horizontal dotted line (representing the 5% confidence level) indicate statistical significance.

6.3.7 Interaction Nullifies the Advantages of Experience

In Section 6.3.1 we have seen that the error rates of participants of Experiment 2 were much higher than those of observers in Experiment 1. In the following sections we investigate the reasons behind this reduction in performance.

Firstly, we analysed how the error rates vary during the experiments. Experience and task familiarisation should improve performance [98] and, therefore, we should expect higher error rates in the earlier part of an experiment than later on. Figure 6.9 shows the mean error rates across participants for the two experiments computed using a simple moving average over 40 consecutive trials when using the 1^{st} (both experiments) or the 2^{nd} (only Experiment 2) responses. To visualise better the trend of the error rates along the experiment, we have fitted a linear regressor to each dataset.

Let us consider the data gathered from the 1^{st} response (red lines in Figure 6.9), which are available for both experiments. When no communication is allowed between participants (i.e., in Experiment 1), the individual performance does increase along the experiment – see Figure 6.9(left). However, when users are allowed to communicate (Experiment 2), surprisingly, we observed the opposite trend, with participants getting worse over time – see Figure 6.9(right). As we have verified with the Kruskal-Wallis test, these error distributions of the two experiments were significantly different ($p = 4.95 \times 10^{-99}$).

Differences between the two error distributions start as early as the first session of the experiments. When participants were isolated, the error distributions in trials 1–10 and 31–40 were similar as users were still familiarising with the task (two-sided Wilcoxon signed-rank $p = 0.75$). On the other hand, the performance

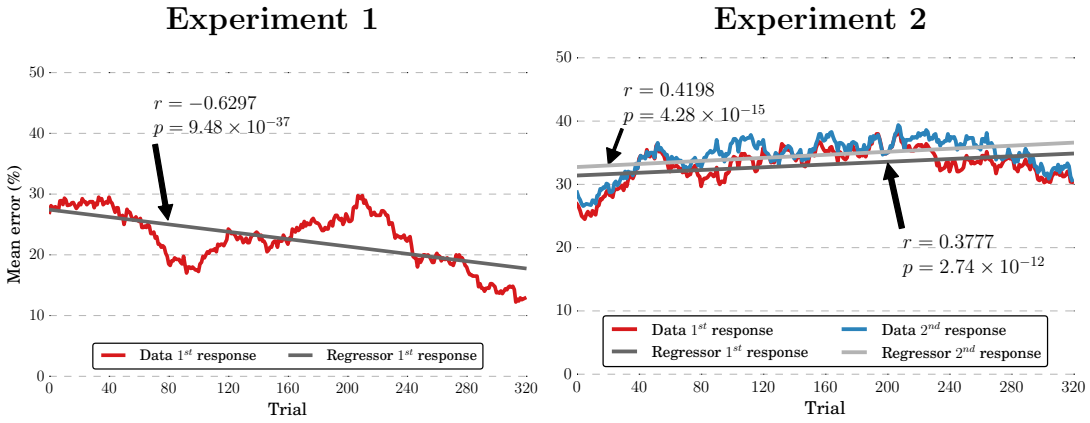

Figure 6.9: Mean error rates across participants for Experiments 1 (left) and 2 (right) computed using a simple moving average on the 1^{st} (red) and 2^{nd} (blue) responses. The grey lines show the linear regressors fitted on the each set of data. The correlation coefficients and the two-sided p-values of the regressors are also indicated.

of communicating participants very rapidly and significantly deteriorates in those trials ($p = 0.04$).

Interestingly, the average performance of participants of Experiments 1 and 2 were almost identical in the first 10 trials and error distributions were not significantly different (Kruskal-Wallis $p = 0.77$). This suggests that the participants had the same initial attitudes and abilities to perform the visual search task. Therefore, their subsequent significantly-different performance was mainly due to the presence or absence of communication (i.e., context).

The average error rates increased even more if we considered the responses provided by the participants of the Experiment 2 *after* seeing the other group member decision and confidence (blue and light-grey lines in Figure 6.9(right)). The two error rate distributions of Experiment 2 (shown in red and blue in Figure 6.9(right)) were significantly different (two-sided Wilcoxon signed-rank $p = 3.87 \times 10^{-49}$).

These results suggest that not only communicating participants are not improving their performance over time, but also that the group interaction (at any stage) negatively affects individual error rates.

6.3.8 Communication Does Not Increase Agreement

One of the main advantages of groups is their intrinsic error correction capabilities, which could be exploited when the decisions made by their members are *diverse* and observations are *not* correlated [181, 77, 36]. In case of pairs, this occurs when the participants give opposite responses, hence generating a tie. The voting method adopted to aggregate the different decisions should then have a tie-breaker strategy (e.g., based on the expertise of the observer) to arrive at a group decision in all cases. Hence, the group performance is not only related to individual accuracy but also to the breaking of ties.

We analysed how the level of agreement of the pairs varied along Experiments 1 and 2 by plotting the mean percentage (across participants) of decisions in which the pair members were disagreeing on a decision (i.e., tie) using either the 1^{st} responses and, for Experiment 2, the 2^{nd} response. The values are averaged across the 45 possible pairs formed with participants of Experiment 1 and the eight pairs of users of Experiment 2. A simple moving average algorithm over 40 consecutive trials has been used to smooth the data. We expected that communicating participants would be more likely to agree on a decision than isolated ones. The results are shown in Figure 6.10.

In Experiment 1 (Figure 6.10(left)), the percentage of trials in which the pair disagrees decreases as the experiment progresses. This is reasonable because,

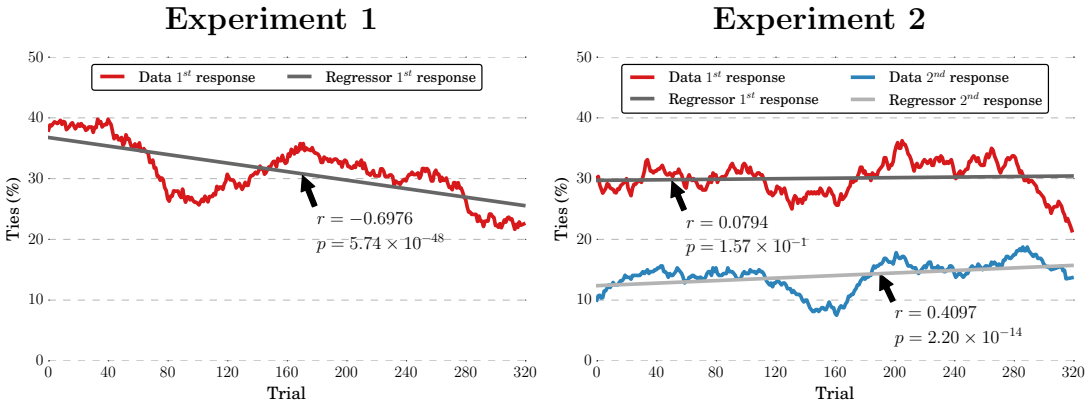

Figure 6.10: Percentage of ties in Experiment 1 (left) and 2 (right). The grey lines show the linear regressor fitted on the data. The correlation coefficients and the two-sided p-values of the regressors are also indicated.

as we have seen before (see Figure 6.9(left)), participants performance improved due to experience. Therefore, some of the ties were actually turned into correct decisions. However, surprisingly, when observers communicated (Experiment 2), their level of agreement remained almost constant – see Figure 6.10(right). We verified with the Kruskal-Wallis test that the two error distributions (red lines in Figure 6.10) were significantly different ($p = 2.32 \times 10^{-80}$). One of the main causes of this significantly different behaviour is that individual performance did not improve over time in Experiment 2 (see previous section).

Experiment 2 gave a chance to participants to change their decisions after sharing information about the other member's decision and reported confidence. In theory, these new decisions would be the result of the increased sensing capabilities and cognition of groups obtained by merging members' knowledge and intelligence. Hence, we expected the level of agreement to be higher and to achieve better performance than that obtained using the decisions made by the participants before sharing any information.

The results shown by the blue line in Figure 6.10(right) confirmed that the level of agreement was much higher (as indicated by fewer ties) when using this 2^{nd} response. However, surprisingly, as shown in Figures 6.9 and 6.3(right), performance was worse. The percentages of erroneous decisions achieved by individual participants and even-sized groups using the majority rule and the 2^{nd} responses (green line) were higher (+2%) than those obtained when using individual decisions provided *before* the constrained form of communication (black line). This suggests that interaction makes participants agree on erroneous decisions.

6.4 Conclusions

Communication in groups is a double edged sword. It is a vital means to reach a consensus, but, for instance, in the presence of strong leadership it can lead to poor group decisions [83]. In this chapter we have investigated the impact of a constrained form of communication on individual and group performance. To do so, we used two realistic visual-search experiments, one where participants were *not* allowed any interaction and one where a constrained form of communication was taking place within pairs of users after each decision, giving the observers the possibility of changing their responses.

Group decisions were obtained by integrating individual responses using either the majority rule or a confidence-based weighted majority, where the weights were estimated by our cBCI introduced in Chapter 3 using EEG signals and RTs.

We have shown that groups make significantly better decisions when assisted by our cBCI than when using the standard majority rule, regardless of the presence or absence of communication.

When a controlled form of communication within pairs was allowed, however, users made many more erroneous decisions than in the experiment where they could not interact. Moreover, communication had a negative impact on the level of agreement (i.e., the number of ties did not decrease over time, hence requiring a better-than-random tie-breaker, like the cBCI, even more) and neural signals (i.e., the patterns that identify confidence became similar to those identifying uncertainty). Furthermore, decisions made by interacting pairs were significantly worse than those made by the average isolated participant. These results suggest that social influence deteriorates individual and group performance in our visual search task. Communicating people trust their gut feelings less than isolated ones [181] and become less prone to risk than required by the task [38], resulting in increased error rates.

The changes in the neural signals caused by interaction made the discrimination between correct and incorrect trials performed by our cBCI more challenging due to the reduction in available information. However, even in these conditions, thanks to its machine learning component and the presence of the RTs in the feature vector (which, we showed were still affected by communication), the cBCI was able to provide a consistent (i.e., results verified with 10-fold cross-validation) and statistically significant improvement in the performance of even-sized groups when compared to traditional groups.

This chapter has also investigated whether it would be possible to replace the confidence estimated by the cBCI with a confidence reported by the participant after making a decision. While this approach works when participants are in isolation, we showed that communication makes reported confidence totally unrelated to the correctness of the decision. These results suggest that the reported

confidence is an unreliable predictor of correctness, while the estimates produced by our cBCI are more robust and consistent.

The results obtained in this chapter suggest that superior group decisions in visual search are achieved when group members are isolated and their decisions are integrated by using our cBCI based on neural signals and RTs. The confidence estimated by the participants could be a good alternative tie-breaker, but should be used cautiously due to its unpredictable reliability.

Chapter 7

A State-Space Model for Cognitive State Estimation

Apart from being used for estimating the decision confidence, physiological and behavioural measures could give an insight into the cognitive processes of a person, which, in turn, are likely to affect decision making. This chapter describes the development of a state-space model based on neural and behavioural signals to estimate the cognitive state of observers undertaking a decision-making task.

7.1 Introduction

Humans and animals have the ability to learn and change their behaviour as a result of the experience gained while undertaking a certain task. Learning is a dynamic process that generally leads to better performance. For example, in a decision-making task, participants usually improve their performance over time thanks to their experience [98]. This learning process does not only have an

impact on the behaviour, but also on neural and other physiological signals [174].

Research has shown that the cognitive load is likely to affect the learning process [183]. EEG signals could be used to monitor the cognitive load of a user [113, 39] and, so, indirectly, to monitor how a decision-maker improves his/her performance due to experience. Moreover, EEG has been used to detect variations in other measures related to decision making, including mental fatigue and attention level [130, 20, 187]. This information could be used as additional inputs to our cBCI to further improve the accuracy of the confidence estimates.

This chapter starts exploring the possibility of using a state-space model to estimate the cognitive state of the decision makers from their neural and behavioural signals. This model could then be plugged into the cBCI (see Figure 7.1) to detect and, possibly, predict changes in the attention level of the user that could affect individual and group performance. Equipped with such a feature, the cBCI could then decide to temporally exclude the tired (and therefore more likely to err) users from the group, leading to further improvement in group performance.

The chapter is organised as follows. Section 7.2 introduces state-space models and defines some notation that will be used across the chapter. Section 7.3 describes a state-space model derived from behavioural measures including the correctness in the decision and the RTs. This model will then be extended in Section 7.4 to also include neural features. Section 7.5 compares different state-space models based on various behavioural and neural features. Finally, Section 7.6 will discuss the potential implications of this work and draw some conclusions.

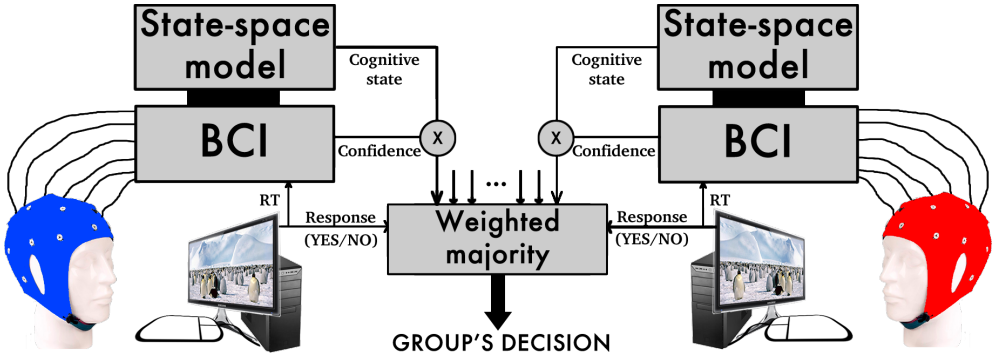

Figure 7.1: Architecture of the decision-making system including the BCI to estimate the decision confidence (as described in Chapter 3) and the additional state-space model to estimate the cognitive state of the user analysed in this chapter. The BCI and the cognitive state modules share the same feature vector (i.e., EEG and RTs).

7.2 State-Space Models

This section briefly introduces state-space models and the methods used in this thesis to estimate their parameters.

7.2.1 Definition and Representation

A Hidden Markov Model (HMM) [154] is a probabilistic model that represents a system as a Markov process with *unobserved, discrete* states over sequences of observations. Let x_t be the hidden state at time t. An HMM is described by the number of possible values that the state could assume K, the number of possible values that the observations could assume M, the state transition matrix A, the observation matrix B and the initial conditions π. In an HMM, the hidden state x_t satisfies the Markov property: the current state x_t is independent of all the states prior to $t-i$, where i is the order of the model [51]. In this thesis, we will consider *first-order* HMM, so that the hidden state at time t only depends on the

state at time $t-1$.

A State-Space Model (SSM) is an HMM where the hidden state modelled is *continuous*, that is, although the progression between one state and another is discrete, the state variable could take any real value ($K \in \mathbb{R}$). A first-order SSM can be written with two equations, a *state equation* and an *observation equation*:

$$\begin{cases} x_t = f(x_{t-1}, \nu_t; w) \\ y_t = h(x_t, n_t; w) \end{cases}, \tag{7.1}$$

where ν_t and n_t are noise processes affecting the state and the observation evolutions, respectively, while f and h are nonlinear functions parametrised via a parameter vector w. The state equation describes how the state *evolves* over time, while the observation equation describes how the hidden state is *observed*.

State-space models, like any HMM, are generally represented with Bayesian networks (Bayes nets), graphs showing the dependencies between the observed and hidden variables of the model. An example of a Bayes net is shown in Figure 7.2.

When equations in (7.1) are both linear and Gaussian, the problem of estimating the parameters of an SSM from a sequence of observations can be solved using the Kalman filter [76]. Various extensions of the Kalman filter have solved the problem of estimating the parameters in the absence of linearity [209] or normal distributions [85].

SSMs have been extensively used in several fields to characterise a process where the state is unobservable. For example, in ecology they have been used to study and predict the animal movements [134], in control theory to control industrial processes [73], and in neuroscience to estimate the cognitive state of

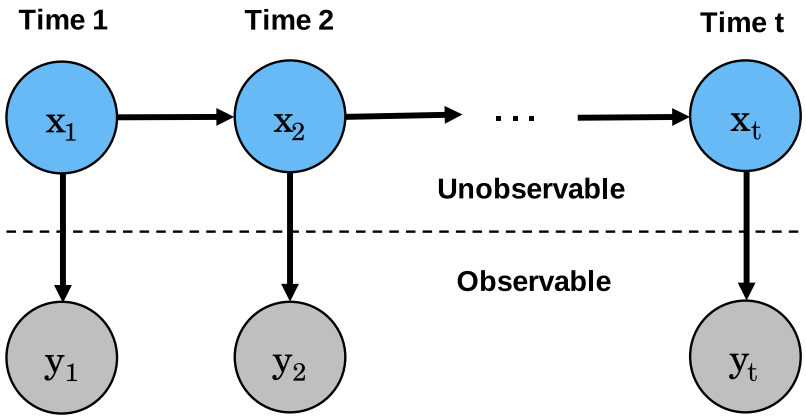

Figure 7.2: Representation of a state-space model using a Bayesian network. Each node represents a model variable (y_t are the observations and x_t the hidden states), while each arrow indicates a dependency between two variables.

the user during a learning task [173, 174, 150, 149]. For example, an SSM model has been used in [174] to characterise learning in behavioural experiments as the probability of a correct response as a function of the trial number. Given the T trials of a behavioural experiment, the SSM developed in [174] was expressed by:

$$\begin{cases} x_t = x_{t-1} + \epsilon_t \\ p(c_t|p_t, x_t) = p_t^{c_t}(1-p_t)^{1-c_t} \end{cases}, \tag{7.2}$$

where c_t denotes the correctness of the response on trial t, ϵ_t are independent Gaussian random variables with mean 0 and variance σ_ϵ^2, and p_t is defined by the logistic equation

$$p_t = \frac{\exp(\mu + x_t)}{1 + \exp(\mu + x_t)}, \tag{7.3}$$

where μ is the chance probability of correct decisions.

In this chapter, we will use and extend the SSM developed in [174] in which the observation model is a *point process*. Point processes are random processes

where realisations are composed by isolated points either in time or in space. The reason behind this choice is that these random processes are a good approximation of what happens in stimulus-response experiments used in ERP-based BCIs, where a stimulus (input) controlled by the experimenter is applied and the response (output) of the human brain (system with a hidden state) is measured, for example via EEG.

7.2.2 Parameter Estimation

This section briefly describes the algorithms used in [174] to estimate the parameters of the state-space model in Equation (7.2).

We firstly set $x_0 = 0$ to set the baseline of the cognitive state of the user before starting the experiment. We then determine μ by using the observation equation in Equation (7.3) to obtain $\mu = \log[p_0(1-p_0)^{-1}]$, where p_0 denotes the probability of a correct response occurring by chance given the experimental setup.

In order to build a forward filter to estimate the state x_t at trial t from the set of observations $N_t = [n_1, \ldots, n_t]$, we need to express the probability density of the state given the observations:

$$p(x_t|N_t) = \frac{p(x_t|N_{t-1})p(n_t|x_t)}{p(n_t|N_{t-1})} \tag{7.4}$$

and the associated one-step prediction probability density obtained using the Chapman-Kolmogorov equation

$$p(x_t|N_{t-1}) = \int p(x_{t-1}|N_{t-1})p(x_t|x_{t-1})dx_{t-1} \tag{7.5}$$

The numerator of Equation (7.4) combines information from the one-step prediction of the state at trial t based on the observation up to through trial $t-1$ (first term) and the observation process (second term). The denominator is simply the normalising constant of the probability density. The one-step prediction density, $p(x_t|N_{t-1})$ is the probability density of the state at trial t given the observations up through trial $t-1$. Equation (7.5) computes this probability density of the state at trial t by "averaging over" the state given the data up to trial $t-1$ defined by $p(x_{t-1}|N_{t-1})$ (first term, i.e., posterior density at $t-1$) and the state transition between trials $t-1$ and t defined by $p(x_t|x_{t-1})$ (second term).

Taken together, Equations (7.4) and (7.5) define a recursion that can be used iteratively to compute the probability of the state given the observations. While this approach might work for low-dimensional models, it becomes less computationally feasible for complex systems [173].

For these reasons, we simplify the problem by computing the Gaussian approximation of Equations (7.4) and (7.5), a process also termed *maximum a posteriori estimation* [174]. A Gaussian probability density is fully defined by its mean and variance. Therefore, to approximate the probability density Equation (7.4) with a Gaussian, we need to compute its mean (or maximum-a-posteriori estimate of x_t) $\hat{\mu}$ and its variance $\hat{\sigma}^2$.

The mean $\hat{\mu}$ describes the maximum of Equation (7.4), while the variance $\hat{\sigma}^2$ defines its curvature. To obtain $\hat{\mu}$, we compute the first derivative of the log of Equation (7.4) with respect to x_t, set it equal to zero and solve for x_t. The variance $\hat{\sigma}^2$ is then obtained by computing the negative inverse of the second derivative of the log posterior probability density with respect to x_t.

Once the Gaussian approximation of Equation (7.4) is given, we can find the

mean $x_{t|t-1}$ (i.e., mean of the state at time t given the states up to time $t-1$) and the variance $\sigma^2_{t|t-1}$ of the Gaussian approximation of Equation (7.5) with standard formula as the integral contains two Gaussian random variables. Given $x_{t-1|t-1}$ and Equation (7.2), we have that

$$x_{t|t-1} = E(x_t|x_{t-1|t-1}) = x_{t-1|t-1} \tag{7.6}$$

and

$$\sigma^2_{t|t-1} = \mathrm{Var}(x_t|x_{t-1|t-1}) = \mathrm{Var}(x_{t-1} + \epsilon_t|x_{t-1|t-1}) = \sigma^2_{t-1|t-1} + \sigma^2_\epsilon. \tag{7.7}$$

With the Gaussian approximation in place, the state-space model of Equation (7.2) could be fully defined by its parameters $\theta = (\mu, \sigma^2_\epsilon, \hat{\mu}, \hat{\sigma}^2)$. These parameters could be estimated by maximum likelihood using the well-known expectation-maximisation (EM) algorithm [29]. Given a set of observations $\mathcal{D} = \{y_1, \ldots, y_T\}$, the maximum likelihood procedure finds the combination of parameters that maximise the likelihood of observing $\mathcal{D}$ and estimating the hidden state $\mathcal{X} = \{x_1, \ldots, x_T\}$ given the set of parameters θ, as described in the following equation:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{T} p(y_i, x_i|\theta) \tag{7.8}$$

Let us define the logarithm of the likelihood as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{T} \log p(y_i, x_i|\theta). \tag{7.9}$$

Because the log is a monotonically-increasing function, the set of parameters

$\hat{\theta}$ that maximise the likelihood also maximise $\mathcal{L}(\theta)$.

We should note that the parameters μ and σ_ϵ^2 are associated to the state-space model itself, while $\hat{\mu}$ and $\hat{\sigma}^2$ are associated to the approximation of the hidden state. Indeed, if the two parameters of the state-space model were known, one could use the Viterbi algorithm [46] to maximise the log likelihood over all possible values of the hidden state X and easily find the values of the other two parameters. Conversely, if the hidden state parameters are known (i.e., all the variables are observable), the computation of the model parameters would be quite easy [51]. However, when all parameters need to be estimated, solving the maximum likelihood (or the maximum log likelihood) problem is usually intractable.

The EM algorithm allows to find the optimal parameters for a lower bound of $\mathcal{L}(\theta)$ [51]. Let $Q(X)$ be a distribution over the hidden variables. We can define a lower bound of $\mathcal{L}(\theta)$ as

$$\mathcal{L}(\theta) > \sum_{i=1}^{K}[Q(x_i)\log p(y_i, x_i|\theta)] - \sum_{i=1}^{K}[Q(x_i)\log Q(x_i)] = \mathcal{F}(Q,\theta). \tag{7.10}$$

When considering the lower bound $\mathcal{F}(Q,\theta)$, we now have two quantities to optimise: (a) the distribution Q, which we want to make the lower bound as more similar to $\mathcal{L}$ as possible, and (b) the set of parameters θ, as our original objective was to find the optimal parameters of the state-space model. The EM procedure iteratively alternates between two steps: the *E step*, where given a set of parameters θ_k finds the best function Q_{k+1} that approximate $\mathcal{L}(\theta)$, and the *M step*, where given a function Q_{k+1} finds the optimal set of parameters θ_{k+1}. The

detailed derivation of the EM algorithm for the model presented in Equation 7.2 can be found in [174].

7.3 Behavioural Model

State-space models have already been used with behavioural experiments to characterise learning from the observations of (a) correctness in a decision and (b) response times [150, 149]. Apart from being related to decision confidence [100], RTs could also indicate the attention level of the user [151]. This section briefly describes the behavioural state-space model developed in [150, 149] and present the results obtained by using that model to estimate the cognitive state of human participants undertaking the realistic visual-search task described in Chapter 5.

The behavioural model defines the unobservable cognitive state of the user x_t with the following state equation:

$$x_t = \rho_0 + \rho x_{t-1} + \upsilon_t, \tag{7.11}$$

where υ_t are independent, zero mean Gaussian random variables with variance σ_υ^2 and $t = 1, \ldots, T$ represent the time steps in which a decision is made. In the case of our visual search task, we have a total of $T = 320$ decisions made by each participant.

The observation model for the RTs is defined as

$$z_t = \log r_t = \alpha + \beta x_t + \epsilon_t, \tag{7.12}$$

where r_t is the RT at trial t, ϵ_t are independent, zero mean Gaussian random

variables with variance σ_ϵ^2, which we assume it is independent from v_t. The parameter α governs the baseline RT, whereas β represents the rate at which the subject reacts as a function of his/her cognitive state. For an experiment in which a subject learns we would expect $\beta < 0$. The use of the logarithmic transformation with the Gaussian error assumption models the empirical observation that larger RTs tend to show greater variability than shorter RTs [149].

Finally, we model the correctness process using a Bernoulli observation model, as the correctness is a binary observation

$$p(c_t|x_t) = p_t^{c_t}(1-p_t)^{1-c_t}, \tag{7.13}$$

where c_t is 1 if the response is correct and 0 if it is incorrect, and p_t is the probability that the process takes the value 1, which is given by

$$p_t = \frac{\exp(\mu + \gamma x_t)}{1 + \exp(\mu + \gamma x_t)}, \tag{7.14}$$

where γ is a modulation parameter which governs the effect of the cognitive state process on the probability of observing the binary outcome, and μ defines the probability of the binary outcome when the state process is zero.

The Bayesian network representing this behavioural model is shown in Figure 7.3. We should note that both observations (i.e., RT and correctness) need to indicate an increase in the cognitive state (which could represent an increase in the attentional level of the user) to make the model predict such change. This means that the cognitive state will rise only when the RT is small *and* the correctness is 1, and decrease when the RT is big *and* the correctness is 0. In the

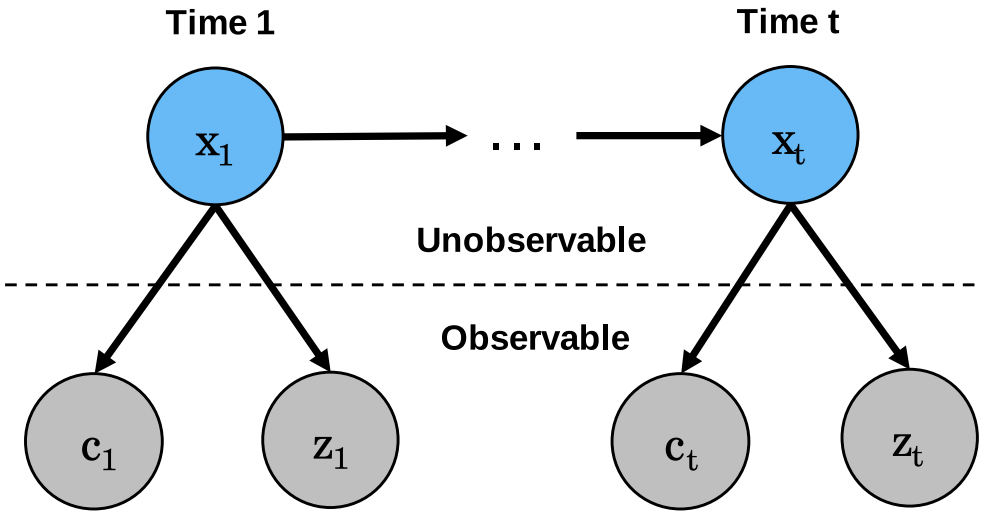

Figure 7.3: Bayesian network representing the behavioural state-space model based on correctness (c_t) and log-transformed RTs (z_t) developed in [149].

other cases (i.e., when the correctness is 1 but the RT is big, or the correctness is 0 and the RT is small), the model will maintain the cognitive state constant.

Let $Z_t = [z_1, \ldots, z_T]$ and $C_t = [c_1, \ldots, c_T]$ be the sequences of log transformations of RTs and decision correctness measures from trials 1 through T, respectively. In order to build a recursive filter to estimate the state x_t at trial t from Z_t and C_t, similarly to what we did for the model based on the sole correctness (see Section 7.2.2), we need to express the probability density of the state given the observations:

$$p(x_t|Z_t, C_t) = \frac{p(x_t|Z_{t-1}, C_{t-1})p(z_t|x_t)p(c_t|x_t)}{p(z_t, c_t|Z_{t-1}, C_{t-1})} \tag{7.15}$$

and the associated one-step prediction probability density (Chapman-Kolmogorov

equation) is

$$p(x_t|Z_{t-1}, C_{t-1}) = \int p(x_{t-1}|Z_{t-1}, C_{t-1})p(x_t|x_{t-1})dx_{t-1} \tag{7.16}$$

The probability densities $p(z_t|x_t)$ and $p(c_t|x_t)$ are the Gaussian and the Bernoulli observation models for the RTs (defined in (7.12)) and the correctness (defined in (7.13) and (7.14)) measures, respectively.

Similarly to the simple correctness-based model described in Section 7.2.2, we estimated the parameters of the behavioural model with the EM algorithm on a participant-by-participant basis. Moreover, similarly to [149] we chose $\rho = 1$ and $\gamma = 1$, to focus the analysis on the estimation of the parameters associated to the RT observations α and β.

7.3.1 Results

Figure 7.4 shows the probability of correct response of each participant at each trial of the realistic visual-search experiment derived using the behavioural model. These probabilities are directly obtained by the cognitive state estimate by the model.

These results confirm that the cognitive state is quite different between participants. For example, participant 1 and 9 seem to slowly increase their cognitive state along the whole experiment, as an effect of experience. Other participants (e.g., 2, 3, 4, 6 and 7) have a peak of performance in sessions 4 and 5 and then their cognitive state drops, probably because of tiredness/boredom. The remaining participants (5, 8 and 10) seem to have constant performance along the experiment.

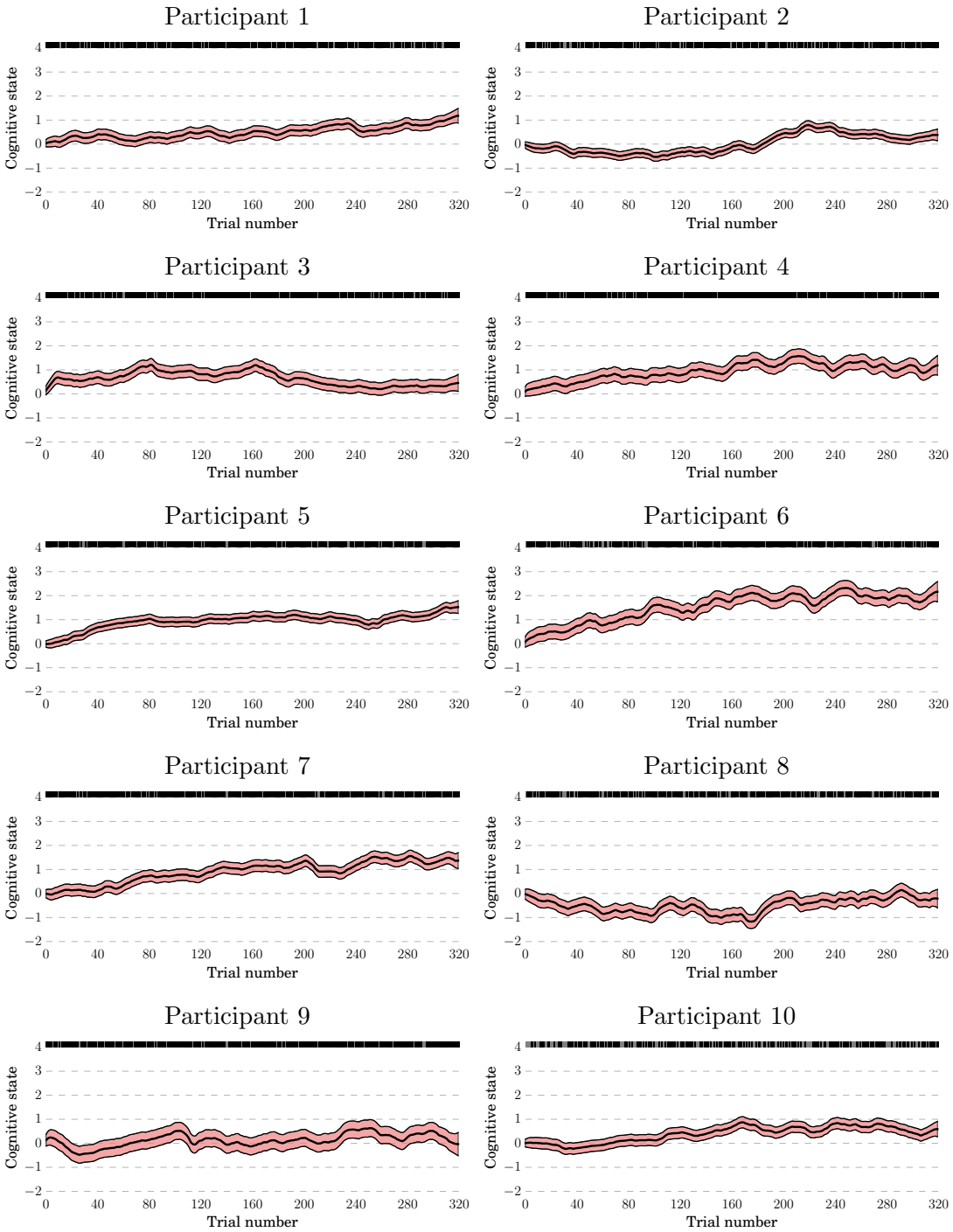

Figure 7.4: Cognitive state evolution for each participant (thick black lines) estimated using the state-space model based on correctness and RTs for the visual search experiment in realistic environments. 95% confidence intervals are shown in light red. Correct (black) and incorrect (grey) decisions for each trial are also shown above each plot.

7.3.2 Between-Trial Comparisons of Performance

Similarly to [149], we compared the performance of learning between pairs of trials to assess how much the cognitive state changes from one trial to another. For each participant and given two trials i and j, we computed the probability that the cognitive state of the observer on trial i is greater than the cognitive state at trial j for all combinations (i, j). To compute these probabilities, we used the Monte Carlo algorithm used in [149].

Figure 7.5 shows a 2D representation of the probabilities that the cognitive state at trial i (abscissas) is greater than the cognitive state at trial j (ordinates) for each participant. The purple areas show the trial comparisons for which $p(x_i > x_j) > 0.95$, while the black areas show the trial comparisons for which such probability is smaller than 0.05. These two areas represent significant variations of the cognitive state of the user along the experiment. Since the cognitive state represents the level of attention and fatigue of the user, we expect to see some purple areas in the middle of the experiment (result of the process of task familiarisation of the participant and high attentional level due to the engagement in the experiment) and some black areas towards the end of the experiment (when the user is likely to be tired).

These results show that, for all participants except for observer 3, there is a high probability that the cognitive state in the trials towards the end of the experiment is higher than the trials at the beginning (cf. purple areas in the bottom-right corner of each plot in Figure 7.5). This confirms the assumption that users increase their attentional level after using the first session to familiarise with the task, then improving their performance due to experience. A different

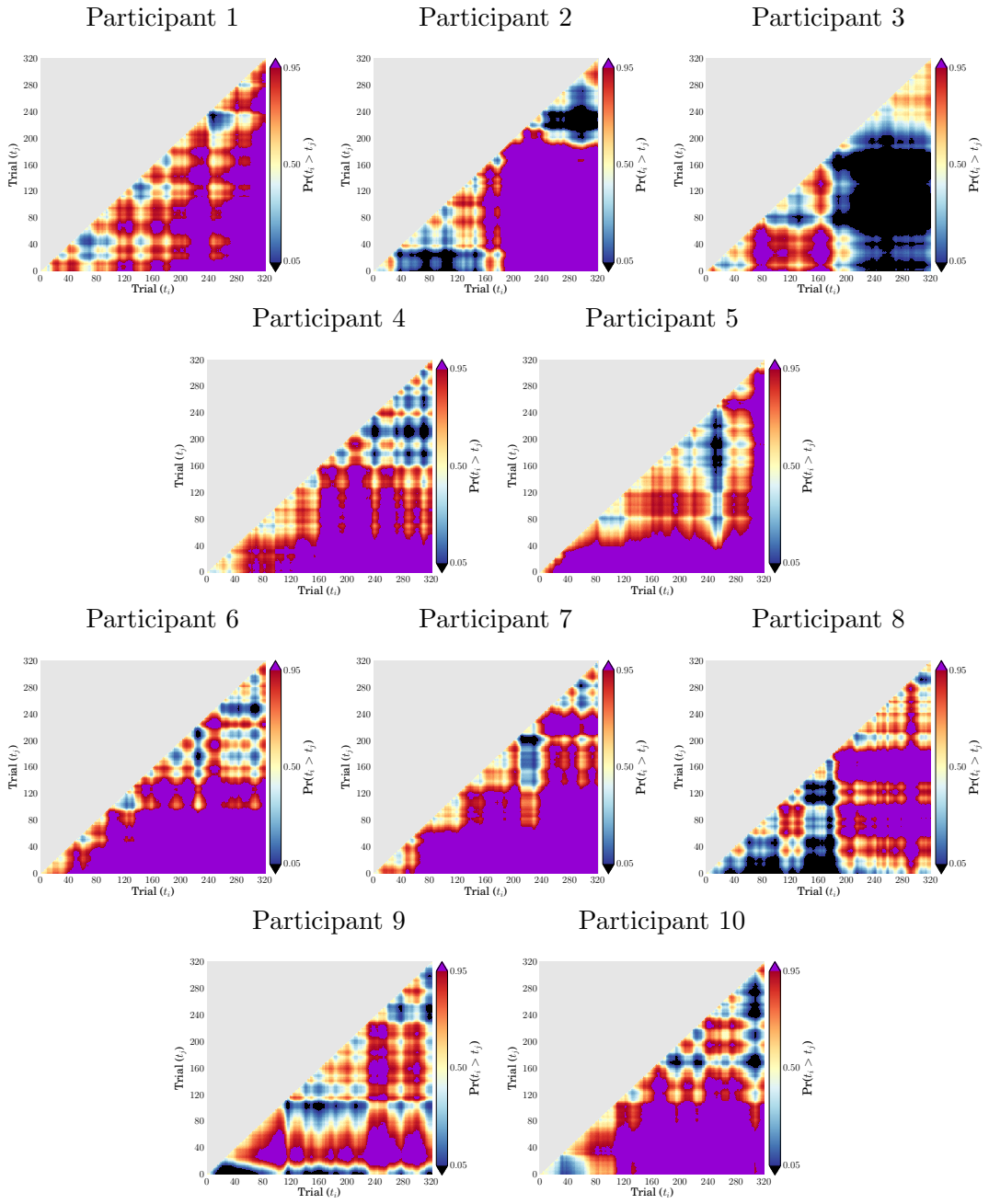

Figure 7.5: Probability $p(x_i > x_j)$ that the cognitive state at trial i (abscissas) estimated using the behavioural state-space model based on correctness and RTs is greater than the cognitive state at trial j (ordinates) for each participant. Comparisons for which this probability is greater than 0.95 or smaller than 0.05 are shown in purple and black, respectively.

situation happens for participant 3, who is very likely to have a low cognitive state in the second half of the experiment with respect to the middle sessions. This suggests that participant 3 started getting tired towards the end of the experiment.

7.4 Neuro-Behavioural Model

This section describes an extension of the behavioural model presented in the previous section that also includes a set of neural features extracted from the EEG signals. We will firstly derive the model mathematically and then we will describe the neural features we have used to estimate the cognitive state. Results obtained with the realistic visual search experiment are also presented.

7.4.1 Observation Model of the EEG Feature

Starting from the behavioural model described in Section 7.3, we firstly need to extend the observation equations to also include a description on how the neural features are observed. Let us assume Ω being the set of different EEG features we want to include in the model. Each feature $e_{j,t}$ is represented by a continuous value at each time step t, with $j = 1, \ldots, |\Omega|$. Similarly to the observation model used for the RTs (see Equation 7.12), the j-th EEG feature recorded at time t is defined as

$$e_{j,t} = \phi_j + \psi_j x_t + \omega_{j,t}, \tag{7.17}$$

where $\omega_{j,t}$ are independent, zero mean Gaussian random variables with variance $\sigma^2_{\omega_j}$ associated to the j-th feature, which we assume it is independent from υ_t and ϵ_t (see Equations (7.11) and (7.12)). The parameter ϕ_j governs the baseline of the j-th EEG feature, whereas ψ_j represents the influence that the cognitive state has on that feature. A positive value of $\psi_j > 0$ means that the cognitive state increases when the j-th EEG feature also increases.

Adding these $|\Omega|$ equations to our behavioural model, we obtain the neuro-behavioural model:

$$\begin{cases} x_t = \rho_0 + \rho x_{t-1} + \upsilon_t \\ z_t = \alpha + \beta x_t + \epsilon_t \\ e_{j,t} = \phi_j + \psi_j x_t + \omega_{j,t}, \ \forall j = 1, \ldots, |\Omega| \\ p(c_t|x_t) = p_t^{c_t}(1-p_t)^{1-c_t} \\ p_t = \frac{\exp(\mu+\gamma x_t)}{1+\exp(\mu+\gamma x_t)}. \end{cases} \tag{7.18}$$

The Bayesian network representing the neuro-behavioural model is shown in Figure 7.6.

7.4.2 Derivation of the Recursive Filter

This section describes how we construct a recursive filter to estimate the state x_t at trial t from the correctness, RTs and EEG features.

Let Z_t and C_t be the sequences of observed RTs and correctness, respectively (as described in Section 7.3) and let $E_{\Omega,t} = \begin{bmatrix} u_{1,1} & \cdots & e_{1,t} \\ \cdots & \cdots & \cdots \\ u_{|\Omega|,1} & \cdots & e_{|\Omega|,t} \end{bmatrix} \in \mathbb{R}^{|\Omega| \times T}$ be the sequences of values for each EEG feature $e_j \in \Omega$ from trials 1 through t. The Equations (7.15) and (7.16) become as follows:

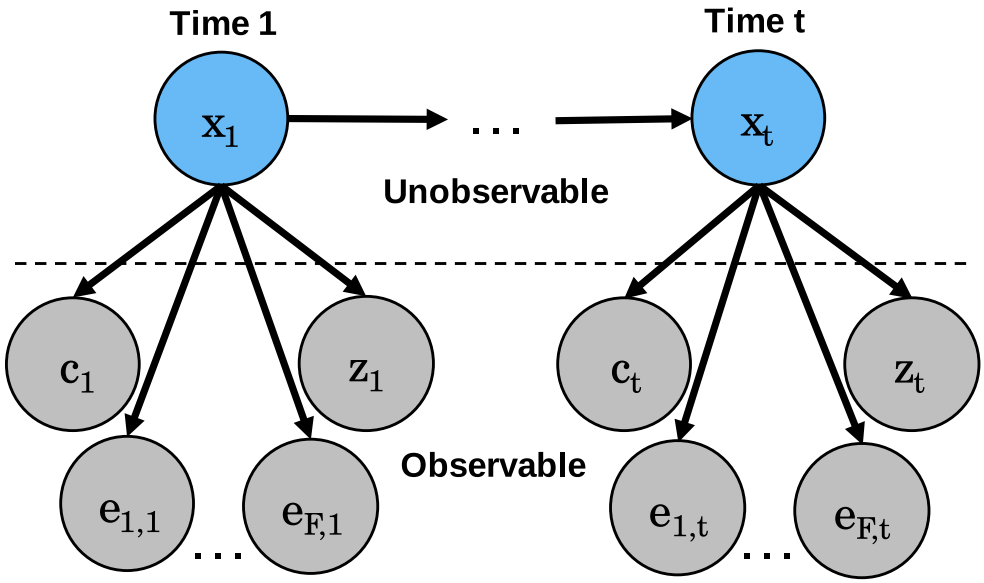

Figure 7.6: Bayesian network representing the neuro-behavioural state-space model based on correctness (c_t), log-transformed RTs (z_t) and $|\Omega|$ EEG features ($e_{f,t}$, with $f \in \Omega$).

$$p(x_t|E_{\Omega,t}, Z_t, C_k) = \frac{p(x_t|E_{\Omega,t-1}, Z_{t-1}, C_{t-1}) \left[\prod_{j=1}^{|\Omega|} p(e_{j,t}|x_t)\right] p(z_t|x_t)p(c_t|x_t)}{p(e_{1,t}, \ldots, e_{|\Omega|,t}, z_t, c_t|E_{\Omega,t-1}, Z_{t-1}, C_{t-1})}, \tag{7.19}$$

$$p(x_t|E_{\Omega,t-1}, Z_{t-1}, C_{t-1}) = \int p(x_{t-1}|E_{\Omega,t-1}, Z_{t-1}, C_{t-1})p(x_t|x_{t-1})dx_{t-1}, \tag{7.20}$$

where $p(e_{j,t}|x_t)$ is the Gaussian observation model for the j-th EEG feature defined in (7.17).

7.4.3 Derivation of the Gaussian Approximation

In order to build the recursive filter to estimate the state x_t at trial t from $E_{\Omega,t}$, Z_t and C_t, we need to follow an approximation process similar to the one used in Section 7.3. To approximate an unimodal probability density $f(x)$ with a Gaussian probability density, we compute its mode $\hat{\mu}$ as the solution of the equation:

$$\frac{\partial \log f(x)}{\partial x}\bigg|_{\hat{\mu}} = 0, \tag{7.21}$$

and its variance $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 = \left[-\frac{\partial^2 \log f(x)}{\partial x^2}\bigg|_{\hat{\mu}} \right]^{-1}. \tag{7.22}$$

We derive the mixed filter algorithm by computing the Gaussian approximation [173] to the posterior density $p(x_t|E_{\Omega,t}, Z_t, C_t)$ in Equation (7.19). At trial t, we assume that the one-step prediction probability density in Equation (7.20) is the Gaussian probability:

$$p(x_t|E_{\Omega,t-1}, Z_{t-1}, C_{t-1}) = (2\pi\sigma^2_{t|t-1})^{-\frac{1}{2}} \exp\{-(2\sigma^2_{t|t-1})^{-1}(x_t - x_{t|t-1})^2\}. \tag{7.23}$$

The probability densities for the EEG and the RTs are the following, respectively:

$$p(e_{j,t}|x_t) = (2\pi\sigma^2_{\omega_j})^{-\frac{1}{2}} \exp\{-(2\sigma^2_{\omega_j})^{-1}(e_{j,t} - \phi_j - \psi_j x_t)^2\}, \tag{7.24}$$

$$p(z_t|x_t) = (2\pi\sigma^2_{\epsilon})^{-\frac{1}{2}} \exp\{-(2\sigma^2_{\epsilon})^{-1}(z_t - \alpha - \beta x_t)^2\}, \tag{7.25}$$

while the probability mass function for the correctness is given by simply rewriting (7.13) as follows:

$$\begin{aligned} p(c_t|x_t) &= p_t^{c_t}(1-p_t)^{1-c_t} = \\ &= \exp(\log(p_t^{c_t}(1-p_t)^{1-c_t})) = \\ &= \exp(\log(p_t^{c_t}) + \log(1-p_t)^{1-c_t}) = \\ &= \exp(c_t \log p_t + (1-c_t)\log(1-p_t)) = \\ &= \exp(c_t \log p_t - c_t \log(1-p_t) + \log(1-p_t)) = \\ &= \exp(c_t(\log p_t - \log(1-p_t)) + \log(1-p_t)) = \\ &= \exp(c_t(\log p_t(1-p_t)^{-1}) + \log(1-p_t)). \end{aligned} \tag{7.26}$$

Substituting Equations (7.23), (7.24), (7.25) and (7.26) into Equation (7.19) gives the following posterior probability density:

$$\begin{aligned} p(x_t|E_{\Omega,t}, Z_t, C_t) \propto \exp\{ &- (2\sigma^2_{t|t-1})^{-1}(x_t - x_{t|t-1})^2 + \\ &- \sum_{j=1}^{|\Omega|}[(2\sigma^2_{\omega_j})^{-1}(e_{j,t} - \phi_j - \psi_j x_t)^2] + \\ &- (2\sigma^2_\epsilon)^{-1}(z_t - \alpha - \beta x_t)^2 + \\ &+ c_t(\log p_t(1-p_t)^{-1}) + \log(1-p_t)\}, \end{aligned} \tag{7.27}$$

where we have ignored the denominator and the other constant terms $(2\pi\sigma^2_{\omega_j})^{-\frac{1}{2}}$, $(2\pi\sigma^2_\epsilon)^{-\frac{1}{2}}$ and $(2\pi\sigma^2_{t|t-1})^{-\frac{1}{2}}$, for $j = 1, \ldots, |\Omega|$.

We can then compute the log posterior probability density as:

$$\begin{aligned}\log p(x_t|E_{\Omega,t}, Z_t, C_t) = &-(2\sigma^2_{t|t-1})^{-1}(x_t - x_{t|t-1})^2 + \\ &- \sum_{j=1}^{|\Omega|}[(2\sigma^2_{\omega_j})^{-1}(e_{j,t} - \phi_j - \psi_j x_t)^2] + \\ &- (2\sigma^2_{\epsilon})^{-1}(z_t - \alpha - \beta x_t)^2 + \\ &+ c_t(\log p_t(1-p_t)^{-1}) + \log(1-p_t).\end{aligned} \tag{7.28}$$

To compute the maximum-a-posteriori estimate of x_t and its associated variance estimate, we compute the first and second derivatives of the log posterior probability density with respect to x_t, which are respectively:

$$\begin{aligned}\frac{\partial \log p(x_t|E_{\Omega,t}, Z_t, C_t)}{\partial x_t} = &-(\sigma^2_{t|t-1})^{-1}(x_t - x_{t|t-1}) + \\ &+ \sum_{j=1}^{|\Omega|}[(\sigma^2_{\omega_j})^{-1}\psi_j(e_{j,t} - \phi_j - \psi_j x_t)] + \\ &+ (\sigma^2_{\epsilon})^{-1}\beta(z_t - \alpha - \beta x_t) + \\ &+ \gamma(c_t - p_t),\end{aligned} \tag{7.29}$$

$$\frac{\partial^2 \log p(x_t|E_{\Omega,t}, Z_t, C_t)}{\partial x_t^2} = -(\sigma^2_{t|t-1})^{-1} - \sum_{j=1}^{|\Omega|}[(\sigma^2_{\omega_j})^{-1}\psi_j^2] - (\sigma^2_{\epsilon})^{-1}\beta^2 - \gamma^2 p_t(1-p_t). \tag{7.30}$$

We now set Equation (7.29) equal to zero and solve for $x_{t|t}$ to obtain the posterior mode or maximum-a-posteriori estimate for x_t. For simplicity, we solve it for $|\Omega| = 1$ (i.e., when there is only one EEG feature), as the generalisation for $|\Omega| > 1$ follows. We obtain:

$$
\begin{aligned}
&\frac{-(x_t - x_{t|t-1})}{\sigma^2_{t|t-1}} + \frac{\psi(e_t - \phi - \psi x_t)}{\sigma^2_\omega} + \frac{\beta(z_t - \alpha - \beta x_t)}{\sigma^2_\epsilon} + \gamma(c_t - p_t) = 0 \\
&x_t \left(\frac{1}{\sigma^2_{t|t-1}} + \frac{\psi^2}{\sigma^2_\omega} + \frac{\beta^2}{\sigma^2_\epsilon} \right) = \frac{x_{t|t-1}}{\sigma^2_{t|t-1}} + \frac{\psi(e_t - \phi)}{\sigma^2_\omega} + \frac{\beta(z_t - \alpha)}{\sigma^2_\epsilon} + \gamma(c_t - p_t) \\
&x_t \left(\frac{\sigma^2_\omega \sigma^2_\epsilon + \sigma^2_\epsilon \sigma^2_{t|t-1}\psi^2 + \sigma^2_\omega \sigma^2_{t|t-1}\beta^2}{\sigma^2_{t|t-1}\sigma^2_\epsilon \sigma^2_\omega} \right) = \frac{x_{t|t-1}}{\sigma^2_{t|t-1}} + \frac{\psi(e_t - \phi)}{\sigma^2_\omega} + \frac{\beta(z_t - \alpha)}{\sigma^2_\epsilon} + \gamma(c_t - p_t) \\
&x_t = \left(\frac{\sigma^2_{t|t-1}\sigma^2_\epsilon \sigma^2_\omega}{\sigma^2_\omega \sigma^2_\epsilon + \sigma^2_\epsilon \sigma^2_{t|t-1}\psi^2 + \sigma^2_\omega \sigma^2_{t|t-1}\beta^2} \right) \left[\frac{x_{t|t-1}}{\sigma^2_{t|t-1}} + \frac{\psi(e_t - \phi)}{\sigma^2_\omega} + \frac{\beta(z_t - \alpha)}{\sigma^2_\epsilon} + \gamma(c_t - p_t) \right].
\end{aligned}
\tag{7.31}
$$

Let us define the gain coefficient $G_t \triangleq \frac{\sigma^2_{t|t-1}}{\sigma^2_\omega \sigma^2_\epsilon + \sigma^2_\epsilon \sigma^2_{t|t-1}\psi^2 + \sigma^2_\omega \sigma^2_{t|t-1}\beta^2}$. Then we obtain

$$
\begin{aligned}
x_t = &\left(\frac{\cancel{\sigma^2_{t|t-1}}\sigma^2_\epsilon \sigma^2_\omega}{\sigma^2_\omega \sigma^2_\epsilon + \sigma^2_\epsilon \sigma^2_{t|t-1}\psi^2 + \sigma^2_\omega \sigma^2_{t|t-1}\beta^2} \right) \frac{x_{t|t-1}}{\cancel{\sigma^2_{t|t-1}}} + \\
&+ G_t \sigma^2_\epsilon \sigma^2_\omega \left[\frac{\psi(e_t - \phi)}{\sigma^2_\omega} + \frac{\beta(z_t - \alpha)}{\sigma^2_\epsilon} + \gamma(c_t - p_t) \right].
\end{aligned}
\tag{7.32}
$$

We then sum and subtract the term $\frac{\sigma^2_\epsilon \sigma^2_{t|t-1}\psi^2 + \sigma^2_\omega \sigma^2_{t|t-1}\beta^2}{\sigma^2_\omega \sigma^2_\epsilon + \sigma^2_\epsilon \sigma^2_{t|t-1}\psi^2 + \sigma^2_\omega \sigma^2_{t|t-1}\beta^2} x_{t|t-1}$ (with G_t in it) to obtain

$$
\begin{aligned}
&x_t = x_{t|t-1} + G_t \left[-(\sigma^2_\epsilon \psi^2 + \sigma^2_\omega \beta^2) x_{t|t-1} + \psi \sigma^2_\epsilon (e_t - \phi) + \beta \sigma^2_\omega (z_t - \alpha) + \sigma^2_\epsilon \sigma^2_\omega \gamma (c_t - p_t) \right] \\
&x_t = x_{t|t-1} + G_t \left[\beta \sigma^2_\omega (z_t - \alpha - \beta x_{t|t-1}) + \psi \sigma^2_\epsilon (e_t - \phi - \psi x_{t|t-1}) + \sigma^2_\epsilon \sigma^2_\omega \gamma (c_t - p_t) \right].
\end{aligned}
\tag{7.33}
$$

Finally, we compute the variance of the Gaussian approximation by replacing Equation (7.30) in Equation (7.22). The one-step prediction and its variance are

obtained as follows:

$$x_{t|t-1} = E(x_t|x_{t-1|t-1}) = \rho_0 + \rho x_{t-1|t-1}, \tag{7.34}$$

$$\sigma^2_{t|t-1} = \mathrm{Var}(x_t|x_{t|t-1}) = \mathrm{Var}(\rho x_{t-1} + \upsilon_t|x_{t|t-1}) = \rho^2\sigma^2_{t-1|t-1} + \sigma^2_\upsilon. \tag{7.35}$$

Putting all together, we obtain the recursive mixed filter algorithm described as follows:

One-Step Prediction
$$x_{t|t-1} = \rho_0 + \rho x_{t-1|t-1}$$

One-Step Variance
$$\sigma^2_{t|t-1} = \sigma^2_{t-1|t-1} + \sigma^2_\upsilon$$

Gain Coefficient
$$G_t = [\sum_{j\in\Omega}(\psi_j^2\sigma^2_{t|t-1}\sigma^2_\epsilon \prod_{i\neq j\in\Omega}\sigma^2_{\omega_i})+ \left(\beta^2\sigma^2_{t|t-1} + \sigma^2_\epsilon\right)\prod_{j\in\Omega}\sigma^2_{\omega_j}\,]^{-1}\,\sigma^2_{t|t-1}$$

Posterior Mode
$$x_{t|t} = x_{t|t-1} + G_t\times \left[\sigma^2_\epsilon\sum_{j\in\Omega}\left(\psi_j(e_{j,t} - \phi_j - \psi_j x_{t|t-1})\prod_{i\neq j\in\Omega}\sigma^2_{\omega_i}\right)+ \left(\beta(z_t - \alpha - \beta x_{t|t-1}) + \gamma\sigma^2_\epsilon(c_t - p_t)\right)\prod_{j\in\Omega}\sigma^2_{\omega_j}\right]$$

Posterior Variance
$$\sigma^2_{t|t} = \left[(\sigma^2_{t|t-1})^{-1} + \gamma^2 p_t(1-p_t) + \sum_{j\in\Omega}[(\sigma^2_{\omega_j})^{-1}\psi_j^2] + (\sigma^2_\epsilon)^{-1}\beta^2\right]^{-1}.$$

7.4.4 Derivation of the EM Algorithm

In this section, we derive the equations of the EM algorithm that has been used for finding the optimal parameters of the neuro-behavioural model. For simplicity, we will consider the initial condition $\rho_0 = 0$ as done in [149].

7.4.4.1 E-step

We use the EM algorithm to compute the maximum likelihood estimates of θ. In order to do that, we need to maximise the expectation of the complete data log-likelihood, which is the joint probability density of E, Z, C and x over the T trials:

$$\begin{aligned} p(E_\Omega, Z, C, x|\theta) = & \prod_{t=1}^{T} p_t^{c_t}(1-p_t)^{1-c_t} \\ & \times \prod_{j\in\Omega}\prod_{t=1}^{T}(2\pi\sigma_{\omega_j}^2)^{-\frac{1}{2}} \exp\{(-2\sigma_{\omega_j}^2)^{-1}(e_{j,t}-\phi_j-\psi_j x_t)^2\} \\ & \times \prod_{t=1}^{T}(2\pi\sigma_\epsilon^2)^{-\frac{1}{2}} \exp\{(-2\sigma_\epsilon^2)^{-1}(z_t-\alpha-\beta x_t)^2\} \\ & \times \prod_{t=1}^{T}(2\pi\sigma_\nu^2)^{-\frac{1}{2}} \exp\{(-2\sigma_\nu^2)^{-1}(x_t-\rho x_{t-1})^2\}, \end{aligned} \tag{7.36}$$

where the first term on the right is defined by the Bernoulli probability mass function in Equation (7.13), the second term is defined by the Gaussian probability density in Equation (7.17) and associated to each EEG feature e_j, the third term is defined by the Gaussian probability density in Equation (7.12), and the fourth term is the joint probability density of the state process defined by the Gaussian model in Equation (7.14).

At iteration $(l+1)$ of the algorithm, in the E-step we compute the expectation of the complete data log likelihood given the observations E_Ω, Z and C across the T trials and $\theta^{(l)} = (\phi^{(l)}_{j\in\Omega}, \psi^{(l)}_{j\in\Omega}, \sigma^{2(l)}_{\omega_{j\in\Omega}}, \alpha^{(l)}, \beta^{(l)}, \sigma^{2(l)}_\nu, \rho^{(l)}, \sigma^{2(l)}_\epsilon, x^{(l)}_0)$, the parameter estimates from iteration l, which is defined as:

$$
\begin{aligned}
&E\{\log[p(E_\Omega, Z, C, x|\theta)]||E_\Omega, Z, C, \theta^{(l)}\} = \\
&= E\left(\sum_{t=1}^{T}\{c_t \log[p_t(1-p_t)^{-1}] + \log(1-p_t)\} \,\middle|\middle|\, E_\Omega, Z, C, \theta^{(l)}\right) \\
&\quad + \sum_{j\in\Omega} E\left[-\frac{1}{2}T\log(2\pi\sigma^2_{\omega_j}) - (2\sigma^2_{\omega_j})^{-1}\sum_{t=1}^{T}(e_{j,t} - \phi_j - \psi_j x_t)^2 \,\middle|\middle|\, E_\Omega, Z, C, \theta^{(l)}\right] \\
&\quad + E\left[-\frac{1}{2}T\log(2\pi\sigma^2_\epsilon) - (2\sigma^2_\epsilon)^{-1}\sum_{t=1}^{T}(z_t - \alpha - \beta x_t)^2 \,\middle|\middle|\, E_\Omega, Z, C, \theta^{(l)}\right] \\
&\quad + E\left[-\frac{1}{2}T\log(2\pi\sigma^2_\nu) - (2\sigma^2_\nu)^{-1}\sum_{t=1}^{T}(x_t - \rho x_{t-1})^2 \,\middle|\middle|\, E_\Omega, Z, C, \theta^{(l)}\right].
\end{aligned}
\tag{7.37}
$$

To evaluate the E-step we have to consider the following terms

$$
\begin{aligned}
x_{t|T} &\equiv E[x_t||E_\Omega, Z, C, \theta^{(l)}] \\
W_{t|T} &\equiv E[x_t^2||E_\Omega, Z, C, \theta^{(l)}] \\
x_{t-1,t|T} &\equiv E[x_t x_{t-1}||E_\Omega, Z, C, \theta^{(l)}],
\end{aligned}
\tag{7.38}
$$

for $t \in \{1, \ldots, T\}$ where the notation $t|T$ denotes the expectation of the state variable at time t given the responses up to time T. To compute these quantities efficiently, we decompose the E-step into three parts [172]: a nonlinear recursive filter algorithm to compute $x_{t|t}$, a fixed interval smoothing algorithm to estimate

$x_{t|T}$, and a state-space covariance algorithm to estimate $W_{t|T}$ and $W_{t,t-1|T}$.

7.4.4.2 Fixed Interval Smoothing Algorithm

Given the sequence of posterior mode estimates $x_{t|t}$ and the variance $\sigma^2_{t|t}$ in Equation 7.36, we use the fixed-interval smoothing algorithm [173] to compute $x_{t|T}$ and $\sigma^2_{t|T}$ as follows:

$$x_{t|T} = x_{t|t} + A_t(x_{t+1|T} - x_{t+1|t}), \tag{7.39}$$

$$A_t = \rho\sigma^2_{t|t}(\sigma^2_{t+1|T})^{-1}, \tag{7.40}$$

$$\sigma^2_{t|T} = \sigma^2_{t|t} + A^2_t(\sigma^2_{t+1|T} - \sigma^2_{t+1|t}), \tag{7.41}$$

for $t = T-1, \ldots, 1$ and initial conditions $x_{t|t}$ and $\sigma^2_{t|t}$.

7.4.4.3 State-Space Covariance Algorithm

The covariance estimate, $\sigma_{t,q|T}$, can be computed from the state-space covariance algorithm and is given as

$$\sigma_{t,q|T} = A_t\sigma_{t+1,q|T} \tag{7.42}$$

for $1 \leq t \leq q \leq T$. It follows that the covariance terms required for the E-step are

$$W_{t|T} = \sigma^2_{t|T} + x^2_{t|T}, \tag{7.43}$$

$$W_{t-1,t|T} = \sigma_{t-1,t|T} + x_{t-1|T}x_{t|T}. \tag{7.44}$$

7.4.4.4 M-step

In the M-step, we maximise the expected value of the complete data log likelihood given by Equation (7.37) with respect of θ^{l+1} obtaining:

(A) State part

$$x_0^{(l+1)} = \rho x_{1|t} \tag{7.45}$$

$$\rho^{(l+1)} = \sum_{t=1}^{T} W_{t-1,t|T} \Big[\sum_{t=1}^{T} W_{t-1|T} \Big]^{-1} \tag{7.46}$$

$$\sigma_v^2 = T^{-1} \sum_{t=1}^{T} [W_{t|T} - 2\rho W_{t-1,t|T} + \rho^2 W_{t-1|T}] \tag{7.47}$$

(B) RT part

$$\begin{aligned} \sigma_\epsilon^{2(l+1)} = {} & T^{-1} \sum_{t=1}^{T} z_t^2 + T\alpha^{2(l+1)} \\ & + \beta^{2(l+1)} \sum_{t=1}^{T} W_{t|T} - 2\alpha^{(l+1)} \sum_{t=1}^{T} z_t \\ & - 2\beta^{(l+1)} \sum_{t=1}^{T} x_{t|T} z_t + 2\alpha^{(l+1)} \beta^{(l+1)} \sum_{t=1}^{T} x_{t|T} \end{aligned} \tag{7.48}$$

$$\begin{bmatrix} \alpha^{(l+1)} \\ \beta^{(l+1)} \end{bmatrix} = \begin{bmatrix} T & \sum_{t=1}^{T} x_{t|T} \\ \sum_{t=1}^{T} x_{t|T} & \sum_{t=1}^{T} W_{t|T} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} z_t \\ \sum_{t=1}^{T} x_{t|T} z_t \end{bmatrix} \tag{7.49}$$

(C) EEG part for feature $j \in \Omega$

$$\begin{aligned} \sigma_{\omega_j}^{2(l+1)} = T^{-1} \sum_{t=1}^{T} e_{j,t}^2 + T\phi_j^{2(l+1)} \\ + \psi_j^{2(l+1)} \sum_{t=1}^{T} W_{t|T} - 2\phi_j^{(l+1)} \sum_{t=1}^{T} e_t \\ - 2\psi_j^{(l+1)} \sum_{t=1}^{T} x_{t|T} e_{j,t} + 2\phi_j^{(l+1)} \psi_j^{(l+1)} \sum_{t=1}^{T} x_{t|T} \end{aligned} \tag{7.50}$$

$$\begin{bmatrix} \phi_j^{(l+1)} \\ \psi_j^{(l+1)} \end{bmatrix} = \begin{bmatrix} T & \sum_{t=1}^{T} x_{t|T} \\ \sum_{t=1}^{T} x_{t|T} & \sum_{t=1}^{T} W_{t|T} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} e_{j,t} \\ \sum_{t=1}^{T} x_{t|T} e_{j,t} \end{bmatrix} \tag{7.51}$$

The algorithm alternates between the E-step of Equation (7.37) and the M-step of Equations (7.45), (7.48) and (7.50), using the filter algorithm, the fixed interval smoothing algorithm and the state-space covariance algorithm to evaluate the E-step. The maximum likelihood estimate of $\hat{\theta} = \theta^{(\infty)}$. The convergence criteria for the algorithm were absolute changes of the parameters of less than 10^2 in consecutive iterations and relative changes of the parameters of less than 10^3 [173].

7.4.5 Selecting the EEG Features

A vital part of the neuro-behavioural model is choosing representations of the EEG signals that best correlate with the latent cognitive state. Several techniques for extracting neural features have been used in the literature, including computing the average power in certain frequency bands and more advanced techniques such as PCA and CSP (see Section 3.6.1).

We decided to start our exploration from *one* EEG feature, namely the average log power in the beta band (16–23 Hz) recorded at electrode Cz. Low values of the lower-beta power have been associated to active thinking and attention [20]. For each stimulus-locked epoch i, the preprocessed EEG signal recorded at electrode Cz s_i has been filtered with a pass-band between 15 and 24 Hz.[1] We used the Welch method [216] to compute the power spectrum of the filtered signal. The neural feature e_i has been computed as the logarithm of the sum of the power spectral density (PSD) between the considered frequencies:

$$e_i = \log \sum_{f=15}^{24} \mathrm{PSD}_i(f).$$

7.4.6 Results

Figure 7.7 shows the cognitive state of each participant obtained using the neuro-behavioural state-space model based on the correctness in the decision, the RT and the EEG feature selected in the previous section.

Let us compare these results with those obtained using the behavioural model (cf. Figures 7.4 and 7.7). For participants 1, 3 and 4, there are no major differences

[1] We used a wider pass-band for the filter than the range of frequencies of interest (16–23 Hz) to reduce transient effects of a non-ideal filter.

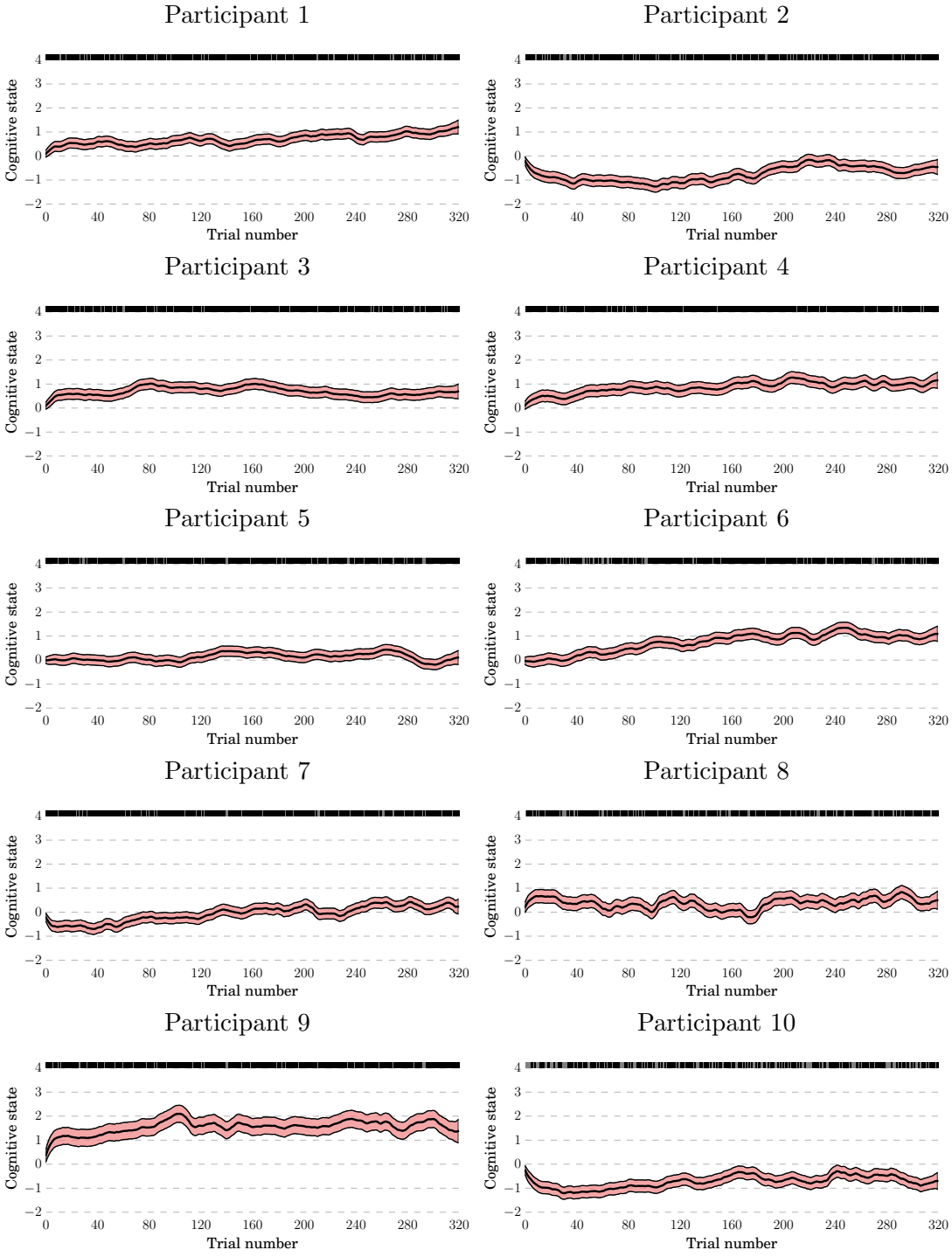

Figure 7.7: Cognitive state at each trial estimated using the neuro-behavioural state-space model based on correctness, RTs and average power in the EEG lower-beta band of each participant for the visual search experiment in realistic environments. 95% confidence intervals are shown in light red. Correct (black) and incorrect (grey) decisions for each trial are also shown above each plot.

in the estimates of the two models. For volunteers 2, 5, 6, 7 and 10, the neuro-behavioural model estimates a much lower cognitive state across the experiment. This is reasonable, for example, for participant 10, as his/her performance across the whole experiment are close to random. For the remaining participants, the neuro-behavioural model estimates they have a higher cognitive state than that estimated by the behavioural model.

These changes are particularly relevant if we consider the potential application of such a model, which is to be able to recognise drops in attention and temporarily ignoring the decisions of certain group members when making group decisions. In the case of participant 10, the performance of the group is likely to be superior if this user is excluded, considering his/her individual performance.

We should note that the neuro-behavioural model is still far from being perfect. For example, it overestimates the cognitive state of participant 9, which has performance close to the average individual performance, and underestimates the cognitive state of participant 4, which is very likely to be correct across the whole experiment. This is likely to be caused by the fact we only used one EEG feature together with behavioural measures, such as RTs and correctness. The addition of extra EEG features (which are supported by the proposed model) and the adoption of more advanced techniques for feature extraction may further improve these results.

7.4.7 Between-Trial Comparisons of Performance

We repeated the trial-by-trial analysis we performed in Section 7.3.2 for the neuro-behavioural state-space model. The results are shown in Figure 7.8.

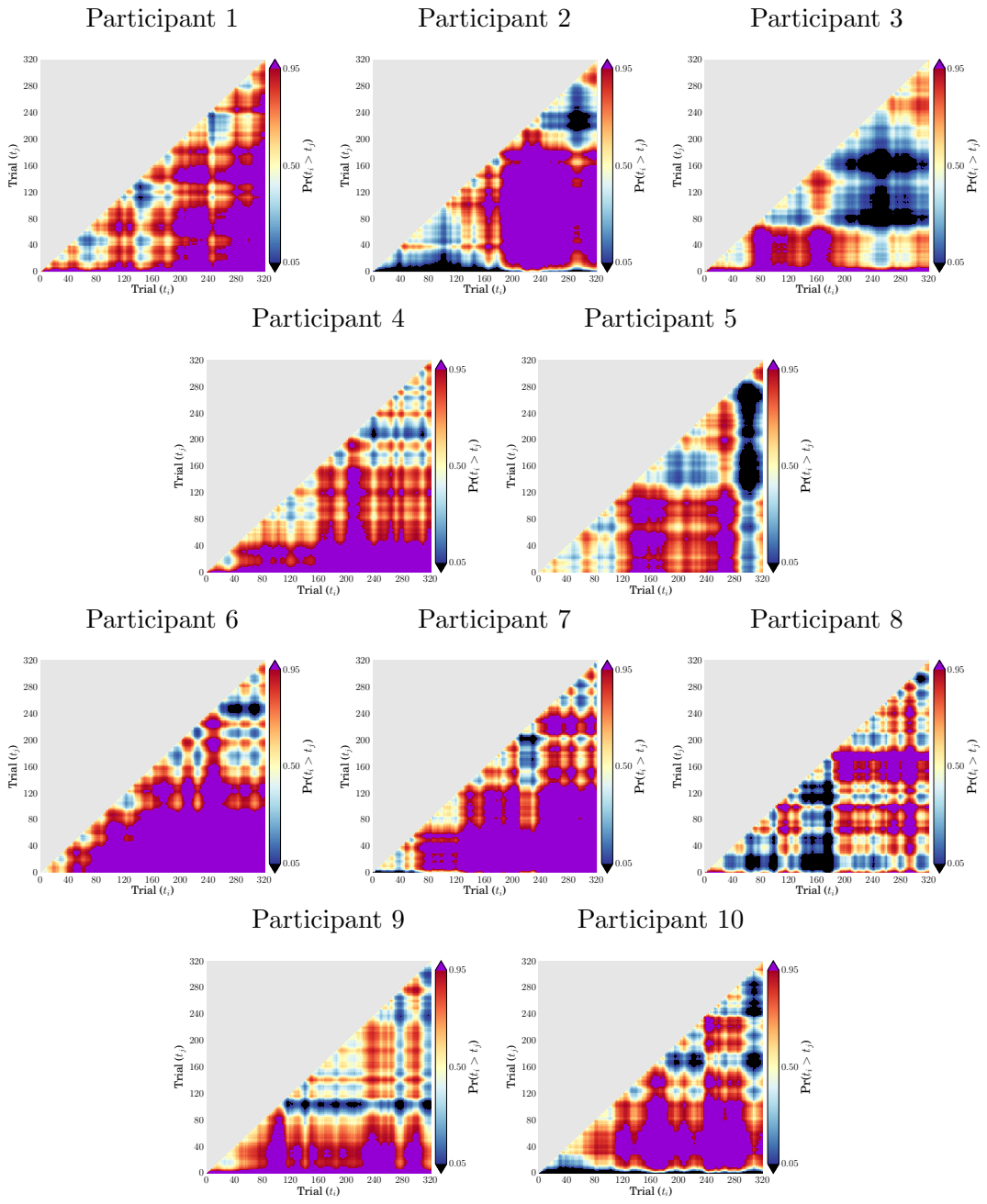

Figure 7.8: Probability $p(x_i > x_j)$ that the cognitive state at trial i (abscissas) estimated using the neuro-behavioural model is greater than the cognitive state at trial j (ordinates) for each participant. Comparisons for which this probability is greater than 0.95 or smaller than 0.05 are shown in purple and black, respectively.

The inclusion of the EEG feature in the model has revealed more processes related to the attention and tiredness of participants. For example, the plot of participant 5 shows that in the last session his/her cognitive state is very likely to be lower than in the previous sessions. This may be due to tiredness, which is indicated by an increase in the beta power and was not visible from only behavioural features. Similar results are obtained for participants 3 (as found also with the behavioural model) and 10. The results for volunteer 9 show that his/her cognitive state reached a maximum in session 3: the probability of the cognitive state in trials in sessions other than 3 to be higher than the cognitive state in trials of that session is very low (cf. black spots along the horizontal line representing session 3 in Figure 7.8). Indeed, this reflects the peak in the cognitive state shown in Figure 7.7.

7.5 Comparison of State-Space Models

This section aims at comparing the goodness-of-fit of the state-space models including different combinations of features developed in this chapter. Since a ground-truth of the cognitive state is not directly available, we evaluated the models on the basis of their ability of predicting the correctness in a decision. In addition to the model based on the sole correctness (Section 7.2), the behavioural model (Section 7.3) and the neuro-behavioural one (Section 7.4), we also studied the performance of a model based on the correctness and the EEG feature (i.e., without RTs). This "neural" model is based on the way the neural feature has been modelled (i.e., similarly to RTs), making it possible to reuse the model described in Section 7.3 by simply replacing the RT observations with the log

Table 7.1: p-values of the likelihood ratio test comparing the goodness of fit of the four models analysed in this chapter based on different combinations of the features, namely the correctness (Cor), RT and correctness (RTCor), EEG and correctness (NeurCor), and EEG, RT and correctness (NeurRTCor). The operator "/" separates the alternative model (first term) from the null model (second term). The difference in the number of free parameters between the two models compared (degrees of freedom) is shown in brackets. p-values below the confidence level 0.05 are shown in boldface and mean that the first model is better than the second one.

Comparison	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
RTCor / Cor (3)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	**0.000**
NeurCor / Cor (3)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	**0.000**	**0.000**
NeurRTCor / Cor (6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**0.000**	1.000	**0.000**
NeurRTCor / RTCor (3)	1.000	**0.000**	1.000	1.000	**0.000**	1.000	1.000	**0.000**	**0.000**	**0.000**
NeurRTCor / NeurCor (3)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

power in the beta band.

Figure 7.9 shows the cognitive state of each participant estimated with the aforementioned four different state-space models.

We used the likelihood ratio test to compare the accuracy of the four models in predicting the correctness in a decision. This test is based on the ratio between the logarithms of the likelihood (Equation 7.9) of two models, one of which (the null model) is a special case of the other (the alternative model). This ratio expresses how many times more likely the data are under the alternative model than under the null one. For each comparison, we computed the test statistic D as:

$$D = 2 \times \log\left(\frac{\text{likelihood for alternative model}}{\text{likelihood for null model}}\right). \tag{7.52}$$

Finally, we computed the probability of the chi-squared approximation of the distribution of D. The results are shown in Table 7.1.

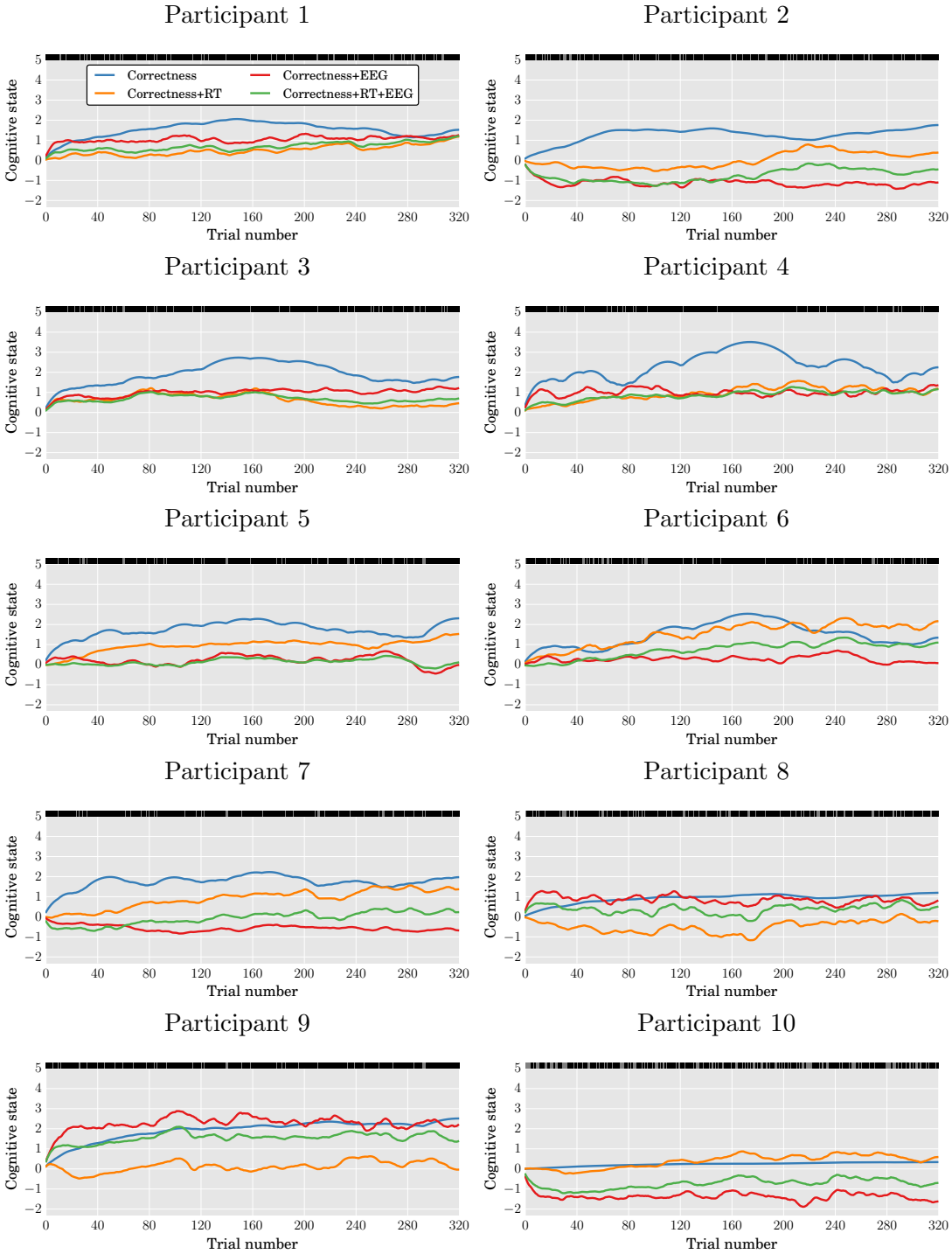

Figure 7.9: Cognitive state processes of each participant estimated with state-space models based on the correctness (blue) and a combination of correctness and (a) RT (orange), (b) neural feature (red), and (c) RT and neural feature (green). Correct (black) and incorrect (grey) decisions for each trial are also shown above each plot.

These results suggest that the state-space models based on the correctness and, either, the RTs or the neural feature are significantly more accurate than the model based on the sole correctness (first two rows in Table 7.1). When combining correctness, RTs and neural features (last three rows), although, the resulting model becomes much better than the models based on the sole correctness or on RTs and correctness for some participants. The neuro-behavioural model performs on par with the model based on the neural feature and the correctness. This suggests that the neural feature used in this chapter and the RTs provide similar information regarding the correctness in the decision, the former being more accurate than the latter with some volunteers.

The likelihood ratio test requires to know the number of free parameters of each model. In the results reported above we empirically-estimated the number of these parameters, although we did not take into account the dependencies between each other. Hence, we have also used the Watanabe-Akaike Information Criterion (WAIC) [215] to select the best state-space model. Similarly to leave-one-out cross-validation (LOO), WAIC is a method for estimating pointwise out-of-sample prediction accuracy (i.e., the quality of the model) from a fitted Bayesian model [50, 205], such as our state-space models. WAIC does not require the estimation of the free parameters of the model, making it a more general method to evaluate a model [50].

WAIC is defined as follows:

$$\text{WAIC} = -2(\widehat{\text{lpd}} - \widehat{p}_{\text{WAIC}}) \tag{7.53}$$

where $\widehat{\text{lpd}}$ is the log pointwise predictive density computed by evaluating the

Table 7.2: WAIC values of the models based on the sole correctness (Cor), the RT and the correctness (RTCor), the neural feature and the correctness (NeurCor), and the RT, the neural feature and the correctness (NeurRTCor) for each participant. The minimum value of WAIC for each volunteer indicates the best model and is reported in boldface, while the second-best model is shown in italics.

Model	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Cor	1.284	**1.311**	1.177	0.971	*1.207*	*1.214*	*1.215*	1.353	*1.210*	**1.376**
RTCor	1.187	*1.475*	1.063	*0.920*	**1.063**	1.383	**1.120**	1.757	1.310	*1.437*
NeurCor	**1.136**	2.513	**1.007**	**0.908**	1.280	1.237	1.915	*1.321*	1.459	2.492
NeurRTCor	*1.142*	2.120	*1.039*	0.930	1.309	**1.154**	1.494	**1.283**	**1.063**	1.787

expectation using draws from the posterior probability, and $\widehat{p}_{\text{WAIC}}$ is estimated effective number of parameters computed using the posterior variance of the log predictive density for each data point. Lower values of WAIC imply higher predictive accuracy of the model [50].

Table 7.2 reports the WAIC of all participants for the four models analysed in this section. The best model of each volunteer is indicated in boldface, while the second best is shown in italics.

The results obtained with WAIC indicate that the two models based also on the neural feature provide the best predictive accuracy of the correctness in the decision on the majority of the participants. For some volunteers (e.g., P2 and P10), the sole correctness is sufficient to provide the best fit. In summary, these results show that there is no best model across participants, suggesting that the selection of the best features for estimating the cognitive state should be conducted on a participant-by-participant basis.

7.6 Conclusions

This chapter has explored the possibility of using first-order state-space models to estimate the cognitive state of a user engaged in a decision-making task from a series of neural and behavioural observations. Detecting changes in the cognitive state could reveal variations in the attentional level and fatigue, which are likely to affect decision-making performance. Our aim was to investigate whether or not our cBCI for group decision making equipped with such models could improve its performance by temporarily excluding the group's members with a low cognitive state from contributing to the group decision, as they are more likely to make an incorrect choice.

We introduced four state-space models based on different combinations of observations, namely the correctness in the decision (Section 7.2), the correctness and RT (Section 7.3), the correctness, RT and a neural feature represented by the log power in the EEG lower-beta band (Section 7.4), as well as a model based on the correctness and the EEG feature. We applied these models to the visual search experiment with realistic stimuli described in Chapter 5 and compared their performance in Section 7.5.

Similar behavioural state-space models have been developed in the literature to track the cognitive state, although they have mainly been applied to learning experiments with animals [149]. In that domain, it is easier to assess the performance of the model, as you can clearly identify when the user or the animal has learnt the task by tracking the correctness in the decisions. Here we track the cognitive state of *human participants* engaged in a *target-detection task including uncertainty*, where even if the volunteer has learned the task properly, he/she can

still make an erroneous decision due to the intrinsic difficulty of the task at hand.

The preliminary results described in this chapter suggest that the integration of an EEG feature in the state-space model could improve the prediction of the cognitive state for certain participants. However, the quantitative analysis conducted with the likelihood ratio test and the WAIC suggests that every participant requires a different combination of features to achieve the best prediction of the cognitive state. Despite these interesting results, we should note that the investigation conducted in this chapter was very preliminary and had the aim of starting exploring the application of state-space models to cBCIs for decision making. Further research is therefore required before being able to draw any conclusions.

Chapter 8

Augmenting Group Performance in Face Recognition

This chapter explores the possibility of using the proposed cBCI to improve performance on face recognition, a task with a broad range of applications in security. Part of the material in this chapter has been published in [196].

8.1 Introduction

Face recognition is a vital task in our everyday lives, especially when applied to security contexts. As seen in Section 2.6.1, BCIs have been used to improve human performance in this taxing and challenging task in a number of experiments. However, the encouraging performance of those BCIs were obtained by performing a particular type of face recognition, that is seeing a sequence of *individual faces* and deciding which ones were target faces. In a real environment, we usually deal with pictures or video frames of *crowded scenes*, possibly taken

from *different viewpoints*, where faces could even be partially occluded. This is the situation in which automatic face recognition usually fails and where BCIs could potentially augment human performance.

This chapter explores the possibility of using the cBCI described in Chapter 3 with a realistic face recognition experiment, where participants have to decide whether a target person was present or not in an image of a crowded environment shown for a very limited time. The aims of this additional experiment are (a) testing the performance of the cBCI described in Chapter 3 with a face recognition task using realistic stimuli and comparing them with those obtained with traditional groups; (b) studying whether confidence reported by participants after each decision could be used to make better group decisions than the cBCI in face recognition; (c) investigating whether traditional and cBCI groups where participants are exposed to different sources of information about the same scene are more accurate than groups where each participant sees the same image.

The chapter is organised as follows. Section 8.2 presents the experimental setup and how group decisions have been obtained in the single and multi-viewpoint approaches. Results are then presented and discussed in Section 8.3. The chapter ends with Section 8.4 drawing some conclusions.

8.2 Methodology

8.2.1 Participants

We gathered data from 10 healthy participants (mean age $\pm$ standard deviation $= 37.8 \pm 4.8$ years old, 7 females, all right-handed) with normal or corrected-to-

normal vision and no reported history of epilepsy.

In addition to the base rate of £16, volunteers were paid an additional rate a_r which depended on their performance as follows:

$$a_r = \begin{cases} £0 & \text{if} \quad acc < 60\% \\ £2 & \text{if} \quad 60\% \leq acc < 80\% \\ £4 & \text{if} \quad acc \geq 80\% \end{cases}$$

where *acc* was the average performance (in %) of the participant across the experiment. The additional rate was adopted in order to further encourage volunteers to focus on the task and achieve the maximum performance.

8.2.2 Experiment

The experiment consisted of a face recognition task where participants had to decide whether a target person was present or not in a picture of a crowded scene shown for a limited amount of time.

The images required for this experiment have been gathered from the sequences P2E_S5 and P2L_S5 of the ChokePoint dataset [226], which was designed for person identification under real-world surveillance conditions. The two sequences consisted in 29 people (six female) walking indoor and passing through two different portals. Three cameras were positioned at the top-left (L), top-center (C) and top-right (R) of each portal, respectively, so that every scene was described by three pictures of size of 800×600 px^2 taken from different viewpoints. Each image contained between 2 and 11 faces.

Since in video sequences consecutive frames contain similar information, we

randomly sampled the 700+ images available in each sequence to select 48 scenes represented by one image for each viewpoint. We then shuffled the selected pictures. This procedure allowed to reduce the possibility that participants used previous knowledge to make decisions. Each image has been converted to greyscale and its histogram has been equalised. Therefore, our dataset was composed of $48 \times 3 \times 2 = 288$ images. The first three rows of Figure 8.1 show a representative image for each sequence and viewpoint.

In each sequence, a different person has been chosen as "target" – see Figure 8.1(bottom). The images have then been labelled as "target" or "non-target" depending on the presence or not of the target person. For each sequence, a total of 36 images (12 per viewpoint) were labelled as "target" and the remaining 108 (36 per viewpoint) as "non-target".

The experiment was split into six sessions of 48 trials each. A session included the presentation of all images taken from a specific combination of sequence and viewpoint, namely (1, L), (1, C), (1, R), (2, L), (2, C), (2, R). Target images were shown in 25% of the trials. The images of each session were shuffled and presented in the same order for each participant, while the order of the sessions was randomised across volunteers. Hence, each stimulus selected as explained before was used exactly once.

Sessions started with a display showing the cropped face of the target person assigned to that session (Figure 8.1(bottom)) and the participant was asked to memorise it. When ready, the user pressed the left mouse button to start the 48 trials of that session. Figure 8.2 shows the sequence of stimuli presented in each trial, which follows the protocol described in Section 3.2 except for the mask, which was not used for this experiment. After the initial fixation cross,

Figure 8.1: Example of images used in the face recognition experiment for the two sequences (columns) and the three viewpoints (first three rows). The last row shows the cropped face of the target person assigned to each sequence.

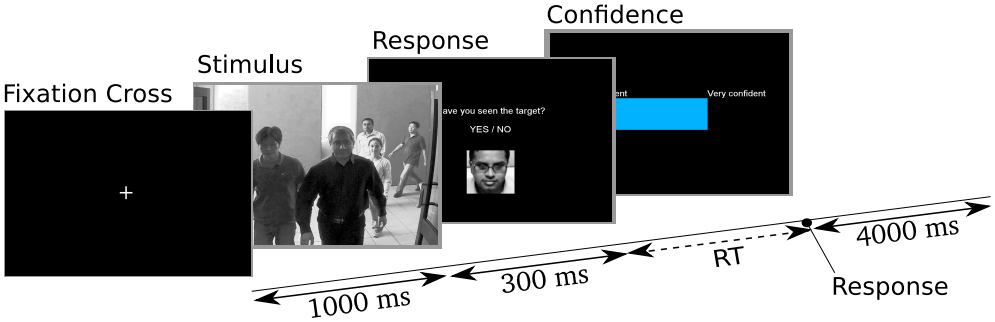

Figure 8.2: Sequence of displays presented in each trial of the face recognition experiment.

an image of a crowded scene was presented for 300 ms in full screen, subtending approximately 14.4 degrees horizontally and 11.0 degrees vertically. After that, a screen showing the target face associated to that session was shown and the user had to decide, as quickly as possible, whether or not the target person was present in the scene, by clicking the left or the right mouse buttons, respectively. After indicating their decision, the participants were asked to indicate the degree of confidence in that decision (0–100%) using the mouse wheel (i.e., scrolling up/down to increase/decrease the confidence by 10%) within a time window of 4 seconds.

The experimental session started with briefing and preparation of the volunteers. Then, two training sessions of 10 trials each were undertaken by the participants to familiarise with the task. Preparation and practice took approximately 45 minutes, while the experiment took about 25 minutes. Participants were comfortably seated at about 80 cm from a LCD screen.

8.2.3 Making Group Decisions

Data were acquired and preprocessed as explained in Chapter 3. For this experiment, we set $p_b = 14$ Hz, $s_b = 16$ Hz and the final sampling rate $s_r = 32$ Hz. Therefore, each stimulus- and response-locked epoch was represented by 48 time samples for each of the 64 EEG channels used.

As done in previous experiments (e.g., see Chapter 6), we split the dataset into a training and a test sets using 10-fold cross-validation. We then used the training set to compute the LTCCSP matrices for the two types of epochs to extract the neural features – see Section 5.2.3. Hence, the cBCI used for the face recognition experiment used 5 features to estimate the decision confidence: 2 LTCCSP neural features extracted from each type of epochs (i.e., stimulus-locked and response-locked) and the RT. Once the BCI confidence was estimated, we computed the confidence weights w by using the negative exponential weighting function described in Equation (3.3).

Group decisions were then made as described in Section 3.8 by using the sign of the weighted sum of members' decisions, where the weights were either the confidence reported by the participants or the confidence weights computed by the cBCI. Group performance were then validated as described in Section 3.9.

In this chapter we tested two approaches for forming groups: a "traditional" one, which has been used in the other experiments and assuming all members of the groups are exposed to the *same* stimuli, and a "multi-viewpoint" approach, where group members are exposed to different sources of information (i.e., images of the same scene taken from different viewpoints). The following sections describe in details these two approaches.

8.2.4 Traditional Approach

Similarly to what has been done in other experiments, we simulated group decisions in which each group's member was exposed to the *same* stimulus. To do so, we saved the order in which the experiment's sessions had been presented (see Table 8.1) to allow reordering the stimuli offline.

With the 10 participants, we were able to assemble $\binom{10}{m}$ groups of size m, for $m = 2, 3, \ldots, 10$. Hence, we computed group decisions for 45 groups of size 2, 120 groups of size 3, and so on.

In this experiment, the stimuli within each session had some shared features (i.e., the sequence and the viewpoint). Hence, we also compared individual and group performance between different sessions. In particular, we looked into the error distributions associated to the three viewpoints (L, C, R), in order to assess whether participants performed better from a certain viewpoint.

8.2.5 Multi-Viewpoint Approach

One of the aims of this experiment was to investigate whether exposing participants to different source of information would improve group performance, as suggested by the literature on group decision making [181]. As described in Section 8.2.2, each scene selected from each sequence was presented in three sessions from different viewpoints. In this experiment, we also simulated group decisions where each group's member was exposed to stimuli representing the same scene seen by other participants but taken from a *different* viewpoint.

When forming groups of size m, we guaranteed that none of the viewpoints was over-represented, in the sense that the number of members viewing images

Table 8.1: Order of the sessions in which each participant has undertook the face recognition experiment. Each session is described by the number of the sequence from which the stimuli has been gathered (i.e., 1 or 2) and the viewpoint of the camera (i.e., "L" for left, "C" for central, "R" for right).

	Session					
Participant	1	2	3	4	5	6
1	(1, R)	(1, L)	(2, L)	(2, C)	(2, R)	(1, C)
2	(2, C)	(1, L)	(1, R)	(1, C)	(2, L)	(2, R)
3	(1, C)	(2, L)	(1, R)	(1, L)	(2, C)	(2, R)
4	(1, R)	(2, R)	(2, C)	(1, C)	(2, L)	(1, L)
5	(2, L)	(1, C)	(2, R)	(2, C)	(1, L)	(1, R)
6	(1, L)	(2, L)	(1, R)	(2, C)	(1, C)	(2, R)
7	(1, L)	(1, R)	(2, L)	(2, C)	(2, R)	(1, C)
8	(2, R)	(2, C)	(1, L)	(1, R)	(1, C)	(2, L)
9	(2, C)	(1, L)	(1, R)	(2, R)	(2, L)	(1, C)
10	(1, R)	(1, C)	(2, C)	(2, R)	(1, L)	(2, L)

from a particular viewpoint never differed by more than 1 from the number of participants viewing images from any other viewpoint. Due to this constraint, the number of possible ways to combine viewpoints v_m for each group size was equal to 1 for $m = 3, 6, 9$ and equal to 3 for the other values of m. The number of groups of size m we could assemble with our $N = 10$ participants was given by the m-permutations of N multiplied by the number of combinations of viewpoints, namely $v_m \frac{N!}{(N-m)!}$. Hence we had 270 groups of size 2, 720 groups of size 3, and so on. The performance of each group was computed on a third of the total number of trials, as only the stimuli from a specific viewpoint were used for the simulation.

It should be noted that the order of the sessions was randomised between participants (see Section 8.2.2) and, therefore, we could consider the samples of

the statistical test used to compare the group performance independent. For this reason, we still use the Wilcoxon signed-rank test for this purpose.

8.3 Results

8.3.1 Individual Performance

Figure 8.3 shows the error rates of each participant in the experiment and the fraction of the overall error rates due to each viewpoint.

The average error rate across participants for the whole experiment was (mean $\pm$ standard deviation) $27.74 \pm 11.98\%$, showing that the face recognition task was extremely difficult for an individual. When considering each viewpoint separately, the average performance across participants was $28.12 \pm 12.25\%$, $28.02 \pm 13.25\%$ and $27.08 \pm 12.07\%$ for the left, center and right camera, respectively.

The average performance was quite similar for the three viewpoints. Indeed, a Kruskal-Wallis test comparing the error rates with each viewpoint showed no statistical differences ($p > 0.7$ for all combinations).

8.3.2 Group Decisions Made from the Same Viewpoint

Figure 8.4 shows the average error rates across all the trials of the experiment for groups of different size making decisions using the standard majority rule (gray line), the confidence-based weighted majority (blue line) and the cBCI-based weighted majority (red line) when participants were seeing images from the same viewpoint.

The results show that the two confidence-based methods perform much better

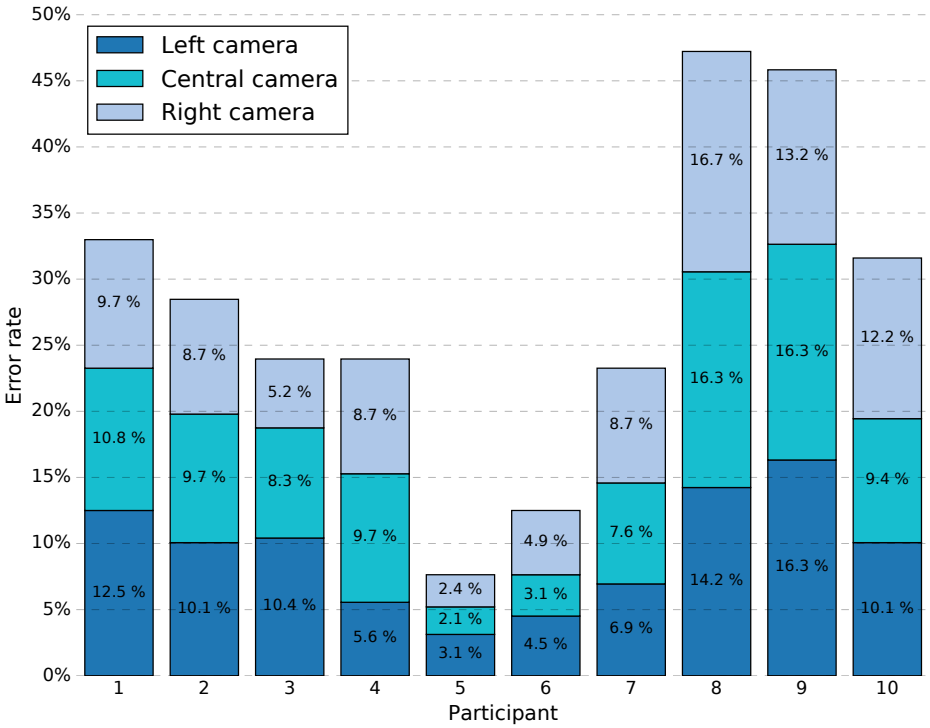

Figure 8.3: Mean error rates for each participant across the 288 trials. The fraction of the overall error rates due to each viewpoint is also indicated.

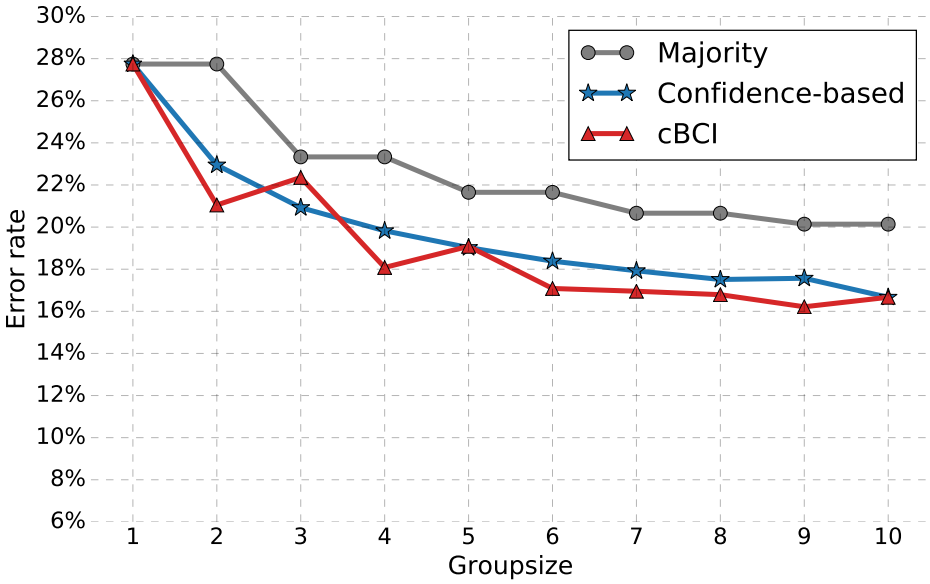

Figure 8.4: Error rates made by groups of different size using the three methods analysed when participants were exposed to stimuli of the same viewpoint.

than the simple majority rule, which confirms our previous findings with other visual experiments [143, 199]. Interestingly, the confidence-based methods do not only improve the performance of even-sized groups, as we were expecting due to their intrinsic ability to break ties in a better way than making a random decision (i.e., the strategy adopted by the majority rule). They also boost the performance of odd-sized groups, hence making their adoption even more advantageous.

The confidence values reported by the participants seem to be quite accurate in predicting when they are making the correct decision. Weighing individual decisions according to subjective estimates reduces the error rates by at least 2% for all group sizes when compared to the error rates obtained with standard majority.

When the decision confidence is estimated by the cBCI using the neural signals and the RTs, however, groups of size 2, 4, 6, 7, 8, 9 are able to further reduce error rates when compared to groups using the confidence values reported by each participant. Particularly interesting is the improvement provided by the cBCI to the performance of pairs, as these are the groups more likely to be used in practice. The cBCI reduces the error rates of traditional pairs (making decisions with the standard majority rule) from 27.7% down to 20.9%.

When analysing the results of odd-sized groups in Figure 8.4, one may wonder why the error rates of groups of size 3 and 5 are higher than the error rates of smaller groups. Ties do not occur in odd-sized groups and, so, to improve performance the cBCI has to allow a minority of users to decide on behalf of the group. For example, in a situation where two group's members made the incorrect decision and one group member made the correct one, the group will make the correct decision only if the cBCI is able to assign a confidence value to

the correct group's member that is higher than the sum of the confidence values assigned to the erroneous group's members. However, for small odd-sized groups this task is quite hard considering the distribution of cBCI weights for the two classes (see Section 8.3.5). This leads to cBCI performance that is closer to (but still significantly better than) the performance obtained by traditional groups.

To compare further the performance of different group sizes making decisions with the three methods analysed, we used the Wilcoxon signed-rank test to compare the different error distributions. The p-values of the Wilcoxon test comparing the overall performance are shown in Table 8.2. It is clear that the performance obtained by the two confidence-based methods is statistically significantly better than that obtained with traditional majority-based groups for all meaningful group sizes (we should note that it is not possible to achieve statistical significance for groups of size 10 as we only have one sample).

When comparing the two confidence-based methods together, we can see from Table 8.2 that the cBCI and the confidence-based methods are complementary, but the cBCI yields significantly better decisions in 6 out of 8 group sizes, while the confidence-based is significantly better than the cBCI only for groups of size 3. The two methods perform on a par for groups of size 5.

These results suggest that both confidence-based methods provide significant improvement in group performance, but the cBCI should be preferred due to its primacy for pairs (the most practical group) and for bigger group sizes (the ones achieving the lowest error rates).

Let us now look at the results obtained by groups when using only the subset of stimuli from one of the viewpoints. Figure 8.5 shows the mean error rates obtained by groups adopting the three decision methods analysed in this chapter

Table 8.2: Statistical comparison of methods for group decisions made using all stimuli for different group sizes. The table reports the p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes adopting different decision methods: standard majority, weighted majority based on the reported confidence (ConfidenceMajority), and cBCI-based weighted majority (cBCI). The p-values below the Bonferroni-corrected statistical significance level $0.05/4 = 0.013$ are in bold face. Sample sizes (the number of groups of each size) are indicated in the last row of the table.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is ConfidenceMajority better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0026**
Is ConfidenceMajority better than cBCI?	0.9968	**0.0000**	1.0000	0.1802	1.0000	1.0000	0.9998	0.9969
Is cBCI better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0029**
Is cBCI better than ConfidenceMajority?	**0.0033**	1.0000	**0.0000**	0.8200	**0.0000**	**0.0000**	**0.0003**	**0.0045**
Sample size	45	120	210	252	210	120	45	10

and using only the subset of the recorded trials associated to a specific viewpoint.

When we look at the error rates obtained with standard majority reported in Figure 8.5 (grey), the group error rate decreases much faster as the group size grows for the right viewpoint than for the centre and the left ones. This is due to the fact that for one sequence of stimuli, people were coming from the top-right corner and, therefore, it was easier for the users to spot the target face from this viewpoint. Also, in the decisions made from the right viewpoint, participants are also more precise in estimating their degree of confidence. In fact, when we compare the performance using the Wilcoxon signed-rank test (Table 8.5), the method based on the reported confidence significantly outperforms the other two methods for almost all group sizes for that viewpoint.

The error rates of groups making decisions using the standard majority from

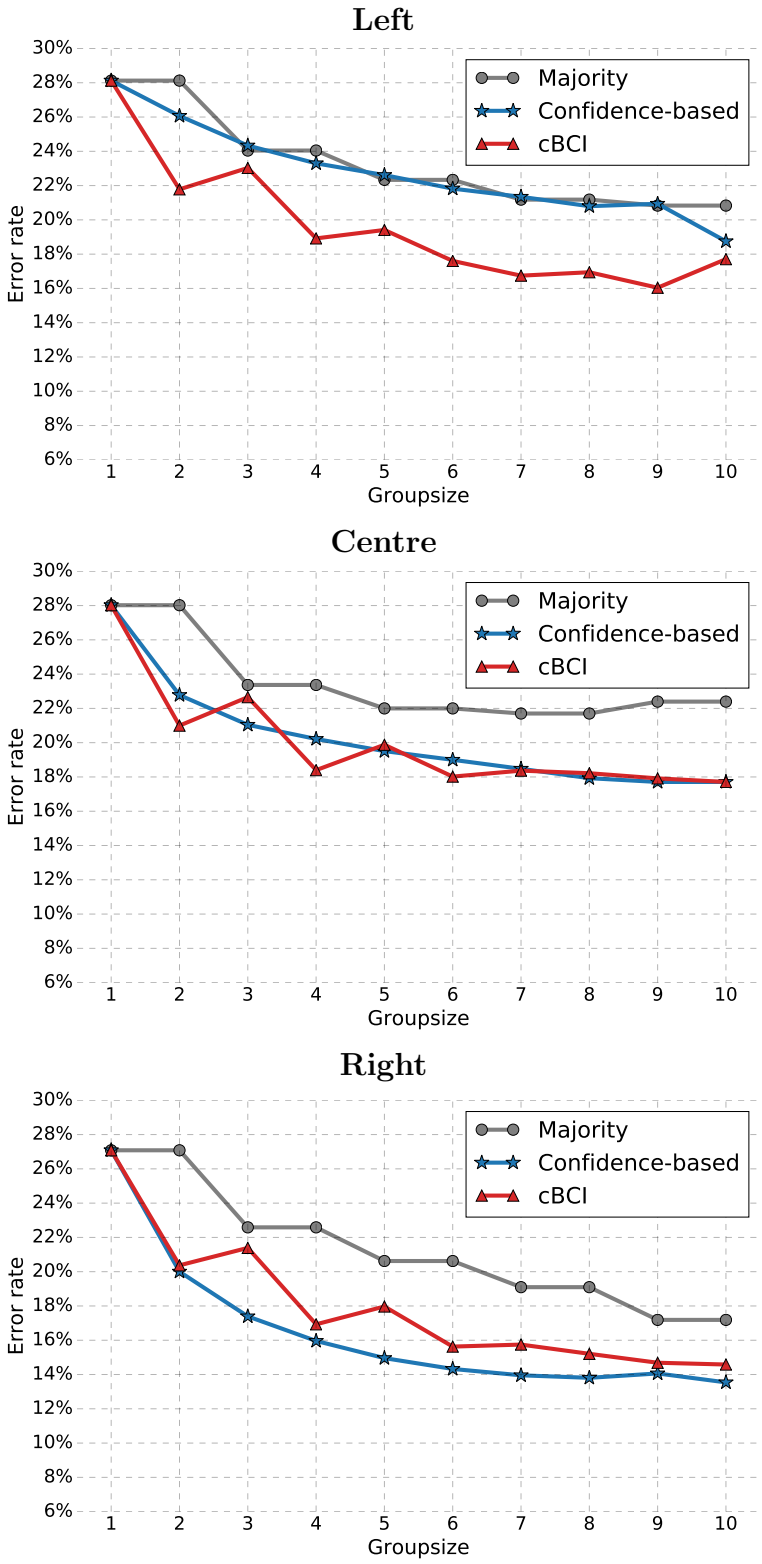

Figure 8.5: Error rates made by groups of different size using the three methods analysed in the study when participants were exposed to stimuli of the left (first row), centre (second row) and right (last row) viewpoints.

the central viewpoint decrease up to group size 5 and then there is no advantage in performance in adding extra group members, with error rates that actually becomes higher than smaller groups for groups of size 9 and 10. This confirms that, in certain circumstances, smaller groups are better than bigger ones [77].

When comparing the performance of the two confidence-based methods with that obtained using the majority rule for the three viewpoints using the Wilcoxon signed-rank test (Tables 8.3–8.5), we can see that the former are almost always significantly better than the latter. However, we should note that for the left viewpoint, the confidence-based method achieves performance that are very similar to the one obtained by traditional groups (statistical differences are present only for groups of size 2, 4 and 6). This is likely to be the other side of the coin of our previous argument: in one of the sequences, people are walking from the top-right corner of the image and, therefore, the left viewpoint is the one containing the lowest information and, therefore, providing more uncertainty.

If we now focus on the performance of the two confidence-based methods themselves, we can see that the cBCI provides a robust and significant improvement over the majority rule in all viewpoints and for all group sizes (i.e., compare the shape of the red curves in Figure 8.4). On the contrary, the method based on the confidence values reported by the participants varies its performance quite a lot depending on the viewpoint. The confidence-based group decisions are significantly better than the cBCI-based ones from the right viewpoint for all group sizes. When considering the central viewpoint, the two confidence-based methods are complementary, with performance on a par for groups of size 2, 7, 8, 9, significantly better performance for the method based on the reported confidence for group sizes 3 and 5, and significantly better performance for the cBCI for groups

Table 8.3: p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes using only the stimuli from the *left* viewpoint adopting the three methods analysed in this chapter. The p-values below the Bonferroni-corrected statistical significance level $0.05/4 = 0.013$ are in bold face. Sample sizes are indicated in the last row of the table.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is ConfidenceMajority better than Majority?	**0.0014**	0.6430	**0.0002**	0.8986	**0.0004**	0.8623	0.0671	0.6374
Is ConfidenceMajority better than cBCI?	1.0000	0.9982	1.0000	1.0000	1.0000	1.0000	1.0000	0.9983
Is cBCI better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0028**
Is cBCI better than ConfidenceMajority?	**0.0000**	**0.0019**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0024**
Sample size	45	120	210	252	210	120	45	10

of size 4 and 6. Finally, when participants see the stimuli from the left viewpoint, confidence-based group decisions are significantly worse than the cBCI-based ones for group sizes 2, 4, 6, 7, 8, 9, significantly better only for groups of size 3 and on a par for groups of size 5.

These results suggest that the reported confidence could be a good predictor of correctness, but it is risky as in some circumstances it is unreliable. On the other hand, the cBCI is able to provide a good estimate of the decision confidence independently from the viewpoint, allowing groups of isolated users to significantly improve their performance.

8.3.3 Group Decisions Made from Different Viewpoints

Figure 8.6 shows the average performance of groups of different sizes when each group member was exposed to stimuli representing the same scene seen by his/her

Table 8.4: p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes using only the stimuli from the *centre* viewpoint adopting the three methods analysed in this chapter. The p-values below the Bonferroni-corrected statistical significance level $0.05/4 = 0.013$ are in bold face. Sample sizes are indicated in the last row of the table.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is ConfidenceMajority better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0028**
Is ConfidenceMajority better than cBCI?	0.9858	**0.0000**	1.0000	**0.0075**	1.0000	0.6963	0.1472	0.3979
Is cBCI better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0028**
Is cBCI better than ConfidenceMajority?	0.0146	1.0000	**0.0000**	0.9926	**0.0000**	0.3047	0.8568	0.6668
Sample size	45	120	210	252	210	120	45	10

Table 8.5: p-values returned by the one-tailed Wilcoxon signed-rank test when comparing the performance of groups of different sizes using only the stimuli from the *right* viewpoint adopting the three methods analysed in this chapter. The p-values below the Bonferroni-corrected statistical significance level $0.05/4 = 0.013$ are in bold face. Sample sizes are indicated in the last row of the table.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is ConfidenceMajority better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0045**
Is ConfidenceMajority better than cBCI?	0.2646	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.0967
Is cBCI better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0027**
Is cBCI better than ConfidenceMajority?	0.7394	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9298
Sample size	45	120	210	252	210	120	45	10

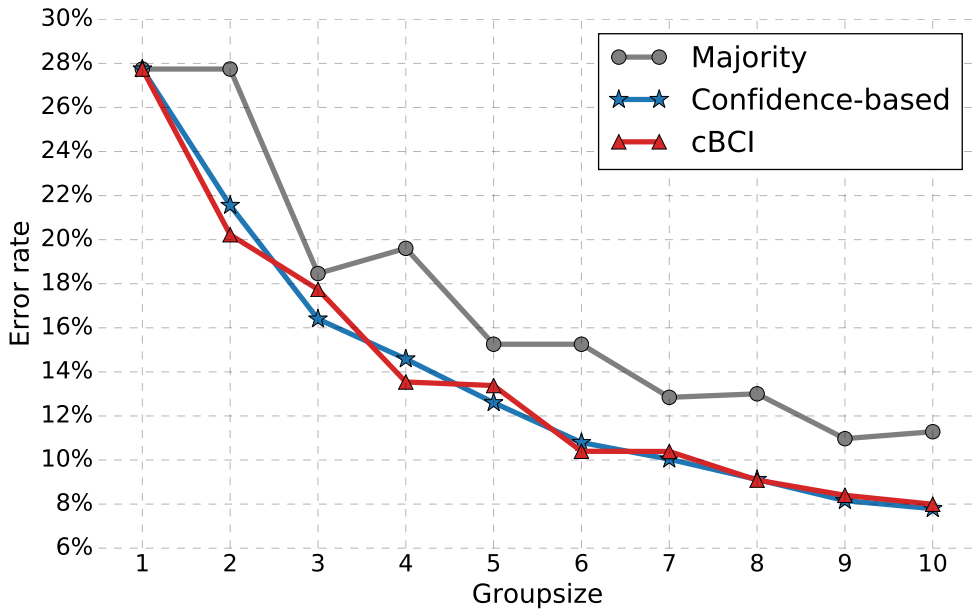

Figure 8.6: Error rates obtained by groups of different size when group's members were exposed by stimuli from different viewpoints. The performance obtained using three methods for making decisions are reported: the majority rule (grey), the confidence-based weighted-majority rule (blue), and our cBCI (red).

colleagues but taken from a different viewpoint. Similarly to the analysis described in Section 8.3.2, we compared the group performance obtained when using: (a) the majority rule (grey line), (b) a weighted-majority rule where individual decisions were weighted according to the confidence value reported by each participant (blue line), and (c) a weighted-majority rule where the weights were obtained from the confidence estimated by the cBCI from the neural signals and RTs (red line).

The results are quite surprising. First of all, we should note the big drop of error rates for traditional groups of size 3 using standard majority when compared to error rates of pairs. When participants were exposed to the same type of stimuli, this reduction was about 4.3%, while here it is almost 10%. This is

quite interesting as one might have expected that exposing group's members to unshared information would have a positive impact on group performance only when users are allowed to communicate and pool information [219, 179, 83, 115], while in this experiment no interaction was allowed between participants. Moreover, volunteers did not know one another and, yet, were able to achieve better group performance than in other studies [58].

Increasing the group size further reduced the error rates, except for groups of size 4 in which the majority rule was performing worse than with groups of size 3. This is due to the combination of two factors: (a) the group members are exposed to different sources of information and, therefore, their decisions will be more uncorrelated, making ties more frequent to happen; (b) the majority rule adopts a random decision in case of ties, which could only happen in even-sized groups. While for bigger even-sized groups these effects are obfuscated by the high number of groups simulated, in groups of size 4 they seem to provide a visible reduction in performance.

Interestingly, the group performance obtained by using the reported confidence to weigh individual responses (blue line in Figure 8.6) are superior than that obtained by traditional majority-based groups (grey line in Figure 8.6). This suggests that the confidence values provided by the participants now correlate much better with the correctness of their decision than before. The performance obtained by groups using these confidence estimates appears also to be much better than the average performance achieved by the cBCI for odd-sized groups. These results indicate that combining participants exposed to different information allows groups to correct individual errors of estimating the decision confidence, hence improving "metacognitive" and decision accuracies.

To further assess these differences, Table 8.6 shows the p-values of the Wilcoxon signed-rank test that has been used to compare the performance of the three methods over different group sizes. Both confidence-based methods are significantly better than simple majority for all group sizes, including groups of size 10 for which we now have more than 10 millions of samples and becomes therefore meaningful to use the Wilcoxon test. Moreover, cBCI-assisted group decisions are significantly better than confidence-based group ones for group sizes 2, 4 and 6, while they are statistically worse for all other group sizes.

Nevertheless, the multi-viewpoint approach allowed groups to reduce error rates down to less than 8%, while the best performance obtained when group members were exposed to the same information was just under 14% (see Figure 8.5(bottom)), which is still worse than what the simple majority rule achieves with the multi-viewpoint approach.

8.3.4 Group Decision Times

Figure 8.7 shows the average time required by groups of different sizes to make a decision when using the same-viewpoint (first four plots) and the multi-viewpoint (last plot) approaches. A group's response time is considered to be the maximum response time recorded across its members.

In Section 8.3.2 we have seen that groups using only the stimuli gathered from the right viewpoint are also the most accurate within the same-viewpoint approaches, as participants could spot the target more easily from this perspective. We have also seen that when people are confident, their RTs is generally lower than when they are not confident [100, 143]. Therefore, for these stimuli we

Table 8.6: p-values returned by the one-tailed Wilcoxon signed-rank test comparing the performance of groups of different sizes adopting the three methods analysed in this chapter when group's member were exposed to stimuli of the same scene taken from different viewpoints. The p-values below the Bonferroni-corrected statistical significance level $0.05/4 = 0.013$ are in bold face. Sample size for group size g is the number of permutations of the g elements picked from the 10 participants $\frac{10!}{(10-g)!}$.

	Group size								
Comparison	2	3	4	5	6	7	8	9	10
Is ConfidenceMajority better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Is ConfidenceMajority better than cBCI?	0.9995	**0.0000**	1.0000	**0.0000**	1.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Is cBCI better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Is cBCI better than ConfidenceMajority?	**0.0005**	1.0000	**0.0000**	1.0000	**0.0000**	1.0000	1.0000	1.0000	1.0000

were expecting RTs to be lower than for the other analysis. However, the results reported in Figure 8.7 show that the average RTs for individuals is very similar (around $1.5s$) in the five cases analysed, including the right viewpoint. This is likely to be due to the randomness with which participants had seen stimuli from the different viewpoints (see Section 8.2.2). The effects of tiredness and learning on RTs [217] could have merged with the effect of correctness in a decision, leading to similar average performance in all conditions. Furthermore, group decision times seem to increase much faster for the right viewpoint and much slower for the left one. This is likely to be due to the higher (lower) standard deviation of RTs for the right (left) viewpoint: bigger groups are more likely to include the slowest participants, which are the ones deciding the group RT.

A different scenario happens when considering the multiple viewpoints ap-

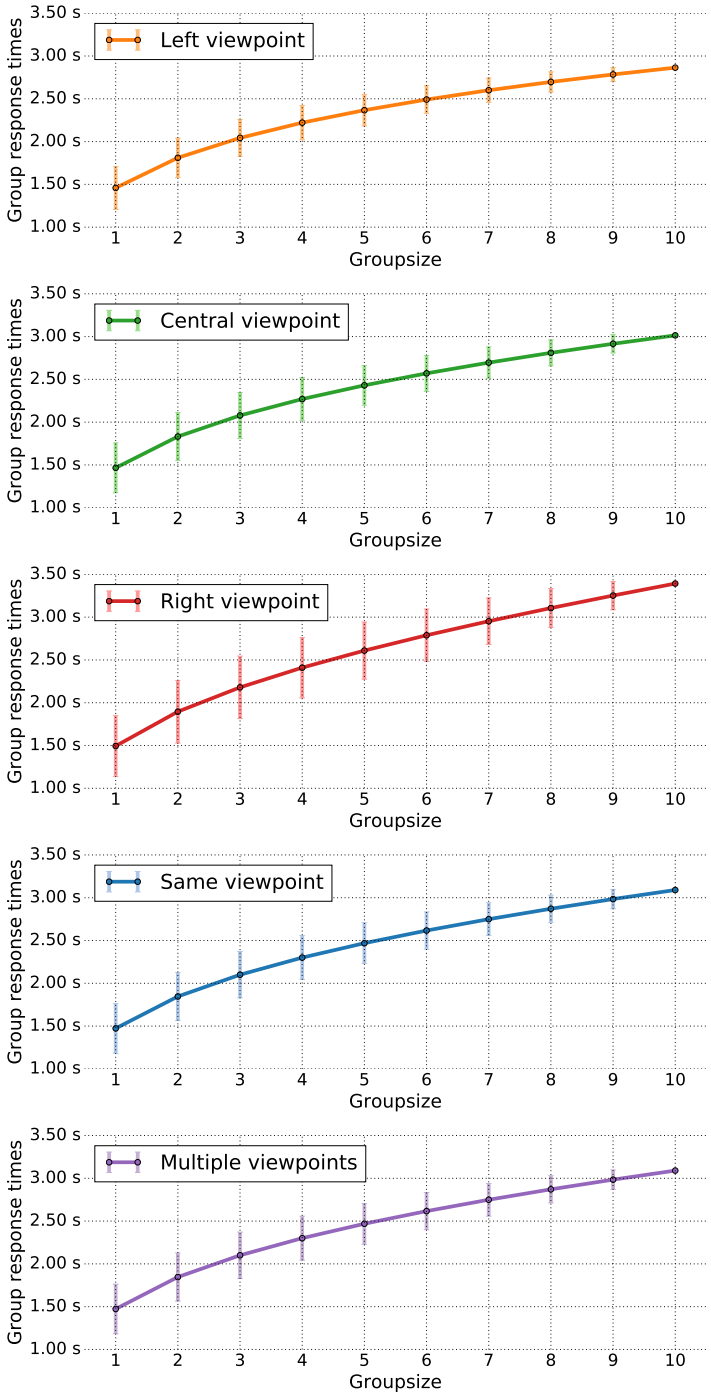

Figure 8.7: Average time required for groups of each size to make a decision when considering only the images from the left, central or right viewpoints (first three rows), when considering all images and having group members seeing the scene from the same viewpoint (fourth row), and when group members are seeing the scene from different viewpoints (last row). The error bars show the standard errors of each group size.

proach (purple plot in Figure 8.7). In this case, group decision times increase at a slower pace than with the right viewpoint, even though we have seen that groups perform much better with this multiple viewpoints approach than with others (see Figure 8.6). Providing participants with different sources of information seems to provide advantages both in terms of performance and speed.

We should note that the average group decision times are the same for groups using the same-viewpoint approach with all stimuli and those using the multi-viewpoint approach (blue and purple plots in Figure 8.7). This is due to the fact that both approaches use all the available stimuli and, therefore, while the multi-viewpoint approach builds many more groups than the same-viewpoint one, *on average* the group response times are the same.

Similarly to what we found in other chapters, in all approaches groups are much slower than the average individual in making a decision. This is because groups need to wait for all members to cast their votes, so that the group decision time is actually given by the RT of its slowest member. In Chapter 4 we have shown that this limitation could be overcome by allowing only the fastest respondents to influence the group's decision. To verify whether this strategy works also in the face recognition task used in this chapter, we applied it to groups seeing stimuli from the same viewpoint across all trials.

For each group size m, we have studied the performance and decision times obtained by groups of size $\hat{m}$ composed by the fastest $\hat{m}$ respondents on each trial, for all $\hat{m} = 1, \ldots, m$. The results obtained by traditional and cBCI-assisted groups are shown in Figure 8.8, where the line colour represents the group size m and the diameter of each circle represents the number of fastest respondents allowed to cast a vote $\hat{m}$.

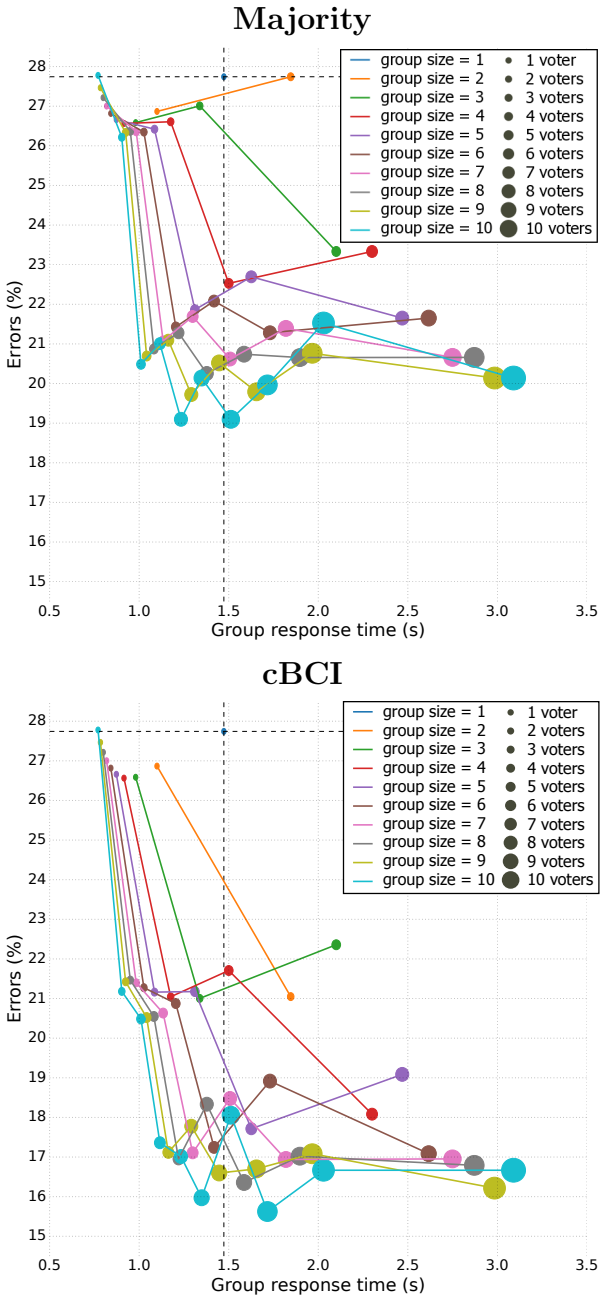

Figure 8.8: Comparison of the error rates and decision times obtained by traditional (top) and cBCI-assisted (bottom) groups of different sizes m when allowing only the fastest $\hat{m}$ voters to influence the group decision, for all $\hat{m} = 1, \ldots, m$. Each colour represents a group size m, while the diameter of the circle represents the number of fastest respondents $\hat{m}$ that contributed to the group decision. The horizontal and vertical dashed lines represent the average individual error rates and decision times, respectively. Therefore, the ideal situation is represented by the bottom-left quadrant where, on average, groups are faster and more accurate than individuals.

In each plot, the average error rate of individuals is represented by the horizontal dashed line, while their average response time is shown by the vertical dashed line. These lines split each plot in four quadrants. The top-right quadrant represents groups that are less accurate and slower than the average participant in making decisions. As expected, no groups fall in this quadrant, confirming that group decisions always provide an advantage in performance. The top-left and bottom-right quadrants represent groups that are faster or more accurate than the average individual in making decisions, respectively. The ideal condition is, finally, represented by the bottom-left quadrant, where groups are both faster and more accurate than the average individual.

These results confirm that it is possible to accelerate group decisions also in the face recognition task by allowing only the fastest respondents to contribute to the group decisions. For all group sizes m, there is at least one value $\hat{m} < m$ for which groups fall in the bottom-left quadrant and, therefore, have lower error rates and faster decision times than the average individual.

Moreover, in this experiment, the fastest respondent of each group size (i.e., smallest circles in each plot) is not always the most accurate, as opposed to what we found with the visual matching task (see Section 4.3.4). In both plots of Figure 8.8, we can see that the error rates of the fastest respondents decrease with the increase of the group sizes until groups of size 4. Then, for bigger group sizes, the average error rate of the fastest respondent increases with the expansion of the groups. This suggests that certain participants (which are more likely to be present in bigger groups) have an inverse relation between decision times and error rates.

Let us now focus on the average error rates. Indeed, for each group size m

the minimum decision time is achieved by the fastest individual (i.e., $\hat{m} = 1$). However, when considering error rates, we can see that the most accurate groups for most group sizes do not include all members in the decision-making process. For example, for $m = 10$, the most accurate group using the majority rule is the one including only the five fastest respondents, while for cBCI-assisted groups the best performance is achieved by considering the eight fastest respondents.

Furthermore, we can see that even adopting the strategy of considering only the fastest respondents in a group, cBCI-assisted sub-groups are almost always more accurate, on average, than equally-sized sub-groups using the majority rule. Also, while the majority rule requires three members in the sub-group to achieve the biggest improvement over smaller groups (i.e., see big drop in error rates in Figure 8.8(top)), only two fastest respondents are needed for cBCI-assisted sub-groups to significantly reduce the error rates.

8.3.5 Comparison of Confidence Estimates

The previous sections have shown that confidence-based group decisions are significantly better than traditional group decisions using the simple majority rule. We hypothesised that the reason behind this performance boost is a correlation between decision confidence and correctness (i.e., higher values of confidence are associated to higher probability of being correct [148]).

To verify this hypothesis, we compared the distributions of the two confidence estimates (i.e., reported by the user and cBCI) between trials in which the participants were correct and those where they were incorrect. The results of these comparisons are shown in Figures 8.9 and 8.10, respectively. We used Kruskal-

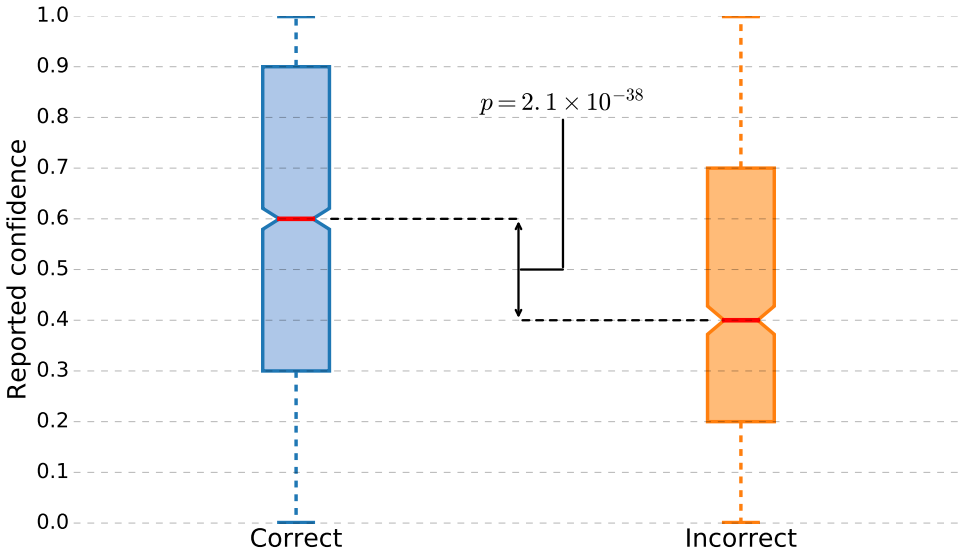

Figure 8.9: Distributions of the confidence values indicated by the participants after each response for the correct and incorrect decisions. The corresponding Kruskal-Wallis p-values comparing the correct and incorrect distributions are also reported.

Wallis test to assess whether or not the correct and incorrect distributions for reported and cBCI confidence estimates were significantly different.

Both confidence estimates have significantly different distributions for correct and incorrect trials, suggesting that they are good predictors of the correctness in a decision. Indeed, groups using confidence-based methods were superior, on average, to groups using standard majority – see Figure 8.4. However, in Section 8.3.2 we have seen that when considering only the stimuli from the left viewpoint, the performance of groups based on reported-confidence weighted-majority was similar to that of traditional groups using the majority rule. Although the distributions of reported confidence values for correct and incorrect trials using left stimuli were significantly different (Kruskal-Wallis $p = 1.3 \times 10^{-6}$), the cBCI seems to provide more robust predictors of correctness in all conditions.

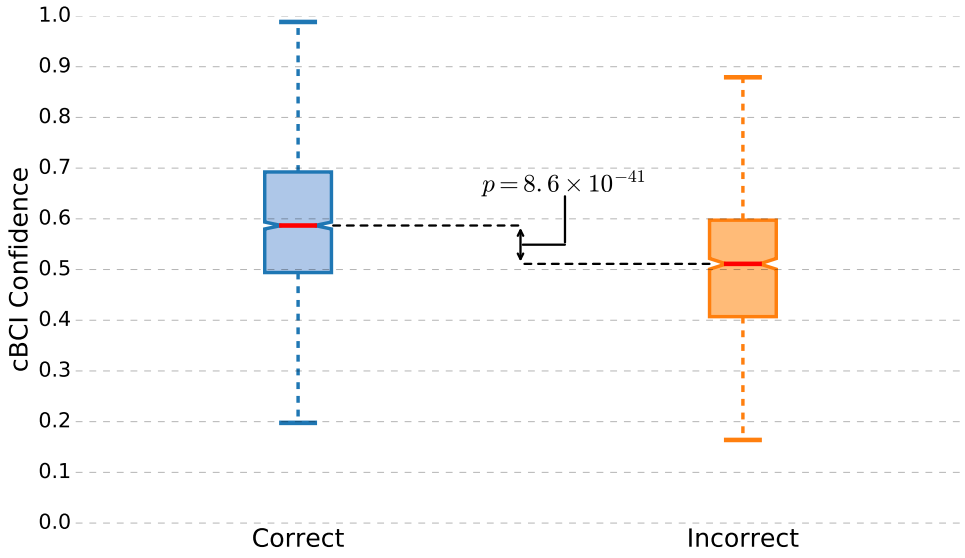

Figure 8.10: Distributions of the confidence weights estimated by the cBCI using neural signals and RTs for the correct and incorrect decisions. The corresponding Kruskal-Wallis p-values comparing the correct and incorrect distributions are also reported. The confidence weights have been divided by 34 for plotting purposes.

8.3.6 Neuro-Behavioural Correlates of Decision Confidence

Since the cBCI uses neural features and RTs to estimate the decision confidence, we expect to find significant differences in these features between correct and incorrect trials, which are, in turn, used by the machine learning algorithms to separate the two classes.

Figure 8.11 shows the distributions of response times for the correct and incorrect trials across all participants. The Kruskal-Wallis test has been used to verify that the two distributions are significantly different. Participants are generally slower in making decisions when they are less confident and, therefore, more likely to be incorrect. This confirms that, also for our face recognition experiment, RTs are good predictors of the correctness in the decision.

To study differences in the neural signals, we computed the grand averages of

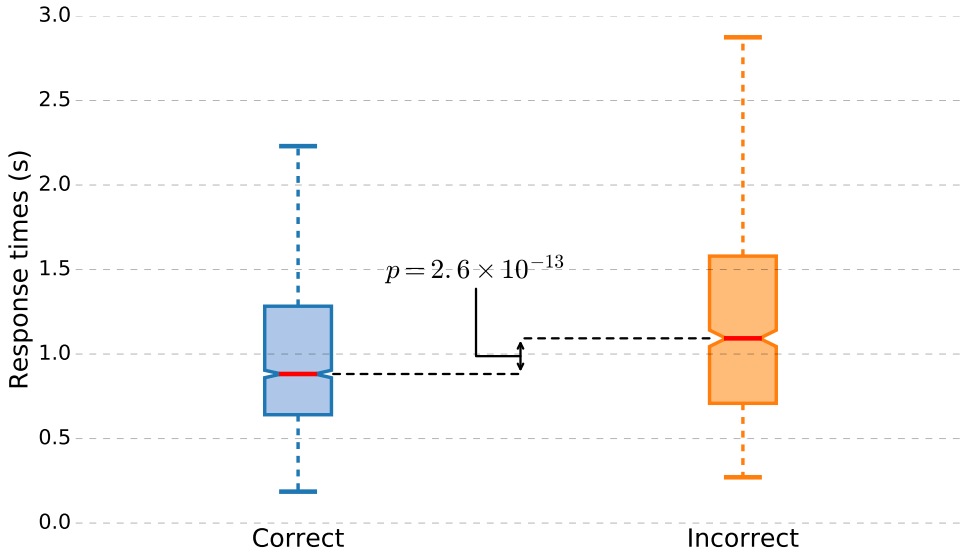

Figure 8.11: Distributions of response times across participants for the correct and incorrect trials. The corresponding Kruskal-Wallis p-values comparing the two distributions are also reported.

the stimulus- and response-locked epochs across the correct and incorrect trials (Figure 8.12). As done in Chapters 4 – 6, we have used the Kruskal-Wallis test to compare the voltages measured in each channel at each time step for the correct and incorrect trials, and the two-tailed Wilcoxon signed-rank test for paired samples to compare the mean ERPs obtained on an individual basis. The p-values of the statistical tests are also shown in Figure 8.12.

Figure 8.13 shows the scalp maps for the stimulus- and response-locked epochs for the difference between the grand averages of correct and incorrect trials (first row) and the corresponding Kruskal-Wallis p-values (last row) at representative time steps.

Let us first analyse the scalp maps. Figure 8.13 shows that, at the selected time steps, there are statistically significant differences at many electrode sites in both stimulus- and response-locked representations. The choice of including

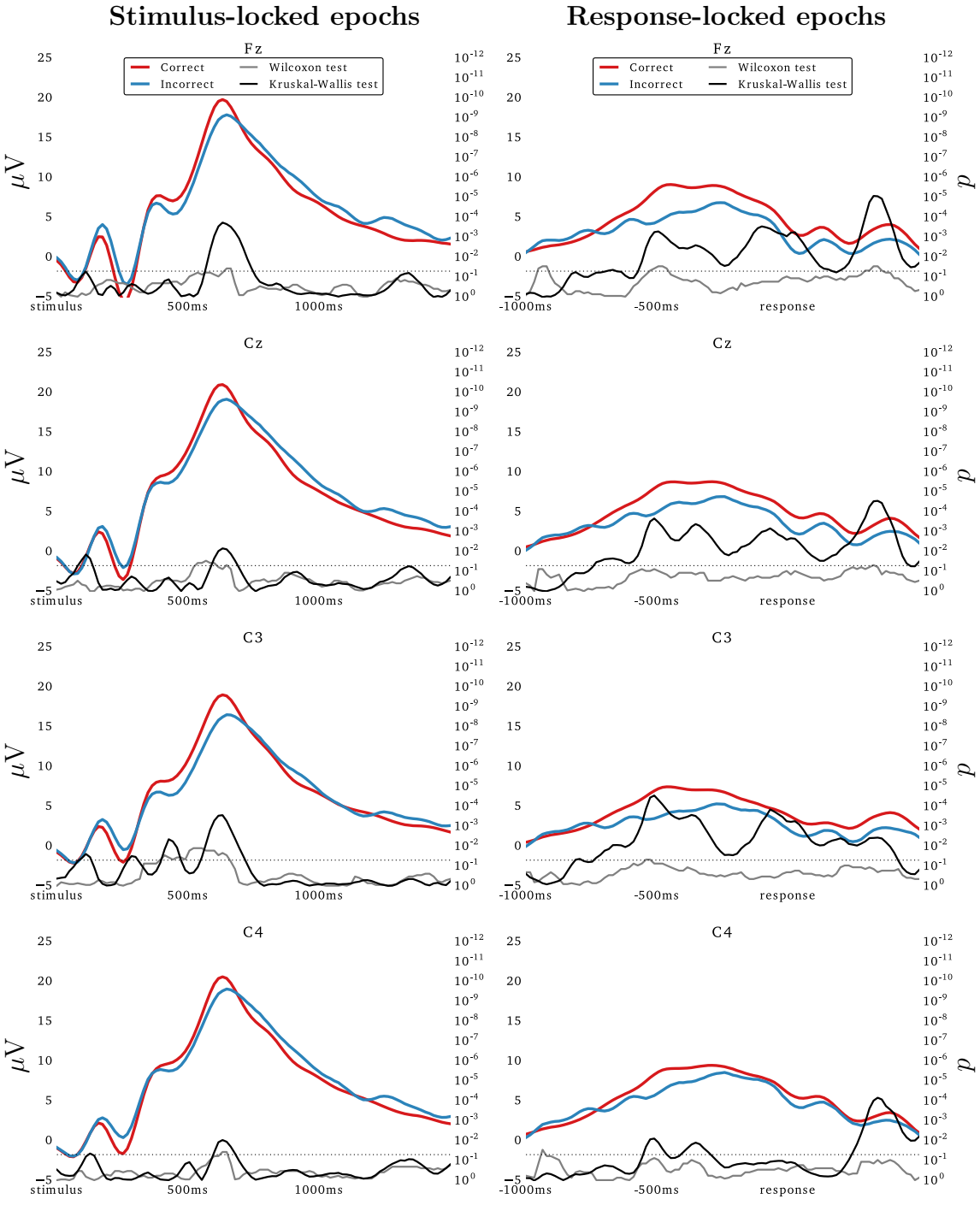

Figure 8.12: Grand averages of stimulus- (left) and response-locked (right) ERPs and corresponding temporal profile of the p-values of the Wilcoxon signed-rank test comparing participant-by-participant averages (grey) and of the Kruskal-Wallis test for all ERPs recorded, irrespective of participant (black), in each error class for representative channels. The horizontal dotted line represents the 5% significance level. The corresponding axes are oriented so that values above that line indicate statistical significance.

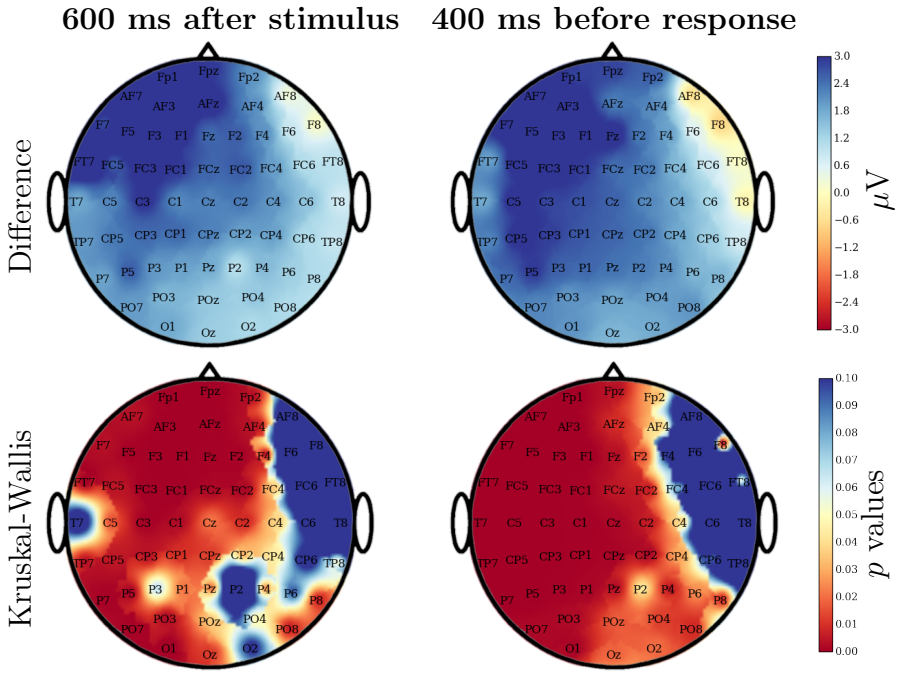

Figure 8.13: Scalp maps of the grand averages of the EEG activity recorded 600 ms after stimulus onset (first column) and 400 ms before the response (second column). Rows represent the difference in the activity between correct and incorrect trials (first row) and the p-values of the Kruskal-Wallis test used to compare the two sets ERPs (last row).

both types of epochs in the cBCI has proven to be beneficial (see similar results in Chapters 5 and 6), as they both provide useful information regarding the decision confidence and they also complement each other (e.g., the stimulus-locked epochs have most of the significant differences between correct and incorrect trials in the front-parietal and left-temporal lobes, while the response-locked epochs present significant differences mainly in the fronto-parietal and occipital lobes).

When looking at the results temporally (Figure 8.12), we can see that most of the differences in the stimulus-locked epochs between the "correct" and "incorrect" classes appear in the range 300–700 ms after the stimulus onset. This is the time range in which the peak of the P300 ERP is likely to occur [102]. Previous

research has shown that the P300 presents significant differences between target and non-target stimuli in face recognition [14]. Hence, we expect the P300 peak to be smaller in the trials where the user did not see the target (non-confident) and higher in the trials where the target was noticed by the observer (confident). Indeed, Figure 8.12 shows this behaviour. However, it should be noted that, in these plots, we are grouping the trials on the basis of the *correctness* in the decision. The "correct" set includes both trials where the target was present (P300 peak) and the user responded "yes" *and* the trials where the target was not present (no P300 peak) and the user responded "no". For these reasons, the differences in the grand averages between the two classes are smaller than expected, as the P300 is generally more associated to the presence of an unexpected event (e.g., the target face [14]) than to the decision confidence. However, these differences are still statistically significant, hence providing the cBCI the required information to estimate the decision confidence.

The P300 is not the only ERP providing useful information to estimate the decision confidence. The response-locked epochs used in our analysis included neural data recorded 1 s before the response *and* 500 ms after it, allowing the cBCI to also capture information related to post-decisional processes, such as evidence accumulation and confidence estimation [122]. Figure 8.12(right) shows that, about 250 ms after the response, the two distributions of "correct" and "incorrect" trials are significantly different. This adds further information about the decision confidence, which is likely to be used by the machine learning module of the cBCI to estimate the probability of being correct.

8.4 Conclusions

This chapter has studied the performance of cBCI-assisted groups in a face recognition task where isolated individuals made, on average, more than a quarter of decisions wrong. The group performance obtained by aggregating individual decisions according to the confidence estimated by the cBCI was compared with the performance achieved by traditional groups using the simple majority rule or a weighted-majority rule where confidence values reported by each participant after the decision were used as weights. We have showed that, when participants are exposed to the *same* stimuli (as they were in our previous tests of the cBCI with other experiments), their decision confidence estimates do not always correlate with the correctness in the decision, while the cBCI is able to provide more robust confidence estimates and significantly improve group performance. Moreover, the cBCI achieves the best performance with pairs, which are the groups that are more likely to be used in practice.

The cBCI predictions rely on two types of features: behavioural and neural. On the one hand, we have verified that response times correlate with the probability of being correct in a decision. On the other hand, we have shown that neural correlates of the decision confidence could be extracted from the EEG recordings by looking at the P300 and other ERPs from both a stimulus- and a response-locked representation. Moreover, post-decisional processing also provides information about the decision confidence.

We also tested the performance obtained by groups using the three decision methods described before when each group member was presented a picture of the same scene but taken from a *different* viewpoint. The exposure of participants to

different information allowed groups to be much more accurate than previously, even when using the simple majority. Moreover, the method using the confidence values reported by participants achieved the best performance for many group sizes. This suggests that, in the presence of unshared information, groups could use the confidence reported by each participant to make better decisions, although the confidence provided by the cBCI allows to further reduce the error rates in even-sized groups.

The confidence reported by participants after each decision should be used carefully. In Chapter 6 we found that group interaction makes these estimates totally unrelated from the correctness in the decision. Here, we have seen that, for the stimuli taken from the left viewpoint (which were also the more difficult ones for individuals), the performance of reported-confidence-based groups were similar to that of majority-based groups, while the cBCI was able to significantly augment group performance even in this condition.

When decision times are critical, we have also shown that group decision making could be accelerated and further improved in accuracy by allowing only the fastest respondents to contribute to the group decisions. Even with this strategy in place, cBCI-assisted groups are generally more accurate and faster than equally-sized groups using standard majority rule.

Face recognition is a task applied to several domains, including security and target detection. The advances of computer vision algorithms have allowed to make face recognition an automatic process for certain applications, although without reaching human-level performance. We believe that the proposed cBCI could be more accurate or, at least, have similar performance than computer-vision-based face-recognition systems, although it requires groups of people and,

therefore, it is not automatic.

Chapter 9

Augmenting Group Performance in Speech Perception

This chapter explores the possibility of applying the cBCI presented in Chapter 3 to groups performing a complex speech-perception task involving recognising target words in audio recordings affected by noise. Part of the research described in this chapter has been included in [195] and [194].

9.1 Introduction

Previous chapters have shown how a collaborative BCI could be used to improve group performance in a simple visual matching task (Chapter 4) and more challenging visual search (Chapters 5 and 6) and face recognition (Chapter 8) tasks. All of these tasks were based on decision tasks involving *visual perception* only, as visual responses are generally easy to detect over the scalp [126].

Certain decisions are taken on the basis of information gathered from senses

other than sight. For example, a soldier might need to establish from a sound whether or not there is a potential threat in the environment [49]. A few studies have used neural signals to improve the detection of target auditory stimuli, such as gunfire events [170, 171], or to spatially localise the source of the sound [119, 165]. Other studies have used auditory stimuli with a modified version of the oddball paradigm to make binary [64] or multi-choice [169] decisions and allow locked-in people to communicate. This suggests that the information used by our cBCI could also be available with auditory tasks. In particular, the P3a ERP seems to also be elicited by auditory stimuli [23]. However, the stimuli used in those studies did not include speech sounds, which also require the user to interpret and understand the meaning of what he/she heard.

One of the main functions of the human auditory system is speech perception, namely mapping sounds to internal linguistic representations [53]. In a broad range of contexts, such as defence and communications, speech perception is a very important task and succeeding in it is sometimes vital. For example, not interpreting correctly the location of the enemy communicated via radio could cause injuries to soldiers. Brain activity could be used with BCIs to augment and improve human performance in this challenging task. For example, in [69] users were listening to digits spoken in Chinese and the BCI was able to recognise from their brain signals the "target" ones. Sellers and Donchin [169] also used a BCI to discriminate between "target" and "non-target" auditory stimuli represented by single words.

In this chapter, we investigate whether or not the cBCI described in Chapter 3 could be successfully applied to a *complex speech-perception task* where participants listened to spoken sentences affected by noise and had to decide whether or

not certain target words are uttered. Individual decisions were aggregated using either the majority rule or a weighted majority based on confidence estimated by (a) the participants after each decision or (b) a cBCI using neural signals and RTs. Group decisions made by these three methods were then compared. We describe the main issues faced in the transition from visual to auditory tasks, including the modifications done to the original cBCI to adapt to auditory stimuli.

9.2 Methodology

9.2.1 Participants

Ten healthy volunteers (average age 24.9 ± 4.9, 2 females) with normal hearing and normal or corrected-to-normal vision participated in the experiment. All participants were native English speakers.

9.2.2 Stimuli and Task

Participants underwent a sequence of 8 blocks of 40 trials each, for a total of 320 trials. The sequence of displays presented in each trial is shown in Figure 9.1. After the usual fixation cross (see Section 3.2), an audio recording was played. Then, participants were asked to decide whether or not one of the following target words was uttered: “route”, “check”, “grid”, “lookout”, “side”, “trucks”, “village”. Decisions were accepted even if made by the participants before the end of the audio recording. After the response, similarly to the visual search experiments described in Chapter 6, participants were asked to report their degree of confidence in that decision, ranging from 0 to 100%, using the mouse wheel.

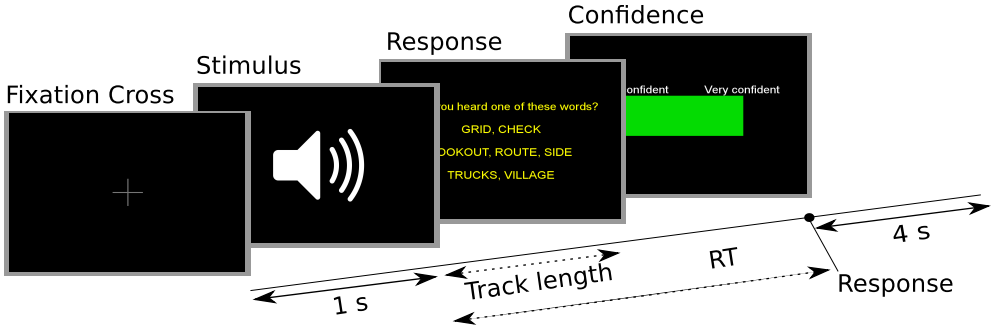

Figure 9.1: Sequence of stimuli presented in a trial of the speech perception task.

An horizontal bar indicated the selected confidence during this 4-s period.

The audio recordings used as stimuli consisted of 41 sentences containing one target word and 42 sentences without any target word. Between 4 and 20 words (average length 9.3 ± 2.8 words) were uttered in each audio recording, which were recorded from a member of the army (male, native-English speaker). The duration of the audio recordings was between 2.19 and 8.75 s (average duration 4.3 ± 1.4 s).

Two sets of stimuli were created from these audio recordings: "standard" and "high-noise". Each set included 415 stimuli, obtained as follows. For each audio recording, we created five versions by superimposing multiple types of noise on the original audio files, in order to make the task of identifying the target words more difficult. Noise types included white noise, environmental noise, volume changes, speed change, change of sampling rate, and audio drop-outs, all of which are typical of real-world military communications. Table 9.1 reports the parameters used for each type of noise in each set of stimuli. The difference between the standard and high-noise sets is that the stimuli in the latter were generally more affected by noise than the former ones. Noise was added using the Pydub library (`www.pydub.com`).

Noise		Set of stimuli	
Applied to track	**Parameter**	**Standard**	**High-noise**
Original	Volume reduction (dB)	`rand`$(6, 12)$	`rand`$(6, 16)$
Original	Speed-up (factor)	`rand`$(1, 1.5)$	`rand`$(1, 1.7)$
White noise	Volume reduction (dB)	`rand`$(5, 23)$	`rand`$(2, 15)$
White noise	Speed-up (factor)	`rand`$(1, 1.7)$	`rand`$(1, 1.7)$
Environmental noise	Volume reduction (dB)	`rand`$(5, 23)$	`rand`$(2, 15)$
Original and noise	Volume reduction (dB)	`rand`$(10, 25)$	`rand`$(12, 26)$
Original and noise	Duration (ms)	`rand`$(0, 700)$	`rand`$(0, 700)$
Output track	Sampling rate (kHz)	`rand`$(9, 17)$	`rand`$(7, 17)$

Table 9.1: Parameters used to add various types of noise to the original audio recording for the two sets of stimuli used in the experiment. The function `rand`(a, b) represents a random float value picked from the range $[a, b)$.

Before the main experiment, participants were asked to memorise the set of target words via a memorisation experiment – see Figure 9.2. In each trial, they were presented a display containing one word randomly chosen from a set of 39 words including the 7 target words, and were asked to indicate whether or not it was a target word by pressing the left or the right mouse buttons, respectively. The memorisation experiment ended as soon as the participant provided a correct answer to 80 questions in a row. If the volunteer made an incorrect response, an "error display" reminding him/her of the set of target words was shown and the memorisation experiment started again.

After completing the memorisation experiment, each participant was familiarised with the speech-perception task by doing 2 training blocks of 10 trials each of the main experiment. During familiarisation, participants had the chance to adjust the volume. Only stimuli from the "standard" set were used in this stage.

Sentences containing one of the target words were used in 50% of the trials.

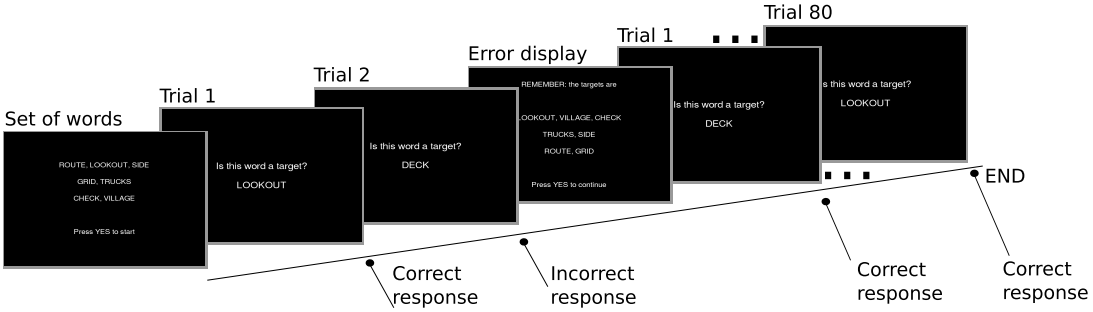

Figure 9.2: Protocol of the memorisation experiment used to help participants memorise the set of target words.

The same sequence of sentences was used in the experiment for all participants to be able to simulate offline concurrent group decisions (see Section 3.2). However, in order to reduce individual performance variations in the task, the difficulty of the audio tracks was dynamically varied by adjusting the proportion of sentences from the "standard" set vs the "high-noise" one. In the first block of trials, stimuli were chosen from the "standard" set for all participant. In the following blocks, a percentage p_s of audio recordings was chosen from the "high-noise" set so as to keep the accuracy of all participants not too far from 80%. More specifically, p_s was varied according to the following formula:

$$p_s = \texttt{min}\{1.0, \texttt{max}\{0.0, p_{s-1} + \texttt{sign}(acc_{s-1} - 0.8) \cdot \texttt{min}\{|acc_{s-1} - 0.8|, 0.2\}\}\} \quad (9.1)$$

where p_{s-1} and acc_{s-1} are the percentage of "high-noise" stimuli and the percentage of correct decisions made by the participant in the previous block, respectively ($p_1 = 0$ and $acc_{s-1} = 0.8$ for the first block). The aim of this formula is to make the speech-perception task gradually more difficult for participants that performed above the target 80% accuracy level in a block of trials by increasing the frequency of "high-noise" stimuli in the following block.

The volunteers were comfortably seated at about 80 cm from a LCD screen and were wearing in-ear earphones. All participants successfully completed the memorisation experiment in less than five minutes. Preparation and task familiarisation took approximately 40 minutes, while the actual experiment took about 35 minutes.

9.2.3 Making Group Decisions

Data were acquired and preprocessed as explained in Chapter 3. We set $p_b = 6$ Hz, $s_b = 8$ Hz and the final sampling rate $s_r = 16$ Hz.

The stimuli used in the speech perception task had different duration and the target word could be uttered at any time within the audio recording. This feature, very typical of realistic speech perception tasks, has two consequences. Firstly, it makes the stimulus-locked epochs used in previous experiments of this thesis (i.e., Chapters 5, 6 and 8) not appropriate to capture the ERPs associated to target detection and decision making (e.g., P300 and N200 [102]). The fact that target words could be uttered in any position of the audio recording makes the detection of such ERPs very difficult from stimulus-locked epochs. Moreover, these ERPs could even be produced after the end of the epoch, which, in other experiments, we considered lasting 1.5 s from the stimulus onset – see Chapter 3. One may suggest to increase the length of such epochs. However, this approach could be a double-edge sword as it will end up including neural data not related to the decision-making task in case of short audio recordings, hence increasing the noise included in the classification problem. Therefore, for simplicity we decided to only extract response-locked epochs starting 1 s before the user's response and

lasting 1.5 s from each trial. Secondly, since RTs are measured from the onset of the stimulus, they will not represent only the reaction time of the user but also the length of the audio recording. This is likely to reduce their correlation with the decision confidence. To partially compensate for this, we subtracted from each RT the duration of the audio recording used in that trial and used the result as RT feature. Indeed this requires the BCI to wait until the end of the audio recording before being able to estimate the decision confidence, hence increasing group decision times, while participants could provide a response before the end of the stimulus. However, we believe this is a reasonable compromise to partially compensate the loss of confidence-related information in RTs due to the realism of the task and increase the accuracy of cBCI-assisted groups.

This new RT feature also includes additional information. If the participant provided the response *before* the end of the audio recording (resulting in a negative RT feature), it is reasonable to think that he/she was particularly sure of having heard a target word, while in non-target trials a participant is more likely to wait until the end of the sentence to give his/her response. This information could further help the machine learning element of the cBCI to predict the confidence of the user.

Considering that the voice recognition task performed by the participants involved word recognition and language comprehension, we expected that key information could be found in the neural signals recorded in the left temporal lobe [232, 231]. Hence, we only used EEG data recorded at locations C5, TP7, T7, FC5 and CP5 for extracting neural features to estimate the confidence in decisions. This reduction in the number of electrodes is likely to promote generalisation and it makes the cBCI much more practical for real applications.

Neural features were extracted from response-locked epochs using LTCCSP as described in Section 5.2.3. Therefore, the cBCI used LARS to estimate the decision confidence from a feature vector of two LTCCSP and one RT features.

Group decisions were then made as described in Section 3.8 by using the sign of the weighted sum of the decisions of its members, where the weights were either the confidence reported by the participants after each decision or the confidence weights computed by the cBCI.

9.3 Results

9.3.1 Individual Performance

The percentages of erroneous decisions made by individuals undertaking the speech perception task are shown in Figure 9.3. The individual performance confirmed the difficulty of the task for a single participant. Many of the errors were false negatives, showing the effectiveness of adding noise to the stimuli to make the task of recognising the target words more challenging.

We should note that the error rates of some participants deviated from the target performance of the algorithm described in Section 9.2 for tuning the difficulty of each block of trials. Despite the use of high-noise stimuli, those participants were still able to perform well in the task.

9.3.2 Group Performance

Figure 9.4 shows the mean error rates obtained by groups of increasing size making their decisions using either the Majority rule (black line) or a weighted majority

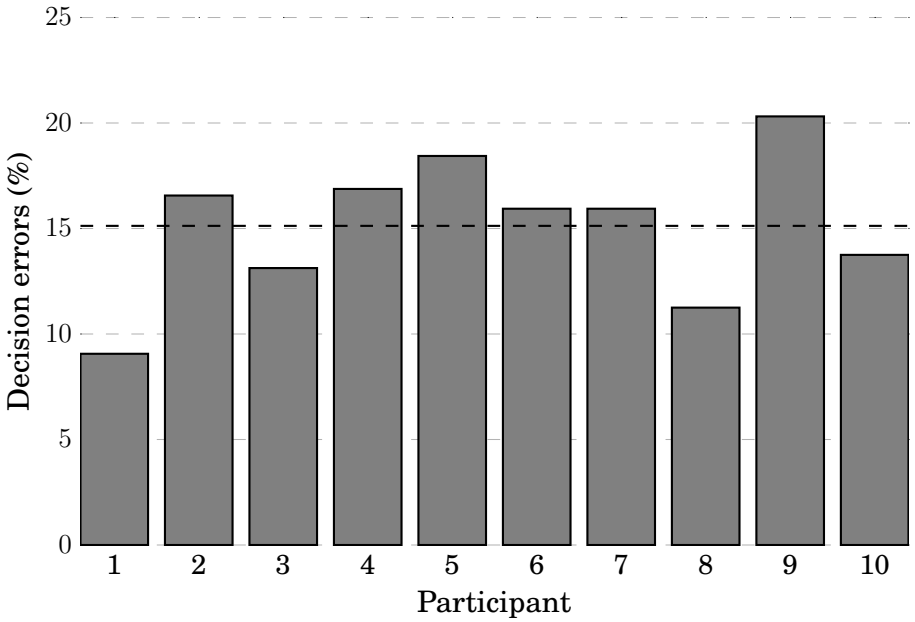

Figure 9.3: Mean decision errors (in %) achieved by participants in the speech perception experiment. The average error rate across the participants is shown by the dashed black line.

where individual decisions were weighed according to the confidence reported by the participants (ConfidenceMajority, blue line) or the confidence estimated by the cBCI (orange line). Table 9.2 shows the results of the statistical comparisons between the three methods made with the Wilcoxon signed-rank test.

Even with the very realistic experiment used in this chapter, groups of almost all sizes assisted by our cBCI were able to achieve significantly superior performance than traditional groups using majority (for groups of size 3, cBCI performance were nearly statistically significance). Similarly to the experiments described in other chapters of this thesis, the cBCI provides most of the advantages over majority for even-sized group, thanks to its tie-breaker ability.

Surprisingly, participants were extremely good in assessing their degree of confidence for this task. When using the reported confidence to weigh individual

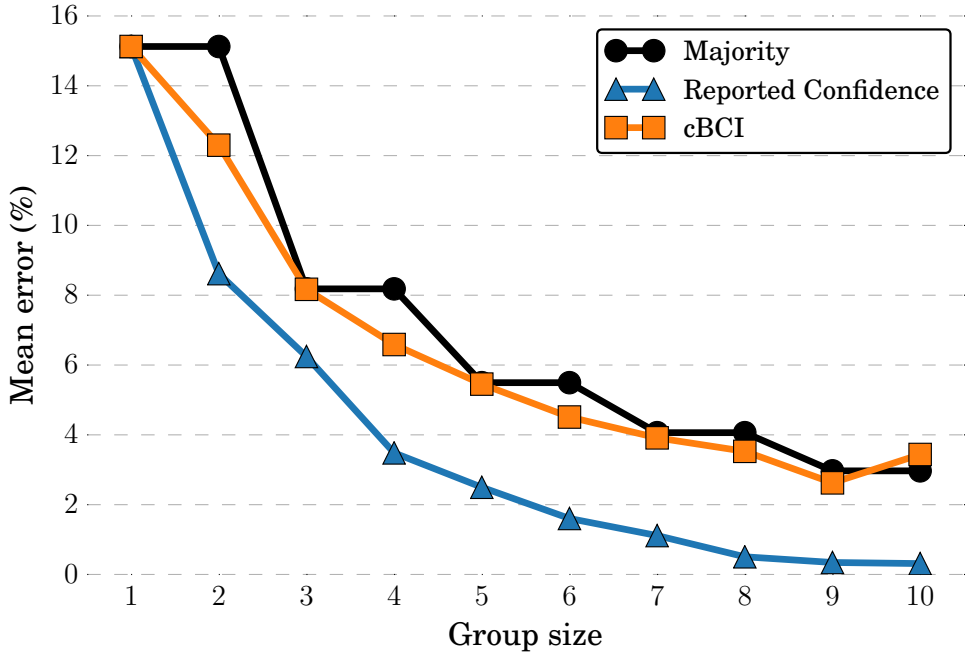

Figure 9.4: Mean decision errors (in %) of groups of different sizes when deciding using the majority rule (black) or a confidence-based weighted majority based on the reported confidence (blue) or the cBCI confidence (orange).

Table 9.2: One-tailed p-values returned by the Wilcoxon signed-rank test when comparing the performance of groups of increasing sizes adopting (a) the majority rule, (b) a weighted majority using the reported confidence (ConfidenceMajority), and (c) a weigthed majority based on the cBCI confidence. The number of groups of each size that could be assembled with 10 participants is indicated in the last row. p-values below the significance level 0.05 are in bold face.

	Group size							
Comparison	2	3	4	5	6	7	8	9
Is ConfidenceMajority better than Majority?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0028**
Is cBCI better than Majority?	**0.0000**	0.0760	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0058**
Is ConfidenceMajority better than cBCI?	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0026**
Sample size	45	120	210	252	210	120	45	10

decisions, groups were significantly more accurate than both those using majority *and* those assisted by our cBCI. Particularly encouraging was the performance of groups of 8+ members, which obtained perfect decisions in almost every trial (error rates $< 1\%$).

9.3.3 Comparisons of Confidence Estimates

Figure 9.5 compares the distributions of confidence values between trials where the participants made correct decisions and trials where they made incorrect ones, for both the reported (left) and the cBCI (right) confidence estimates.

As expected considering the results showed in the previous section, the confidence reported by the participants is well separated between the two sets, with median values for incorrect decisions being half (0.5) of those for correct ones (1.0). Interestingly, the median value for correct responses is 1.0, which is the ideal value to achieve an optimal metacognitive accuracy (i.e., when the participant made a correct decision, we want to give his/her response the maximum weight in the weighted majority rule used to obtain the group decision).

The distributions of the values in the "correct" and "incorrect" sets of trials are more similar for the cBCI confidence, although significantly different. These results are also confirmed by the plots of the density functions shown in Figure 9.5(bottom). The distributions of the cBCI confidence values overlap much more than those of the reported confidence.

The particularly good results obtained by groups with the reported confidence should be taken with caution. On the one hand, they represent a much harder yardstick for the cBCI than majority, and this can promote further research and

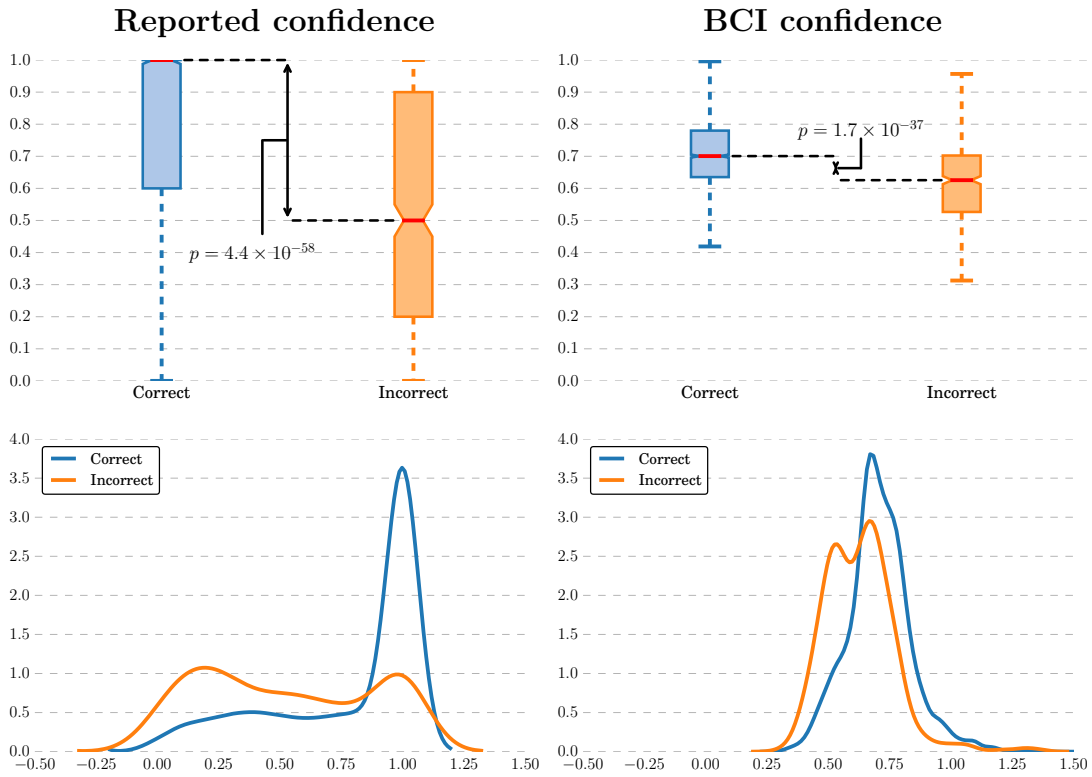

Figure 9.5: Box plots (top) representing the distributions of the confidence values reported by participants (left) and estimated by the cBCI (right) for correct and incorrect decisions and corresponding probability density functions (bottom) estimated via Gaussian kernel density estimation. The corresponding p-values of the Kruskal-Wallis test used to compare the "correct" and "incorrect" distributions are also shown. The cBCI confidence values have been divided by 36 for plotting purposes.

improve the exploitation of neural correlates of the decision confidence. Indeed, moving towards real-world decision-making applications comes at a cost: in this study, the cBCI could only count on response-locked epochs and RT features only partially correlated with correctness. On the other hand, there are circumstances in which these subjective confidence estimates might be totally unrelated to the correctness in a decision, especially when individuals are not very accurate [96, 132], in the presence of difficult stimuli (see Chapter 8), or when communication

between participants is allowed [99], as verified in Chapter 6.

To understand whether there is a risk in using the reported confidence in the auditory experiment, we looked at the distributions of confidence values for correct and incorrect trials *on a participant-by-participant basis*. Table 9.3 shows the results of this analysis. As can be seen, the confidence values reported by some participants were not correlating with the correctness in the decision. For example, participants 3 and 9 were overconfident, reporting high confidence values most of the times, even when they were incorrect. Conversely, participants 5 and 8 were underconfident, as they reported low confidence values even when they made the correct decisions. Groups using these confidence estimates to make decisions were very accurate because of their intrinsic ability of correcting errors (wisdom of crowds). However, the reported confidence is an unreliable predictor of correctness in a significant proportion of the participants.

9.3.4 ERP Analysis

Figure 9.6 shows the grand averages of the ERPs recorded in the "correct" and "incorrect" trials, as well as the temporal profiles of the p-values of the Wilcoxon signed-rank test comparing participant-by-participant averages and of the Kruskal-Wallis test comparing all ERPs recorded in each error class.

The plots clearly show that there are statistically significant differences in the neural signals between the two classes. These are mainly located in proximity of the response. However, other significant differences are also present at an earlier stage (e.g., around 500 ms before the response on channel TP7).

Our cBCI uses brain signals recorded from only five electrodes out of the 64

Table 9.3: Percentage of trials in which different participants (x axis) reported each value of the subjective confidence for the correct (top) and incorrect (bottom) trials. The last two rows of each table (grey) show the mean and the median confidence values of each participant in each set.

Correct

Confidence	1	2	3	4	5	6	7	8	9	10
1.0	63.9%	51.3%	88.8%	40.6%	34.5%	59.1%	68.4%	33.5%	87.8%	52.2%
0.9	1.0%	0.7%	0.7%	1.1%	3.8%	3.3%	8.9%	3.5%	4.3%	14.1%
0.8	3.4%	6.4%	1.1%	4.1%	4.2%	8.6%	4.8%	3.2%	3.5%	11.6%
0.7	4.8%	5.2%	1.1%	7.9%	3.8%	5.2%	4.1%	3.5%	2.0%	7.6%
0.6	4.5%	5.6%	1.1%	7.5%	7.3%	4.1%	1.9%	3.2%	1.2%	5.1%
0.5	5.5%	1.1%	1.1%	10.9%	4.6%	3.3%	1.5%	8.8%	0.8%	8.3%
0.4	2.4%	4.5%	0.7%	10.5%	7.7%	1.1%	1.1%	23.2%	0.0%	0.4%
0.3	4.8%	7.5%	0.7%	8.3%	8.0%	0.7%	1.9%	16.5%	0.4%	0.0%
0.2	4.5%	6.7%	1.1%	6.4%	13.4%	1.9%	2.2%	4.6%	0.0%	0.0%
0.1	1.7%	10.9%	2.5%	2.3%	12.3%	2.2%	4.5%	0.0%	0.0%	0.7%
0.0	3.4%	0.0%	1.1%	0.4%	0.4%	10.4%	0.7%	0.0%	0.0%	0.0%
MEAN	0.8	0.7	0.9	0.7	0.6	0.8	0.9	0.6	1.0	0.9
MEDIAN	1.0	1.0	1.0	0.7	0.6	1.0	1.0	0.5	1.0	1.0
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Participant

100% 80% 60% 40% 20% 0%

Incorrect

Confidence	1	2	3	4	5	6	7	8	9	10
1.0	10.3%	11.3%	69.0%	9.3%	0.0%	7.8%	11.8%	0.0%	80.0%	15.9%
0.9	0.0%	0.0%	0.0%	0.0%	0.0%	5.9%	5.9%	0.0%	1.5%	15.9%
0.8	0.0%	7.5%	7.1%	5.6%	3.4%	3.9%	13.7%	0.0%	4.6%	11.4%
0.7	10.3%	3.8%	2.4%	9.3%	1.7%	3.9%	9.8%	2.8%	3.1%	15.9%
0.6	3.4%	7.5%	2.4%	16.7%	5.1%	9.8%	3.9%	2.8%	3.1%	15.9%
0.5	17.2%	3.8%	0.0%	16.7%	5.1%	13.7%	7.8%	8.3%	3.1%	15.9%
0.4	3.4%	3.8%	2.4%	5.6%	6.8%	2.0%	7.8%	25.0%	0.0%	4.5%
0.3	20.7%	7.5%	0.0%	11.1%	13.6%	9.8%	7.8%	41.7%	3.1%	2.3%
0.2	10.3%	20.8%	7.1%	16.7%	23.7%	11.8%	9.8%	16.7%	0.0%	2.3%
0.1	10.3%	32.1%	2.4%	9.3%	35.6%	3.9%	19.6%	0.0%	0.0%	0.0%
0.0	13.8%	1.9%	7.1%	0.0%	5.1%	27.5%	2.0%	2.8%	1.5%	0.0%
MEAN	0.4	0.4	0.8	0.5	0.2	0.4	0.5	0.3	0.9	0.7
MEDIAN	0.3	0.2	1.0	0.5	0.2	0.3	0.5	0.3	1.0	0.7
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Participant

100% 80% 60% 40% 20% 0%

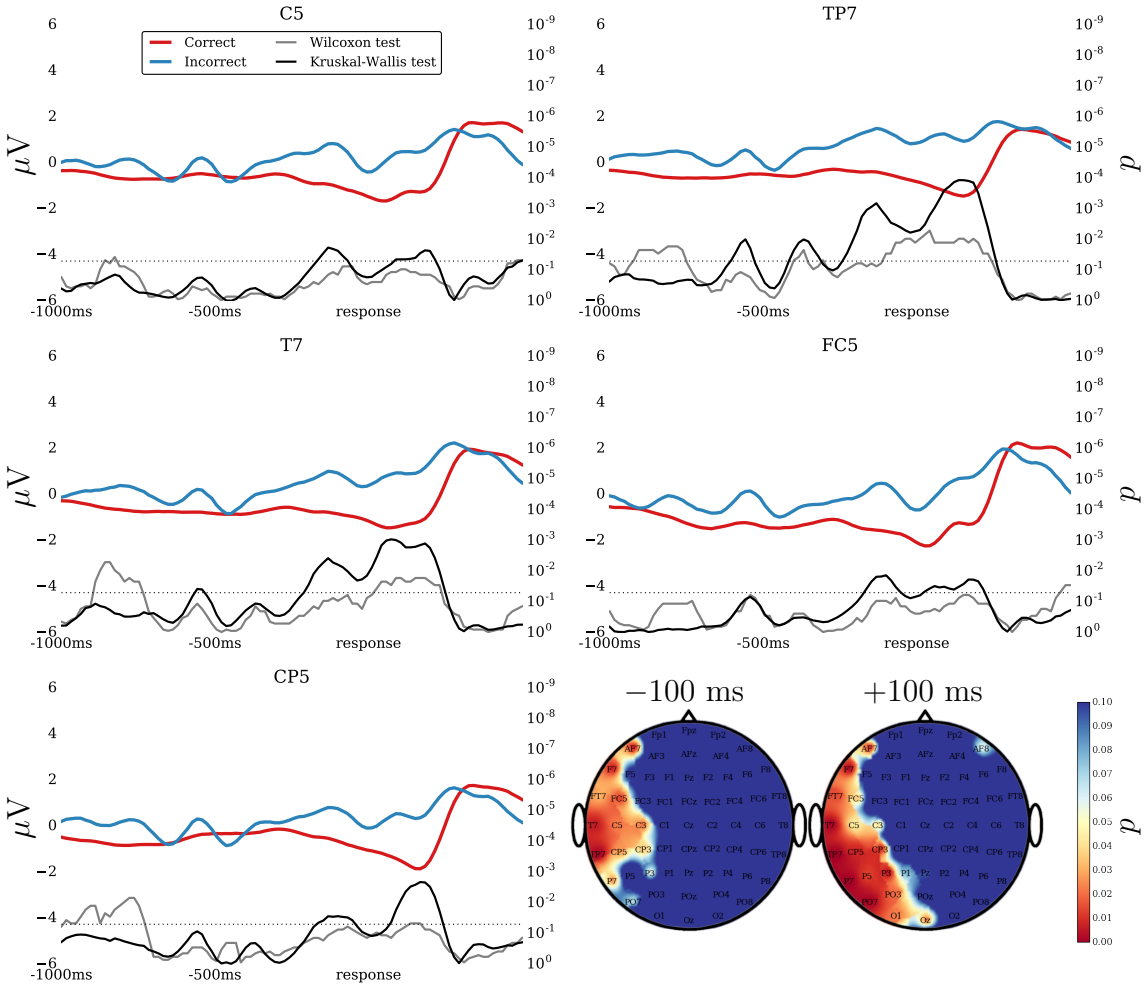

Figure 9.6: Grand averages of the response-locked epochs recorded at electrodes C5, TP7, T7, FC5 and CP5 and corresponding temporal profile of the p-values of the Wilcoxon signed-rank test comparing participant-by-participant averages and of the Kruskal-Wallis test comparing all ERPs recorded in each error class. The scalp maps (bottom right) show the p-values of the Kruskal-Wallis test used to compare the grand averages of the EEG activity recorded 100 ms before and after the user's response in each error class.

channels available to estimate the confidence. Figure 9.6(bottom right) shows the p-values of the Kruskal-Wallis test used to compare the grand averages of the EEG activity recorded 100 ms before and after the user's response in each error class for all electrodes available. These scalp maps justify our choice of electrodes: the neural information about the decision confidence is mainly concentrated in the left temporal lobe, where the primary auditory cortex is located. This part of the brain has several important functions related to speech perception, including low-level auditory processing and language comprehension [231, 232].

9.4 Conclusions

This chapter has analysed the possibility of using a cBCI to improve group decision making in a speech perception task. Participants were asked to listen to audio recordings of spoken sentences highly affected by noise and recognise target words.

The transition between the visual tasks used in Chapters 4, 5, 6 and 8 and the speech perception task used in this chapter has required some adaptations. Firstly, auditory tasks are perceptually and cognitively very different from the visual ones and generate different ERPs, such as the N100 and N200 [102]. Secondly, while the visual experiments conducted previously presented stimuli for a constant period of time, the audio recordings used here were of different duration and target words could be uttered in any position of the sentence. For these reasons, the cBCI could not use stimulus-locked EEG epochs to estimate the decision confidence. Response times were also affected by this less-constrained type of stimuli. Thirdly, to promote generalisation and practicality for future ap-

plications, we decided to only use the brain signals recorded from five electrode locations instead of using all 64 available channels as in our previous experiments.

With these changes, the cBCI was able to provide significantly better group decisions than standard majority for almost all group sizes. We have, therefore, verified that our cBCI improve group decision making also with auditory stimuli.

This chapter has also described the results obtained by groups making decisions using the confidence estimated by the participants themselves. Surprisingly, these confidence-based groups were significantly superior to both majority and cBCI groups. However, we showed how most of these very good results were due to the intrinsic error correction capabilities of groups, and not to the high metacognitive accuracy of the users. In fact, many participants were either underconfident or overconfident, hence reporting confidence values unrelated from the correctness in the decision. These results further corroborate our previous finding (see Chapter 6): the reported confidence is an unreliable predictor of correctness.

Chapter 10

Conclusions

This chapter summarises the main contributions of this thesis, looks at the degree to which it has addressed its research questions, and suggests possible future avenues of further studies in the area of collaborative BCIs for improving group decision making.

10.1 Main Contributions

Making correct decisions is, of course, very important in multiple contexts and has triggered research to study new techniques to assist humans in this delicate process. One of the most frequently used approaches to improve the quality of decisions is to act in groups, as these have augmented cognition capabilities due to the integration of the different perspectives of their members. However, in certain circumstances groups fail to provide advantages, especially when time constraints are present.

While BCIs are traditionally used as assistive technologies, in recent years

they have been applied to the area of human augmentation, especially with a collaborative approach. Researchers have started using brain signals from multiple users to improve the performance of single-user BCIs. Moreover, in decision making, collaborative BCIs have been able to outperform individuals in simple target-detection tasks [36].

This thesis has explored the possibility of using hybrid cBCIs to improve group decisions in a number of difficult tasks using either visual or auditory stimuli, including visual matching, traditional and realistic visual search, face recognition and speech perception. Instead of predicting the decision of the user from his/her brain signals, the hybrid cBCI firstly records the response of the participant via mouse clicks and, then, uses a combination of physiological and behavioural measures to estimate how *confident* the user is in making that decision. These confidence estimates correlate directly with decision correctness and can then be used to weigh individual responses and obtain better group decisions.

Seven experiments of increasing difficulty and realism have been conducted to test this approach. In all cases, cBCI-assisted groups were able to achieve superior performance than both individuals and equally-sized groups making decisions via the majority rule.

We have also studied the impact of a constrained form of communication on individual and group performance. Pairs of users were allowed to exchange information related to each other's opinion and degree of confidence before being able to choose whether or not changing their responses. We showed that this approach led individuals and groups to be much more erroneous than when participants were acting in isolation. However, the proposed cBCI was still able to significantly boost the group performance.

Furthermore, we showed how the confidence estimated by the cBCI was reliable across tasks and experiments, unlike the confidence reported by the participants after each decision, which yielded superior group decisions in the speech perception experiment but significantly deteriorated group performance in the visual search experiment with communicating pairs.

10.2 Progress towards Answering the Research Questions of this Thesis

At the beginning of this thesis (Section 1.3), several research questions were set. On the basis of the evidence gathered from the experimental work and analyses conducted, this section provides tentative answers to those questions.

Q1. Can group decision making based on neural, physiological and behavioural features achieve better levels of accuracy than traditional majority voting across a range of tasks?

In all decision tasks adopted in the various experiments described in this thesis (i.e., visual matching, visual search, face recognition and speech perception), groups assisted by the proposed hybrid cBCI based on neural features, RTs and, possibly, eye movements were able to achieve significantly better performance than equally-sized groups using majority voting to obtain group decisions.

While these results were obtained by the cBCI offline, they were also confirmed in the presence of a constrained form of communication within pairs of users performing the task concurrently. This suggests that the proposed cBCI achieves

better group decisions than traditional majority voting across a range of tasks *and* settings.

Q2. What is the best set of physiological and behavioural features acting as confidence indicators?

We firstly verified a finding from decades ago [100]: RTs are very informative in relation to estimating decision confidence and, therefore, they were always included in our feature set. Secondly, we found that neural signals always provided additional information to the cBCI. In particular, in Chapter 4 we showed that a cBCI based on both RTs and neural features achieved better performance than a cBCI based only on one of these two features. Furthermore, we found that, in decision-making tasks using visual stimuli, the vertical component of the eye movements could be used to extract eye features that correlate with the decision confidence.

These analyses have identified a set of physiological and behavioural correlates of decision confidence that works across tasks and settings.

Q3. What are the neural features that are the most relevant for the proposed hybrid cBCI for group decision making?

We have experimented with a number of techniques for extracting neural features correlating with the decision confidence, including stimulus- and response-locked ERP analysis, PCA and LTCCSP. The results obtained in this thesis suggest that the response-locked epochs are generally more informative about the decision confidence than the stimulus-locked ones, but the combined use of neural data from both epochs leads to the best group decisions.

The results obtained in visual matching and visual search also suggest that neural features extracted using LTCCSP are more informative than those obtained using the classic PCA transform. This is also reasonable as LTCCSP is a supervised method for feature extraction. However, it is very encouraging that 2 LTCCSP features seem to be more informative than 24 PCA ones, as this would also help the cBCI to scale up.

Q4. Is the confidence estimate provided by the cBCI more reliable than a confidence reported by the user?

We asked participants to report their degree of confidence in four experiments out of the seven conducted, namely in realistic visual search with and without group interaction (Chapter 6), face recognition (Chapter 8) and speech perception (Chapter 9). In all experiments the confidence estimated by the cBCI was able to provide advantages to groups over the majority rule. On the contrary, the confidence reported by the participants had very variable performance. In the auditory experiment, the reported confidence was far superior than the cBCI one (see Chapter 9). In visual search with non-communicating volunteers (Chapter 6), groups using these confidence estimates did achieve better performance over traditional majority groups, but were worse than cBCI-assisted groups, especially for even group sizes. Similar results were obtained in face recognition (Chapter 8). Finally, in visual search with communicating participants (Chapter 6), the reported confidence was totally uncorrelated with the correctness of the decision, making groups using these estimates even less accurate than majority groups.

These results confirm previous findings in the literature regarding the high

variability of metacognitive accuracy [122], depending on the participants themselves [123] or the context in which decisions are made [132]. With reported confidence, one needs to check whether it correlates with the correctness in a decision on a task-by-task basis, and comparing group performance using these estimates with that of majority-based and cBCI-assisted groups. If the reported confidence correlates with the correctness, then one could use these estimates and achieve higher group performance with less complexity (e.g., no needs of wearing EEG cap, eye tracker, etc.). However, in only one out of four experiments conducted this was the case. This suggests that when the aim is to provide better decisions in a variety of tasks and conditions, the confidence estimated by the cBCI should be preferred for its ability to provide significant advantages over traditional majority groups.

Q5. Can collaborative BCIs lead to faster decisions than average human reaction times?

Traditionally, groups are slower in making decisions than the average individual, as they require time for discussion and to collect the opinions of all members. Indeed, this happened in our experiments too, as the RT of the group was equal to the RT of its slowest member.

However, in Chapter 4 we have proposed a strategy that allow groups to become faster than the average individual with very minor loss in terms of group accuracy. We verified that a similar strategy worked also with groups undertaking a face recognition task (Chapter 8). Since RTs correlate with individual decisions being correct [100], we studied how the performance of groups of different size varies when allowing only the fastest members to contribute to the group decision.

This approach led to cBCI groups that were both faster and more accurate than the average individual and equally-sized majority groups, even when the latter were using the same strategy. Therefore, we can argue that the answer to this research question is in the positive.

Q6. Are there optimal scenarios for which BCI group decision making is most suited?

All decision tasks considered in this thesis share some common features, including uncertainty (e.g., due to the stimulus being shown for a very limited time or being affected by a high level of noise) and no time for discussion within the group. In all experiments, the cBCI was able to provide significant advantages over majority groups, especially for even-sized groups. However, we should note that the best improvement in performance was obtained in *visual search with non-communicating participants*, while the worst one was associated to auditory stimuli. These findings are reasonable as it is usually easier to extract neural information from the visual cortex [126], as a large part of the brain is devoted to visual processing and, therefore, the cBCI could rely on many EEG signals related to that activity.

Given that the cBCI provides a significant improvement already for very small groups, e.g., pairs, this is the most likely setup for practical applications of this technology. For instance, we can envisage a scenario where two users assisted by our cBCI look for threats in images gathered from a surveillance camera in all situations where the added security achieved through the cBCI is of primary importance (e.g., security control at the airport).

Q7. What is the impact of group interaction on cBCI performance?

Chapter 6 has shown how a constrained form of group interaction negatively affects individual and group performance. Since our cBCI only decides the weight to assign to behavioural responses (i.e., it cannot change individual decisions), indeed group communication also negatively affected the cBCI performance. We have also shown that the interaction between participants had a negative impact on the neural correlates of the decision confidence, i.e., on the ability of the cBCI to predict the likelihood of the user being correct in the decision. Despite these adverse conditions, the cBCI was still able to achieve significantly better group decisions than traditional groups.

Q8. In what ways does the exposure of different observers to various sources of information modify optimal group sizes, accuracy, and speed of decisions?

To start addressing this question, we formed groups with observers undertaking a face recognition task that were presented with images of the same scene from different viewpoints – see Chapter 8. We found that group accuracy was very much boosted by this approach when compared to the traditional strategy where all group members were seeing the same stimuli. Majority groups of size 9 were able to reduce the error rates from 20% to 11% when using the multi-viewpoint approach, while cBCI groups went from 16% down to 8%. These results are reasonable as the multi-viewpoint approach allows groups to integrate unique information provided by each member [181]. When concerning optimal group sizes and speed of decisions, no major effects were found by using this multi-viewpoint approach instead of the traditional one.

Moreover, in all experiments conducted in this thesis we found that group error rate decreases monotonically as the group size grows. When minimising

the accuracy is the only objective, one should therefore prefer bigger groups. However, when practicality and low decision times are also important, we envisage that pairs or groups of four people assisted by our cBCI are the optimal groups, as they are significantly more accurate than individuals without requiring long decision times.

10.3 Future Work

The research conducted in this thesis has proposed a hybrid cBCI framework to improve group decision making and tested it with several decision-making tasks. The positive results and issues faced during this work have opened up different pathways for future research.

10.3.1 Online Validation

All experiments conducted in this thesis were *offline*. Individual responses of the participants performing the various decision-making tasks were collected in different sessions and then aggregated, at a later stage, to simulate group decisions. However, BCI studies should always be validated online.

Future research should, therefore, be pointed at developing experiments where participants simultaneously make decisions while the cBCI estimates their decision confidence in real-time, so that the resulting group decisions could be presented to the volunteers immediately after their responses. The analyses of these online results should focus on how these settings affect metacognitive accuracy, individual and group performance in a range of tasks.

10.3.2 Full Communication between Participants

Chapter 6 has investigated the impact on performance of a *constrained* form of communication. We showed that both communicating individuals and groups were significantly less accurate than when the task was performed by isolated users. However, our computer-mediated communication did not allow participants to discuss and agree on a decision, but only consisted in sharing opinions and decision confidence followed by the possibility of changing the response. Moreover, interaction only occurred between pairs of users. Studying the impact of a more natural form of communication between participants in pairs and larger groups would be interesting to see if the results presented in Chapter 6 still hold.

10.3.3 Expand the Feature Set

The speech perception task analysed in Chapter 9 was the only experiment where the confidence reported by the participants together with the error correction capability of groups was providing significantly superior performance than the cBCI. This stimulates to conduct more research in order to improve the quality of the cBCI confidence estimates and leading to better group decisions.

One of the core components of the cBCI that could be improved is the feature set. Future research should (a) explore other methods for extracting neural features in the time, frequency and time-frequency domains (e.g., wavelet analysis), and (b) investigate other physiological measures related to decision making (e.g., skin conductance and pupil dilation [26]) that could complement our feature set and lead to better confidence estimates. This additional research would make another step towards identifying the best feature sets for estimating the decision

confidence.

10.3.4 Developing Advanced State-Space Models for Cognitive State Estimation

This thesis has also explored the possibility of estimating the cognitive state of a decision maker from a series of observations using state-space models. The neuro-behavioural model presented in Chapter 7 could be applied to our cBCI to temporarily exclude the group members with a low cognitive state from contributing to the group decision, as they are more likely to make an incorrect choice. This could lead to significant improvement in group performance.

Future research should be focused on investigating this integration process. Moreover, the accuracy of such state-space model in predicting the cognitive state could be enhanced by using advanced methods for extracting neural features, including the promising Gaussian-process factor analysis [227], which takes into account both temporal and spatial information.

Furthermore, while these state-space models aim at estimating the cognitive state of *single* users, one could also assume that the group itself has a dynamic cognitive state. Therefore, another interesting avenue of research would be to use state-space models to *track the group cognitive state* based on the observations related to its members (e.g., correctness, RTs and EEG signals of all the members). The dynamics of the cognitive state process could then be used for *group selection*, i.e., to identify which participants are more effective together. This, in turn, could further improve group performance in decision making.

10.3.5 Broaden the Range of Tasks

This thesis has applied the proposed cBCI to various decision-making tasks involving visual and auditory stimuli. Moreover, it paves the way to a number of real-world applications of cBCIs, especially when reducing the decision errors is vital.

Future research should investigate the performance of the proposed cBCI with decision-making tasks using (a) different auditory stimuli, for example where the user has to listen to an audio recording, understand the command issued and, possibly, execute it, (b) video streams as stimuli, for example as a natural extension of our face recognition experiment, and (c) multisensory stimuli (e.g., video and audio), to study whether or not the combination of multiple modalities impacts on the cBCI estimates of decision confidence.

Moreover, it will be important to bring the cBCI out of the lab and apply it to a real scenario. Also, more complex decision-making tasks should be adopted, including those requiring reasoning and not providing only two possible options (i.e., "yes" and "no"). For example, the performance of the cBCI could be studied when applied to the financial market, where two brokers assisted by such a system have to decide whether or not a certain stock should be bought. The performance could be evaluated in terms of amount of money lost instead of just as a number of erroneous decisions made. This is likely to trigger more interest on applying cBCIs to critical decision making and increase the probability that, in the near future, we will be able to use this technology to reduce our misjudgements.

Bibliography

[1] Laurence Aitchison, Dan Bang, Bahador Bahrami, and Peter E. Latham. Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLOS Computational Biology*, 11(10):1–23, 2015.

[2] George A. Alvarez and Patrick Cavanagh. The Capacity of Visual Short-Term Memory is Set Both by Visual Information Load and by Number of Objects. *Psychological Science*, 15(2):106–111, 2004.

[3] Fabio Babiloni and Laura Astolfi. Social neuroscience and hyperscanning techniques: Past, present and future. *Neuroscience and Biobehavioral Reviews*, 44:76–93, 2014.

[4] Fabio Babiloni, Febo Cincotti, Maria Grazia Marciani, Serenella Salinari, Laura Astolfi, Andrea Tocci, Fabio Aloise, Fabrizio De Vico Fallani, Simona Bufalari, and Donatella Mattia. The estimation of cortical activity for brain-computer interface: Applications in a domotic context. *Computational Intelligence and Neuroscience*, 2007(91651):1–7, 2007.

[5] Bahador Bahrami, Karsten Olsen, Peter E. Latham, Andreas Roepstorff, Geraint Rees, and Chris D. Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

[6] Dean C Barnlund. A comparative study of individual, majority, and group judgment. *The Journal of Abnormal and Social Psychology*, 58(1):55–60, 1959.

[7] Simone Benedetto, Marco Pedrotti, Luca Minin, Thierry Baccino, Alessandra Re, and Roberto Montanari. Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(3):199–208, 2011.

[8] Hans Berger. Über das elektrenkephalogramm des menschen. *European Archives of Psychiatry and Clinical Neuroscience*, 87(1):527–570, 1929.

[9] Nima Bigdely-Shamlo, Andrey Vankov, Rey R. Ramirez, and Scott Makeig. Brain activity-based image classification from rapid serial visual presentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(5):432–441, 2008.

[10] Niels Birbaumer, N. Ghanayim, Thilo Hinterberger, I. Iversen, B. Kotchoubey, Andrea Kübler, J. Perelmouter, E. Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[11] Laurent Bonnet, Fabien Lotte, and Anatole Lécuyer. Two Brains, One Game: Design and Evaluation of a Multi-User BCI Video Game Based on Motor Imagery. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(2):185–198, 2013.

[12] Leonard Branson, Nathan L. Steele, and Chung-Hsein Sung. When two heads are worse than one: Impact of group style and information type

on performance evaluation. *Journal of Business and Behavioral Sciences*, 22(1):75–84, 2010.

[13] Jeffrey B. Brookings, Glenn F. Wilson, and Carolyne R. Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, 42(3):361–77, 1996.

[14] Bangyu Cai, Siyuan Xiao, Lei Jiang, Yiwen Wang, and Xiaoxiang Zheng. A rapid face recognition BCI system using single-trial ERP. In *6th Annual International IEEE EMBS Conference on Neural Engineering*, pages 89–92, 2013.

[15] Matthew S. Cain, Edward Vul, Kait Clark, and Stephen R. Mitroff. A bayesian optimal foraging model of human visual search. *Psychological Science*, 23(9):1047–54, 2012.

[16] Hubert Cecotti and Bertrand Rivet. Performance estimation of a cooperative brain-computer interface based on the detection of steady-state visual evoked potentials. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2059–2063, 2014.

[17] Hubert Cecotti and Bertrand Rivet. Subject Combination and Electrode Selection in Cooperative Brain-Computer Interface Based on Event Related Potentials. *Brain Sciences*, 4(2):335–55, 2014.

[18] John K. Chapin, Karen A. Moxon, Ronald S. Markowitz, and Miguel A. L. Nicolelis. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neuroscience*, 2(7):664–670, 1999.

[19] Long Chen, Jing Jin, Yu Zhang, Xingyu Wang, and Andrzej Cichocki. A survey of the dummy face and human face stimuli used in BCI paradigm. *Journal of Neuroscience Methods*, 239:18–27, 2015.

[20] Shyh-Yueh Cheng and Hong-Te Hsu. Mental Fatigue Measurement Using EEG. In Giancarlo Nota, editor, *Risk Management Trends*, pages 203–228. InTech, 2011.

[21] Caterina Cinel, Glyn W. Humphreys, and Riccardo Poli. Cross-Modal Illusory Conjunctions Between Vision and Touch. *Journal of Experimental Psychology: Human Perception and Performance*, 28(5):1243–1266, 2002.

[22] Luca Citi, Riccardo Poli, Caterina Cinel, and Francisco Sepulveda. P300-based BCI mouse with genetically-optimized analogue control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(1):51–61, 2008.

[23] Marco D. Comerchero and John Polich. P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology*, 110(1):24–30, 1999.

[24] Eveline A. Crone, Riek J. M. Somsen, Bert Van Beek, and Maurits W. Van Der Molen. Heart rate and skin conductance analysis of antecendents and consequences of decision making. *Psychophysiology*, 41(4):531–40, 2004.

[25] James H. Davis. Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, 80:97–125, 1973.

[26] Archy O. de Berker, Robb B. Rutledge, Christoph Mathys, Louise Marshall, Raymond J. Dolan, and Sven Bestmann. Computations of uncer-

tainty mediate acute stress responses in humans. *Nature Communications*, 7(10996):1–11, 2016.

[27] Vincent de Gardelle and Pascal Mamassian. Weighting Mean and Variability during Confidence Judgments. *PLOS ONE*, 10(3):e0120870, mar 2015.

[28] Fabrizio De Vico Fallani, Vincenzo Nicosia, Roberta Sinatra, Laura Astolfi, Febo Cincotti, Donatella Mattia, Christopher Wilke, Alex Doud, Vito Latora, Bin He, and Fabio Babiloni. Defecting or Not Defecting: How to Read Human Behavior during Cooperative Games by EEG Measurements. *PLOS ONE*, 5(12):e14187, dec 2010.

[29] Arthur 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):1–38, 1977.

[30] Joseph Dien, Kevin M. Spencer, and Emanuel Donchin. Localization of the event-related potential novelty response as defined by principal components analysis. *Cognitive Brain Research*, 17(3):637–650, 2003.

[31] Joseph DiVita, Richard Obermayer, William Nugent, and James M. Linville. Verification of the Change Blindness Phenomenon While Managing Critical Events on a Combat Information Display. *Human Factors*, 46(2):205–218, 2004.

[32] Emanuel Donchin and Yael Arbel. P300 Based Brain Computer Interfaces: A Progress Report. In *Proceedings of the 5th International Conference*

on Foundations of Augmented Cognition, pages 724–731. Springer Berlin Heidelberg, 2009.

[33] Emanuel Donchin and E. F. Heffley. Multivariate Analysis of Event-Related Potential Data: A Tutorial Review. In Dave Otto, editor, *Multidisciplinary Perspectives in Event-Related Brain Potential Research*, pages 555–572. U.S. Government Printing Office, Washington, D.C., 1978.

[34] Mathew Dyson, Francisco Sepulveda, and John Q. Gan. Localisation of cognitive tasks used in EEG-based BCIs. *Clinical Neurophysiology*, 121(9):1481–1493, 2010.

[35] Miguel P. Eckstein. Visual search: A retrospective. *Journal of Vision*, 11(5):1–36, 2011.

[36] Miguel P. Eckstein, Koel Das, Binh T. Pham, Matthew F. Peterson, and Craig K. Abbey. Neural decoding of collective wisdom with multi-brain computing. *NeuroImage*, 59(1):94–108, 2012.

[37] Bradley Efron and Trevor Hastie. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[38] Kfir Eliaz, Debraj Ray, and Ronny Razin. Choice shifts in groups: A decision-theoretic basis. *American Economic Review*, 96(4):1321–1332, 2006.

[39] Deniz Erdogmus, Andre Adami, Michael Pavel, Tian Lan, Santosh Mathan, Stephen Whitlow, and Michael Dorneich. Cognitive State Estimation Based

on EEG for Augmented Cognition. In *2nd International IEEE EMBS Conference on Neural Engineering*, pages 566–569. IEEE, 2005.

[40] Michael W. Eysenck and Mark T. Keane. *Cognitive psychology: A student's handbook*. Psychology Press, 6th edition, 2010.

[41] Georg E. Fabiani, Dennis J. McFarland, Jonathan R. Wolpaw, and Gert Pfurtscheller. Conversion of EEG activity into cursor movement by a brain-computer interface (BCI). *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 12(3):331–338, 2004.

[42] Lawrence A. Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

[43] Reza Fazel-Rezai, Brendan Z. Allison, Christoph Guger, Eric W. Sellers, Sonja C. Kleih, and Andrea Kübler. P300 brain computer interface: current challenges and emerging trends. *Frontiers in Neuroengineering*, 5:14, 2012.

[44] Bernd Figner and Ryan O. Murphy. Using skin conductance in judgment and decision making research. In Michael Schulte-Mecklenbeck, Anton Kuehberger, and Rob Ranyard, editors, *A handbook of process tracing methods for decision research*, chapter 7, pages 163–184. Psychology Press, New York, NY, USA, 2011.

[45] Tomas Folke, Catrine Jacobsen, Stephen M. Fleming, and Benedetto De Martino. Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1:1–8, nov 2016.

[46] G. David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[47] Richard S. J. Frackowiak, Karl J. Friston, Christopher D. Frith, Raymond J. Dolan, Cathy J. Price, Semir Zeki, John T. Ashburner, and William D. Penny. *Human brain function*. Academic Press, 2nd edition, 2004.

[48] Ferran Galán, Marnix Nuttin, Eileen Lew, Pierre W. Ferrez, Gerolf Vanacker, Johan Philips, and José del R. Millán. A brain-actuated wheelchair: Asynchronous and non-invasive Braincomputer interfaces for continuous control of robots. *Clinical Neurophysiology*, 119(9):2159–2169, 2008.

[49] Jeremy R. Gaston and Tomasz R. Letowski. Listener perception of single-shot small arms fire. *Noise Control Engineering Journal*, 60(3):236–245, 2012.

[50] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[51] Zoubin Ghahramani. An Introduction to Hidden Markov Models and Bayesian Networks, 2001.

[52] Avniel Singh Ghuman, Nicolas M. Brunet, Yuanning Li, Roma O. Konecky, John A. Pyles, Shawn A. Walls, Vincent Destefino, Wei Wang, and R. Mark Richardson. Dynamic Encoding of Face Information in the Human Fusiform Gyrus. *Nature Communications*, 5(5672):1–22, 2014.

[53] Anne-Lise Giraud and David Poeppel. Speech Perception from a Neurophysiological Perspective. In David Poeppel, editor, *The Human Auditory Cortex*, pages 225–260. Springer Hand, 2012.

[54] Massimo Girelli and Steven J. Luck. Are the Same Attentional Mechanisms Used to Detect Visual Search Targets Defined by Color, Orientation, and Motion? *Journal of Cognitive Neuroscience*, 9(2):238–53, 1997.

[55] Daniel Göhring, David Latotzky, Miao Wang, and Raúl Rojas. Semi-autonomous Car Control Using Brain Computer Interfaces. In *Proceedings of the 12th International Conference on Intelligent Autonomous Systems (IAS)*, pages 393–408. Springer Berlin Heidelberg, 2013.

[56] Joshua I. Gold and Michael N. Shadlen. The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1):535–574, jul 2007.

[57] Piercesare Grimaldi, Hakwan Lau, and Michele A. Basso. There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience and Biobehavioral Reviews*, 55(2015):88–97, aug 2015.

[58] Deborah H. Gruenfeld, Elizabeth A. Mannix, Katherine Y. Williams, and Margaret A. Neale. Group Composition and Decision Making: How Member Familiarity and Information Distribution Affect Process and Performance. *Organizational Behavior and Human Decision Processes*, 67(1):1–15, 1996.

[59] Arzu Güneysu and H. Levent Akin. An SSVEP based BCI to control a humanoid robot by using portable EEG device. In *35th Annual International Conference of the IEEE EMBS*, pages 6905–6908, 2013.

[60] Burcu Gürçay, Barbara A. Mellers, and Jonathan Baron. The Power of Social Influence on Estimation Accuracy. *Journal of Behavioral Decision Making*, 28:250–261, 2014.

[61] Galen F. Hagen, James R. Gatherwright, Brian A. Lopez, and John Polich. P3a from visual stimuli: Task difficulty effects. *International Journal of Psychophysiology*, 59(1):8–14, 2006.

[62] Sowon Hahn, Curt Carlson, Shawn Singer, and Scott D. Gronlund. Aging and visual search: Automatic and controlled attentional bias to threat faces. *Acta Psychologica*, 123(3):312–336, 2006.

[63] Matti Hämäläinen, Riitta Hari, Risto J. Ilmoniemi, Jukka Knuutila, and Olli V. Lounasmaa. Magnetoencephalography theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, 1993.

[64] N. Jeremy Hill, Thomas Navin Lal, Karin Bierig, Niels Birbaumer, and Bernhard Schölkopf. An Auditory Paradigm for Brain-Computer Interfaces. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pages 569–576, 2004.

[65] Steven A. Hillyard and Lourdes Anllo-Vento. Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences*, 95(3):781–787, 1998.

[66] Steven A. Hillyard and Marta Kutas. Electrophysiology of cognitive processing. *Annual Review of Psychology*, 34(1):33–61, 1983.

[67] Leigh R. Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y. Masse, John D. Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S. Cash, Patrick van der Smagt, and John P. Donoghue. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–7, 2012.

[68] Alex O. Holcombe. Seeing slow and seeing fast: two limits on perception. *Trends in Cognitive Sciences*, 13(5):216–221, 2009.

[69] Bo Hong, Bin Lou, Jing Guo, and Shangkai Gao. Adaptive Active Auditory Brain Computer Interface. In *31st Annual International Conference of the IEEE EMBS*, pages 4531–4534, 2009.

[70] Gordon F. Hughes. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968.

[71] Sara Ilstedt Hjelm and Carolina Browall. Brainball - using brain activity for cool competition. In *NordiCHI2000 Proceedings*, pages 177–188, 2000.

[72] Lei Jiang, Yun Wang, Bangyu Cai, Yiwen Wang, Weidong Chen, and Xiaoxiang Zheng. Rapid Face Recognition Based on Single-Trial Event-Related Potential Detection over Multiple Brains. In *7th Annual International IEEE EMBS Conference on Neural Engineering*, pages 106–109, 2015.

[73] Jionghua Jin and Jianjun Shi. State Space Modeling of Sheet Metal Assembly for Dimensional Control. *Journal of Manufacturing Science and Engineering*, 121(4):756–762, 1999.

[74] G. Juckel, S. Karch, W. Kawohl, V. Kirsch, L. Jäger, G. Leicht, J. Lutz, A. Stammel, O. Pogarell, M. Ertl, M. Reiser, U. Hegerl, H. J. Möller, and C. Mulert. Age effects on the P300 potential and the corresponding fMRI BOLD-signal. *NeuroImage*, 60(4):2027–34, 2012.

[75] Daniel Kahneman and Shane Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics of Intuitive Judgment: Extensions and Applications*. Cambridge University Press, 2002.

[76] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[77] Albert B. Kao and Iain D. Couzin. Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 281(1784):1–8, apr 2014.

[78] Christoph Kapeller, Rupert Ortner, Gunther Krausz, Markus Bruckner, Brendan Z. Allison, Christoph Guger, and Günter Edlinger. Toward multibrain communication: Collaborative spelling with a P300 BCI. In *International Conference on Augmented Cognition*, pages 47–54. Springer International Publishing, 2014.

[79] Kapil D. Katyal, Matthew S. Johannes, Spencer Kellis, Tyson Aflalo, Christian Klaes, Timothy G. McGee, Matthew P. Para, Ying Shi, Brian Lee,

Kelsie Pejsa, Charles Liu, Brock A. Wester, Francesco Tenore, James D. Beaty, Alan D. Ravitz, Richard A. Andersen, and Michael P. McLoughlin. A collaborative BCI approach to autonomous control of a prosthetic limb system. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 1479–1482, 2014.

[80] Tobias Kaufmann, E. M. Hammer, and Andrea Kübler. ERPs contributing to classification in the "P300" BCI. *5th International Brain-Computer Interface Conference*, pages 136–9, 2011.

[81] Julian Paul Keenan, Bruce McCutcheon, Stefanie Freund, Gordon G. Gallup, Glenn Sanders, and Alvaro Pascual-Leone. Left hand advantage in a self-face recognition task. *Neuropsychologia*, 37(12):1421–1425, 1999.

[82] Norbert L. Kerr, Robert J. Maccoun, and Geoffrey P. Kramer. Bias in judgment: Comparing individuals and groups. *Psychological Review*, 103(4):687–719, 1996.

[83] Norbert L. Kerr and R. Scott Tindale. Group Performance and Decision Making. *Annual Review of Psychology*, 55(1):623–655, 2004.

[84] Andrew J. King, Lawrence Cheng, Sandra D. Starke, and Julia P. Myatt. Is the true 'wisdom of the crowd' to copy successful individuals? *Biology Letters*, 8(2):197–200, 2012.

[85] Genshiro Kitagawa. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[86] Louis Korczowski, Marco Congedo, and Christian Jutten. Single-Trial Classification of Multi-User P300-Based Brain-Computer Interface Using Riemannian Geometry. In *37th Annual International Conference of IEEE Engineering in Medicine and Biology Society*, pages 1769–1772, 2015.

[87] Agustín Lage-Castellanos, Juan I. Nieto, Ileana Quiñones, and Eduardo Martínez-Montes. A zero-training algorithm for EEG single-trial classification applied to a face recognition ERP experiment. In *32nd Annual International Conference of the IEEE EMBS*, pages 4209–4212, 2010.

[88] Saroj K. L. Lal and Ashley Craig. A critical review of the psychophysiology of driver fatigue. *Biological Psychology*, 55(3):173–94, 2001.

[89] Patrick R. Laughlin, Bryan L. Bonner, and Andrew G. Miner. Groups perform better than the best individuals on Letters-to-Numbers problems. *Organizational Behavior and Human Decision Processes*, 88(2):605–620, 2002.

[90] Patrick R. Laughlin, Erin C. Hatch, Jonathan S. Silver, and Lee Boh. Groups Perform Better Than the Best Individuals on Letters-to-Numbers Problems: Effects of Group Size. *Journal of Personality and Social Psychology*, 90(4):644–651, 2006.

[91] Sylvain Le Groux, Jonatas Manzolli, and Paul F. M. J. Verschure. Disembodied and Collaborative Musical Interaction in the Multimodal Brain Orchestra. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 309–314, 2010.

[92] Michael B. Lewis and Hadyn D. Ellis. How We Detect a Face: A Survey of Psychological Evidence. *International Journal of Imaging Systems and Technology*, 13(1):3–7, 2003.

[93] Junhua Li, Ye Liu, Zhen Lu, and Liqing Zhang. A Competitive Brain Computer Interface: Multi-person Car Racing System. In *35th Annual International Conference of the IEEE EMBS*, pages 2200–2203, 2013.

[94] Yueqing Li and Chang S Nam. A Collaborative Brain-Computer Interface (BCI) for ALS Patients. In *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting*, pages 716–720, 2015.

[95] Yueqing Li and Chang S Nam. Collaborative Brain-Computer Interface for People with Motor Disabilities. *IEEE Computational Intelligence Magazine*, 11(3):56–66, 2016.

[96] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips. Calibration of Probabilities: The State of the Art. In *Decision making and change in human affairs.*, pages 275–324. Springer Netherlands, 1977.

[97] Peilun Ling and Aleksandra Vučković. Competitive and Collaborative Multiuser BCI. In *Proceedings of the 6th International Brain-Computer Interface Meeting*, page 228, Asilomar, CA, USA, 2016.

[98] Glenn Littlepage, William Robison, and Kelly Reddington. Effects of Task Experience and Group Experience on Group Performance, Member Ability, and Recognition of Expertise. *Organizational Behavior and Human Decision Processes*, 69(2):133–147, 1997.

[99] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, may 2011.

[100] Robert Duncan Luce. *Response Times: Their Role in Inferring Elementary Mental Organization*, volume 8. Oxford University Press, 1986.

[101] Steven J. Luck. Ten Simple Rules for Designing and Interpreting ERP Experiments. In T. C. Handy, editor, *Event-Related Potentials: A Methods Handbook*. 2004.

[102] Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, 2nd edition, 2014.

[103] Steven J. Luck and Steven A. Hillyard. Electrophysiological correlates of feature analysis during visual search. *Psychophysiology*, 31(3):291–308, 1994.

[104] Steven J. Luck and E. S. Kappenman. *The Oxford Handbook of Event-Related Potential Components*. Oxford Library of Psychology. Oxford University Press, USA, 2011.

[105] Steven J. Luck, Geoffrey F. Woodman, and Edward K. Vogel. Event-related potential studies of attention. *Trends in Cognitive Sciences*, 4(11):432–440, 2000.

[106] Arien Mack and Irvin Rock. *Inattentional Blindness*. MIT Press, Cambridge, MA, 1998.

[107] Jaakko A. Malmivuo and Veikko E. Suihko. Effect of Skull Resistivity on the Spatial Resolutions of EEG and MEG. *IEEE Transactions on Biomedical Engineering*, 51(7):1276–1280, 2004.

[108] René Marois and Jason Ivanoff. Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, 9(6):296–305, 2005.

[109] Ana Matran-Fernandez and Riccardo Poli. Collaborative brain-computer interfaces for target localisation in rapid serial visual presentation. In *6th Computer Science and Electronic Engineering Conference*, pages 127–132, 2014.

[110] Ana Matran-Fernandez and Riccardo Poli. Event-Related Potentials induced by Cuts in Feature Movies and their Exploitation for Understanding Cut Efficacy. In *7th International IEEE EMBS Neural Engineering Conference*, pages 22–24, 2015.

[111] Ana Matran-Fernandez and Riccardo Poli. BrainComputer Interfaces for Detection and Localization of Targets in Aerial Images. *IEEE Transactions on Biomedical Engineering*, 64(4):959–969, apr 2017.

[112] Ana Matran-Fernandez, Riccardo Poli, and Caterina Cinel. Collaborative Brain-Computer Interfaces for the Automatic Classification of Images. In *6th International IEEE/EMBS Conference on Neural Engineering*, pages 1096–1099, 2013.

[113] Gerald Matthews, Lauren E. Reinerman-Jones, Daniel J. Barber, and Julian Abich. The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent. *Human Factors*, 57(1):125–143, 2014.

[114] James G. May, Robert S. Kennedy, Mary C. Williams, William P. Dunlap, and Julie R. Brannan. Eye Movements as an Index of Mental Workload. *Acta Psychologica*, 75(1):75–89, 1986.

[115] Martha L. Maznevski. Understanding Our Differences: Performance in Decision-Making Groups with Diverse Members. *Human Relations*, 47(5):531–552, 1994.

[116] James H. McClellan, Thomas W. Parks, and Lawrence R. Rabiner. A Computer Program for Designing Optimum FIR Linear Phase Digital Filters. *IEEE Transactions on Audio and Electroacoustics*, 21(6):506–526, 1973.

[117] Florent Meyniel, Daniel Schlunegger, and Stanislas Dehaene. The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLOS Computational Biology*, 11(6):e1004305, jun 2015.

[118] Vojkan Mihajlovic, Bernard Grundlehner, Ruud Vullers, and Julien Penders. Wearable, wireless EEG solutions in daily life applications: What are we missing? *IEEE Journal of Biomedical and Health Informatics*, 19(1):6–21, 2015.

[119] Isao Nambu, Masashi Ebisawa, Masumi Kogure, Shohei Yano, Haruhide Hokari, and Yasuhiro Wada. Estimating the Intended Sound Direction of the User: Toward an Auditory Brain-Computer Interface Using Out-of-Head Sound Localization. *PLOS ONE*, 8(2):e57174, feb 2013.

[120] Noman Naseer and Keum-Shik Hong. fNIRS-based brain-computer interfaces: a review. *Frontiers in Human Neuroscience*, 9:3, 2015.

[121] Joaquin Navajas, Maryam Ahmadi, and Rodrigo Quian Quiroga. Uncovering the Mechanisms of Conscious Face Perception: A Single-Trial Study of the N170 Responses. *Journal of Neuroscience*, 33(4):1337–1343, 2013.

[122] Joaquin Navajas, Bahador Bahrami, and Peter E. Latham. Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences*, 11:55–60, oct 2016.

[123] Joaquin Navajas, Chandni Hindocha, Hebah Foda, Mehdi Keramati, and Peter E Latham. The idiosyncratic nature of confidence. *bioRxiv*, page 102269, 2017.

[124] Joaquin Navajas and Lisandro N. Kaunitz. Late EEG Responses Are Absent for Conscious But Task-Irrelevant Stimuli. *Journal of Neuroscience*, 36(1):4–6, 2016.

[125] Joaquin Navajas, Mariano Sigman, and Juan E. Kamienkowski. Dynamics of Visibility, Confidence, and Choice During Eye Movements. *Journal of Experimental Psychology: Human Perception and Performance*, 40(3):1213–27, 2014.

[126] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012.

[127] Sander Nieuwenhuis, Gary Aston-Jones, and Jonathan D. Cohen. Decision Making, the P3, and the Locus Coeruleus*Norepinephrine System. *Psychological Bulletin*, 131(4):510–532, 2005.

[128] Sander Nieuwenhuis, K. Richard Ridderinkhof, Jos Blom, Guido P. H. Band, and Albert Kok. Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, 38(5):752–760, 2001.

[129] Anton Nijholt and Mannes Poel. Multi-Brain BCI: Characteristics and Social Interactions. In *Proceedings of the 10th International Conference on Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*, pages 79–90, 2016.

[130] B. S. Oken, M. C. Salinsky, and S. M. Elsas. Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical Neurophysiology*, 117(9):1885–1901, 2006.

[131] Mayra L. Padilla, Richard A. Wood, Laura A. Hale, and Robert T. Knight. Lapses in a PrefrontalExtrastriate Preparatory Attention Network Predict Mistakes. *Journal of Cognitive Neuroscience*, 18(9):1–11, 2006.

[132] Paul W. Paese and Michael A. Feuer. Decisions, Actions, and the Appropriateness of Confidence in Knowledge. *Journal of Behavioral Decision Making*, 4(1):1–16, 1991.

[133] John Palmer. Attentional Limits on the Perception and Memory of Visual Information. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):332–350, 1990.

[134] Toby A. Patterson, Len Thomas, Chris Wilcox, Otso Ovaskainen, and Jason Matthiopoulos. State-space models of individual animal movement. *Trends in Ecology and Evolution*, 23(2):87–94, 2008.

[135] Michael Peters and Jason Ivanoff. Performance Asymmetries in Computer Mouse Control of Right-Handers, and Left-Handers with Left- and Right-Handed Mouse Experience. *Journal of Motor Behavior*, 31(1):86–94, 1999.

[136] Gert Pfurtscheller, Doris Flotzinger, and Joachim Kalcher. Brain-Computer Interface - a new communication device for handicapped persons. *Journal of Microcomputer Applications*, 16:293–299, 1993.

[137] Terence W. Picton. The P300 Wave of the Human Event-Related Potential. *Journal of Clinical Neurophysiology*, 9(4):456–479, 1992.

[138] Michael A. Pitts, Antígona Martínez, and Steven A. Hillyard. Visual Processing of Contour Patterns under Conditions of Inattentional Blindness. *Journal of Cognitive Neuroscience*, 24(2):287–303, 2012.

[139] Michael A. Pitts, Stephen Metzler, and Steven A. Hillyard. Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. *Frontiers in Psychology*, 5:1–16, 2014.

[140] Riccardo Poli, Caterina Cinel, Luca Citi, and Francisco Sepulveda. Reaction-time binning: A simple method for increasing the resolving power of ERP averages. *Psychophysiology*, 47(3):467–485, 2010.

[141] Riccardo Poli, Caterina Cinel, Ana Matran-Fernandez, Francisco Sepulveda, and Adrian Stoica. Towards Cooperative Brain-Computer Interfaces for Space Navigation. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, pages 149–160, New York, USA, mar 2013. ACM Press.

[142] Riccardo Poli, Caterina Cinel, Francisco Sepulveda, and Adrian Stoica. Improving Decision-making based on Visual Perception via a Collaborative Brain-Computer Interface. In *2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 1–8. IEEE, 2013.

[143] Riccardo Poli, Davide Valeriani, and Caterina Cinel. Collaborative Brain-Computer Interface for Aiding Decision-Making. *PLOS ONE*, 9(7):e102693, jul 2014.

[144] John Polich. Task difficulty, probability, and inter-stimulus interval as determinants of P300 from auditory stimuli. *Electroencephalography and Clinical Neurophysiology/ Evoked Potentials*, 68(4):311–320, 1987.

[145] John Polich. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007.

[146] John Polich. Neuropsychology of P300. In *The Oxford Handbook of Event-Related Potential Components*, pages 159–188. 2012.

[147] John Polich, Lawrence Howard, and Arnold Starr. Effects of Age on the P300 Component of the Event-Related Potential from Auditory Stimuli: Peak Definition, Variation, and Measurement. *Journal of Gerontology*, 40(6):721–6, nov 1985.

[148] Alexandre Pouget, Jan Drugowitsch, and Adam Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3):366–374, 2016.

[149] Michael J. Prerau, Anne C. Smith, Uri T. Eden, Y. Kubota, Marianna Yanike, Wendy A. Suzuki, Ann M. Graybiel, and Emery N. Brown. Characterizing learning by simultaneous analysis of continuous and binary measures of performance. *Journal of Neurophysiology*, 102(5):3060–72, 2009.

[150] Michael J. Prerau, Anne C. Smith, Uri T. Eden, Marianna Yanike, Wendy A. Suzuki, and Emery N. Brown. A mixed filter algorithm for cognitive state estimation from simultaneously recorded continuous and binary measures of performance. *Biological Cybernetics*, 99(1):1–14, 2008.

[151] William Prinzmetal, Christin McCool, and Samuel Park. Attention: Reaction Time and Accuracy Reveal Different Mechanisms. *Journal of Experimental Psychology: General*, 134(1):73–92, 2005.

[152] Walter S. Pritchard. Psychophysiology of P300. *Psychological Bulletin*, 89(3):506–40, 1981.

[153] P. M. Quilter, B. B. MacGillivray, and D. G. Wadbrook. The removal of eye movement artefact from EEG signals using correlation techniques. In *Random Signal Analysis, IEEE Conference Publication*, volume 159, pages 93–100, 1977.

[154] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[155] Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal Spatial Filtering of Single Trial EEG During Imagined Hand Movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–6, 2000.

[156] Miguel A Recarte and Luis M Nunes. Mental Workload While Driving: Effects on Visual Search, Discrimination, and Decision Making. *Journal of Experimental Psychology: Applied*, 9(2):119–137, 2003.

[157] Arbora Resulaj, Roozbeh Kiani, Daniel M. Wolpert, and Michael N. Shadlen. Changes of mind in decision-making. *Nature*, 461(7261):263–266, 2009.

[158] Markus Ringnér. What is principal component analysis? *Nature Biotechnology*, 26(3):303–304, 2008.

[159] Neethu Robinson, A. P. Vinod, Kai Keng Ang, Keng Peng Tee, and Cuntai T. Guan. EEG-Based Classification of Fast and Slow Hand Movements Using Wavelet-CSP Algorithm. *IEEE Transactions on Biomedical Engineering*, 60(8):2123–2132, aug 2013.

[160] John W. Rohrbaugh, Emanuel Donchin, and Charles W. Eriksen. Decision making and the P300 component of the cortical evoked response. *Perception & Psychophysics*, 15(2):368–374, 1974.

[161] Bruno Rossion and Corentin Jacques. Th N170: Understanding the Time Course of Face Perception in the Human Brain. In *The Oxford handbook of ERP components*, pages 115–142. Oxford University Press, 2011.

[162] Michael D. Rugg and Michael G. H. Coles. *Electrophysiology of Mind: Event-related Brain Potentials and Cognition.* Oxford University Press, 1996.

[163] Kazuhiko Sagara, Kunihiko Kido, and Kuniaki Ozawa. Portable single-channel NIRS-based BMI system for motor disabilities' communication tools. In *Proceeding of the 31st Annual International IEEE EMBS Conference*, pages 602–605, 2009.

[164] Mathew Salvaris, Caterina Cinel, Luca Citi, and Riccardo Poli. Novel Protocols for P300-Based BrainComputer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(1):8–17, 2012.

[165] Martijn Schreuder, Benjamin Blankertz, and Michael Tangermann. A New Auditory Multi-Class Brain-Computer Interface Paradigm: Spatial Hearing as an Informative Cue. *PLOS ONE*, 5(4):e9813, apr 2010.

[166] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[167] Rafael Schultze-Kraft, Kai Görgen, M. Wenzel, J.-D. Haynes, and Benjamin Blankertz. Cooperating Brains: Joint Control of a Dual-BCI. In *Proceedings of the Fifth International Brain-Computer Interface Meeting*, 2013.

[168] A. Selimbeyoglu, Y. Keskin-ergen, and T. Demiralp. Clinical Neurophysiology What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision. *Clinical Neurophysiology*, 123(6):1158–1167, 2012.

[169] Eric W. Sellers and Emanuel Donchin. A P300-based brain-computer interface: Initial tests by ALS patients. *Clinical Neurophysiology*, 117(3):538–548, 2006.

[170] Jason Sherwin and Jeremy Gaston. Soldiers and marksmen under fire: monitoring performance with neural correlates of small arms fire localization. *Frontiers in Human Neuroscience*, 7(67):1–14, 2013.

[171] Jason Samuel Sherwin and Jeremy Rodney Gaston. Experience Does Not Equal Expertise in Recognizing Infrequent Incoming Gunfire: Neural Markers for Experience and Task Expertise at Peak Behavioral Performance. *PLOS ONE*, 10(2):e0115629, feb 2015.

[172] Robert H. Shumway and David S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

[173] Anne C. Smith and Emery N. Brown. Estimating a State-Space Model from Point Process Observations. *Neural Computation*, 15(5):965–991, 2003.

[174] Anne C. Smith, Loren M. Frank, Sylvia Wirth, Marianna Yanike, Dan Hu, Yasuo Jubota, Ann M. Graybiel, Wendy A. Suzuki, and Emery N. Brown. Dynamic Analysis of Learning in Behavioral Experiments. *The Journal of Neuroscience*, 24(2):447–461, 2004.

[175] Elaine Snyder and Steven A. Hillyard. Long-Latency Evoked Potentials to Irrelevant, Deviant Stimuli. *Behavioral Biology*, 16(3):319–331, 1976.

[176] Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545, 2008.

[177] Robert D. Sorkin, Christopher J. Hays, and Ryan West. Signal-Detection Analysis of Group Decision Making. *Psychological Review*, 108(1):183, 2001.

[178] Nancy K. Squires, Kenneth C. Squires, and Steven A. Hillyard. Two Varieties of Long-Latency Waves Evoked by Unpredictable Auditory Stimuli in Man. *Electroencephalography and Clinical Neurophysiology*, 38(4):387–401, 1975.

[179] Garold Stasser and William Titus. Pooling of Unshared Information in Group Decision Making: Biased Information Sampling During Discussion. *Journal of Personality & Social Psychology*, 48(6):1467–1478, 1985.

[180] Adrian Stoica, Ana Matran-Fernandez, Dimitrios Andreou, Riccardo Poli, Caterina Cinel, Y Iwashita, and C Padgett. Multi-brain fusion and applications to intelligence analysis. In *Proceedings of the SPIE*, volume 8756, pages 1–8, 2013.

[181] James Surowiecki. *The Wisdom of Crowds*. Anchor, New York, 2005.

[182] Samuel Sutton, Margery Braren, Joseph Zubin, and E. R. John. Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, 150(3700):1187–1188, 1965.

[183] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[184] Yaniv Taigman, Ming Yang, Marc Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[185] Xiaoyang Tan, Songcan Chen, Zhi Hua Zhou, and Fuyan Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39(9):1725–1745, 2006.

[186] James W. Tanaka, Tim Curran, Albert L. Porterfield, and Daniel Collins. Activation of Preexisting and Acquired Face Representations: The N250 Event-related Potential as an Index of Face Familiarity. *Journal of Cognitive Neuroscience*, 18(9):1488–1497, 2006.

[187] Fumihiko Taya, Yu Sun, Fabio Babiloni, Nitish Thakor, and Anastasios Bezerianos. Brain enhancement through cognitive training: a new insight from brain connectome. *Frontiers in Systems Neuroscience*, 9(April):44, apr 2015.

[188] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[189] Manoj Thulasidas, Cuntai Guan, and Jiankang Wu. Robust Classification of EEG Signal for Brain-Computer Interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(1):24–29, 2006.

[190] R. Scott Tindale, Tatsuya Kameda, and Verlin B. Hinsz. Group Decision Making. In M. A. Hogg and J. Cooper, editors, *Sage handbook of social psychology*, pages 381–403. Sage Publications, Inc., London, 2003.

[191] George Townsend and Valerie Platsko. Pushing the P300-based braincomputer interface beyond 100 bpm: extending performance guided constraints into the temporal domain. *Journal of Neural Engineering*, 13(2):1–15, apr 2016.

[192] Athina Tzovara, Micah M. Murray, Nicolas Bourdaud, Ricardo Chavarriaga, José del R. Millán, and Marzia De Lucia. The timing of exploratory decision-making revealed by single-trial topographic EEG analyses. *NeuroImage*, 60(4):1959–1969, 2012.

[193] Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences of the United States of America*, 113(10):1–6, 2016.

[194] Davide Valeriani, Caterina Cinel, and Riccardo Poli. Hybrid Collaborative Brain-Computer Interfaces to Augment Group Decision Making. In *1st International Neuroergonomics Conference*, 2016.

[195] Davide Valeriani, Caterina Cinel, and Riccardo Poli. Improving Speech Perception with Collaborative Brain-Computer Interfaces. In *38th Annual International IEEE EMBS Conference*, 2016.

[196] Davide Valeriani, Caterina Cinel, and Riccardo Poli. Augmenting Group Performance in Target-Face Recognition via Collaborative Brain-Computer Interfaces for Surveillance Applications. In *8th International IEEE EMBS Neural Engineering Conference*, 2017.

[197] Davide Valeriani and Ana Matran-Fernandez. Past and Future of Multi-Mind Brain-Computer Interfaces. In Chang Soo Nam, Anton Nijholt, and

Fabien Lotte, editors, *Brain-Computer Interfaces Handbook: Technological and Theoretical Advances*, chapter 16. CRC Press, 2017.

[198] Davide Valeriani, Riccardo Poli, and Caterina Cinel. A Collaborative Brain-Computer Interface for Improving Group Detection of Visual Targets in Complex Natural Environments. In *7th International IEEE EMBS Neural Engineering Conference*, pages 25–28, 2015.

[199] Davide Valeriani, Riccardo Poli, and Caterina Cinel. Enhancement of Group Perception via a Collaborative Brain-Computer Interface. *IEEE Transactions on Biomedical Engineering*, 64(6):1238–1248, 2016.

[200] Bram van de Laar, Hayrettin Gürkök, Danny Plass-Oude Bos, Mannes Poel, and Anton Nijholt. Experiencing BCI control in a popular computer game. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(2):176–184, 2013.

[201] Berry van den Berg, Lawrence G. Appelbaum, Kait Clark, Monicque M. Lorist, and Marty G. Woldorff. Visual search performance is predicted by both prestimulus and poststimulus electrical brain activity. *Scientific Reports*, 6:37718, nov 2016.

[202] Jan B. F. van Erp, Fabien Lotte, and Michael Tangermann. Brain-Computer Interfaces: Beyond Medical Applications. *Computer*, 45(4):26–34, 2012.

[203] Jr. Vaughan and G. Herbert. The relationship of brain activity to scalp recordings of event-related potentials. In Emanuel Donchin and Donald B.

Lindsley, editors, *Average Evoked Potentials: Methods, Results and Evaluations.*, pages 45–75. U.S. Government Printing Office, Washington, D.C., xvii edition, 1969.

[204] Giovanni Vecchiato, Jlenia Toppi, Febo Cincotti, Laura Astolfi, Fabrizio De Vico Fallani, Fabio Aloise, Donatella Mattia, S. Bocale, F. Vernucci, and Fabio Babiloni. Neuropolitics: EEG spectral maps related to a political vote based on the first impression of the candidate's face. In *32nd Annual International Conference of the IEEE EMBS*, pages 2902–2905, 2010.

[205] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, pages 1–20, aug 2016.

[206] Douglas Vickers. Where does the balance of evidence lie with respect to confidence? In *Proceedings of the 17th Annual Meeting of the International Society for Psychophysics*, pages 148–153, 2001.

[207] Paul Viola and Michael J. Jones. Robust Real-Time Face Detection PAUL. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[208] Laura Walker Renninger, Preeti Verghese, and James Coughlan. Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7(3):1–17, 2007.

[209] Eric A. Wan and Rudolph van der Merwe. The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 153–158. IEEE, 2000.

[210] Haixian Wang and Wenming Zheng. Local Temporal Common Spatial Patterns for Robust Single-Trial EEG Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(2):131–139, 2009.

[211] Yijun Wang and Tzyy-Ping Jung. A collaborative framework for brain-computer interfaces. In *Society for Neuroscience San Diego*, San Diego, 2010. Society of Neuroscience.

[212] Yijun Wang and Tzyy-Ping Jung. A collaborative brain-computer interface for improving human performance. *PLOS ONE*, 6(5):e20422, 2011.

[213] Yijun Wang, Yu-Te Wang, Tzyy-Ping Jung, Xiaorong Gao, and Shangkai Gao. A Collaborative Brain-Computer Interface. In *4th International Conference on Biomedical Engineering and Informatics (BMEI)*, pages 583–586, 2011.

[214] Yiwen Wang, Lei Jiang, Yun Wang, Bangyu Cai, Yueming Wang, Weidong Chen, Sanyuan Zhang, and Xiaoxiang Zheng. An Iterative Approach for EEG-Based Rapid Face Search: A Refined Retrieval by Brain Computer Interfaces. *IEEE Transactions on Autonomous Mental Development*, 7(3):211–222, 2015.

[215] Sumio Watanabe. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[216] Peter D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified

periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

[217] Alan Traviss Welford. Choice reaction time: Basic concepts. In *Reaction Times*, chapter 3, pages 73–128. Academic Press, London, 1980.

[218] Andreas Widmann and Erich Schröger. Filter Effects and Filter Artifacts in the Analysis of Electrophysiological Data. *Frontiers in Psychology*, 3(July):1–5, 2012.

[219] Jennifer R. Winquist and James R. Larson. Information Pooling: When It Impacts Group Decision Making. *Journal of Personality and Social Psychology*, 74(2):371–377, 1998.

[220] Martijn E. Wokke, Axel Cleeremans, and K. Richard Ridderinkhof. Sure I'm Sure: Prefrontal Oscillations Support Metacognitive Monitoring of Decision Making. *The Journal of Neuroscience*, 37(4):781–789, 2017.

[221] Jeremy M. Wolfe. Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–38, jun 1994.

[222] Jeremy M Wolfe. Visual Search. In H. Pashler, editor, *Attention*, pages 1–41. University College London Press, London, UK, 1998.

[223] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

[224] Jonathan R. Wolpaw and Dennis J. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface

in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17849–54, dec 2004.

[225] Jonathan R. Wolpaw, Dennis J. McFarland, Gregory W. Neat, and Catherine A. Forneris. An EEG-based brain-computer interface for cursor control. *Electroencephalography and Clinical Neurophysiology*, 78(3):252–259, 1991.

[226] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, Brian C Lovell, P O Box, and St Lucia. Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 74–81, 2011.

[227] Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *Journal of Neurophysiology*, 102(1):614–635, 2009.

[228] Shuli Yu, Timothy J. Pleskac, and Matthew D. Zeigenfuse. Dynamics of Postdecisional Processing of Confidence Shuli. *Journal of Experimental Psychology: General*, 144(2):489–510, 2015.

[229] Peng Yuan, Yijun Wang, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. A Collaborative Brain-Computer Interface for Accelerating Human Decision Making. In Constantine Stephanidis and Margherita Antona, editors, *International Conference on Universal Access in Human-Computer Interaction*, pages 672–681. Springer Berlin Heidelberg, 2013.

[230] Peng Yuan, Yijun Wang, Wei Wu, Honglai Xu, Xiaorong Gao, and Shangkai Gao. Study on an Online Collaborative BCI to Accelerate Response to Visual Targets. In *34th Annual International Conference of the IEEE EMBS*, pages 1736–1739, 2012.

[231] Robert J. Zatorre. Neural Specializations for Tonal Processing. *Annals of the New York Academy of Sciences*, 930(1):193–210, 2012.

[232] Robert J. Zatorre and Pascal Belin. Spectral and Temporal Processing in Human Auditory Cortex. *Cerebral cortex*, 11(10):946–53, 2001.

[233] Rui Zhang, Peng Xu, Tiejun Liu, Yangsong Zhang, Lanjin Guo, Peiyang Li, and Dezhong Yao. Local Temporal Correlation Common Spatial Patterns for Single Trial EEG Classification during Motor Imagery. *Computational and Mathematical Methods in Medicine*, 2013:1–7, 2013.

[234] Yu Zhang, Qibin Zhao, Jin Jing, Xingyu Wang, and Andrzej Cichocki. A novel BCI based on ERP components sensitive to configural processing of human faces. *Journal of Neural Engineering*, 9(2):026018, 2012.

[235] Yu Zhang, Qibin Zhao, Guoxu Zhou, Xingyu Wang, and Andrzej Cichocki. Regularized CSP with Fisher's criterion to improve classification of single-trial ERPs for BCI. In *9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 891–895, 2012.

[236] Wenyi Zhao, Rama Chellappa, P. Jonathon Phillips, and Azriel Rosenfeld. Face Recognition: A Literature Survey. *ACM Computing Survey*, 35(4):399–458, 2003.

[237] Ariel Zylberberg, Pieter R. Roelfsema, and Mariano Sigman. Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition*, 27(1):246–253, 2014.