Causal Modelling of Evoked Brain Responses

by

Marta Isabel Figueiredo Garrido

Wellcome Trust Centre for Neuroimaging
Institute of Neurology

A thesis submitted for the degree of Doctor of Philosophy
University College London
April, 2008

Primary supervisor: Professor Karl J Friston

Secondary supervisor: Dr. James M Kilner

Declaration

I, Marta Isabel Figueiredo Garrido, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

"What we observe is not nature itself, but nature exposed to our method of questioning." (Heisenberg, 1958)

Abstract

The aim of this thesis was to test predictive coding as a model of cortical organization and function using a specific brain response, the mismatch negativity (MMN), and a novel tool for connectivity analysis, dynamic causal modelling (DCM). Predictive coding models state that the brain perceives and makes inferences about the world by recursively updating predictions about sensory input. Thus, perception would result from comparing bottom-up input from the environment with top-down predictions. The generation of the MMN, an event-related response elicited by violations in the regularity of a structured auditory sequence, has been discussed extensively in the literature. This thesis discusses the generation of the MMN in the light of predictive coding, in other words, the MMN could reflect prediction error, occurring whenever the current input does not match a previously learnt rule. This interpretation is tested using DCM, a methodological approach which assumes the activity in one cortical area is caused by the activity in another cortical area. In brief, this thesis assesses the validity of DCM, shows the usefulness of DCM in explaining how cortical activity is expressed at the scalp level and exploits the potential of DCM for testing hierarchical models underlying the MMN. The first part of this thesis is concerned with technical issues and establishing the validity of DCM. The second part addresses hierarchical cortical organization in MMN generation, plausible network models or mechanisms underlying the MMN, and finally, the effect of repetition or learning on the connectivity parameters of the causal model.

Acknowledgments

I owe immense gratitude to my principal and secondary supervisors, Karl Friston and James Kilner, for providing excellent supervision and being so enjoyable to work with. James has guided me from the first steps of data acquisition and analysis, to the final stages of writing up. He has dedicated me generous hours of teaching and advice, always with extreme patience and wisdom. Karl has the perfect balance of support and freedom. He has been a constant source of inspiration, encouragement, and knowledge, flavoured by a brilliant sense of humour.

I also thank my co-authors Stefan Kiebel, Klaas Stephan, and Torsten Baldeweg for fruitful discussions and advice. Stefan has given crucial support in DCM.

I would like to thank the subjects for their time and cooperation in the studies that made this thesis possible. Thanks to David Bradbury for valuable support during data collection.

Thanks to Eric Featherstone for help in stimulus coding.

I sincerely thank the fellows in the Neurobiology and in the Methods group for insightful and humorous discussions every week.

Thanks to the multitude of people at the FIL for creating such an inspiring and friendly working atmosphere.

Special thanks to Tobias Overath for putting up with sharing offices since our very first day at the FIL, for coffee arrangements and especially for friendship.

I would like to thank Marcia Bennett for kindly dealing with my multiple travel arrangements during these years.

Many thanks to Peter Aston, Michelle Oliver, and Karen Lewis for having the greatest uplifting morning *hello* ever!

I am deeply grateful to Eduardo Ducla-Soares for his contagious enthusiasm in the Physics of the brain and for his continuous support since my years of undergrad.

This work was funded by the Wellcome Trust and the Portuguese Foundation for Science and Technology.

Table of contents

Abbreviations

A1	primary Auditory cortex
Bf	Bayes factor
DCM	Dynamic Causal Modelling
ECD	Equivalent Current Dipole
EEG	Electroencephalography
EM	Expectation-Maximization
ER	Evoked Response
ERF	Event-Related Field
ERP	Event-Related Potential
EOG	Electro-OculoGrams
fMRI	functional Magnetic Resonance Imaging
IFG	Inferior Frontal Gyrus
LGN	Lateral Geniculate Nucleus
MEG	Magnetoencephalography
MMN	MisMatch Negativity
MNI	Montreal Neurological Institute
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
STG	Superior Temporal Gyrus
SVD	Singular Value Decomposition

Figures and Tables

Outline and aims of this thesis

The aim of this thesis was to assess predictive coding as an explanation of cortical organization and function using a specific brain response, the mismatch negativity (MMN), and a novel tool for connectivity analysis, dynamic causal modelling (DCM) of event-related responses (ERs). The first part of the work described in this thesis will focus on the validation of DCM with real group data. In the second part, specific questions are formulated in terms of mechanistic hypotheses that map onto DCMs. Bayesian model comparison was the key for selecting the DCM, amongst the models tested, that best explains the data. The MMN was the paradigm selected for this research because it fits theoretically within the predictive coding framework and because it is a robust and prominent response in the ERP literature.

This thesis is organized as follows:

Chapter 1 – Introduction – is divided into two parts. The first part outlines the theoretical framework under which the research in this thesis was carried out. The second part presents a review of the genesis of a specific cortical response, the MMN, which is a concrete example of an experimental response that can be framed in the light of predictive coding.

Chapter 2 – DCM of ERPs: Methods – describes DCM, a novel tool which is validated and used for hypothesis testing in subsequent chapters.

Chapters 3-6 – Results chapters – describe the experimental work: the aims, the hypothesis or models tested, the set up and the outcomes of four studies. The specific goals of each study were the following:

- To access the validity of DCM for ERPs in terms of its reproducibility across a real multi-subject data set (**Chapter 3**)
- To investigate the role of backward connections in ERP generation and source activity as a function of time (**Chapter 4**)

- To identify an underlying connectivity model for the MMN generation (**Chapter 5**)
- To explore the effect of learning by repetition in the connectivity parameters of the underlying causal model (**Chapter 6**)

Chapter 7 – General Discussion and Conclusion – provides a general discussion and the conclusions of this work; presents its contributions to the field; and indicates directions for future research.

Chapter 1

Introduction

We experience the world through our senses. This sensory information from the environment is processed by the brain which constructs an inner model, or our perception, of what the world seems to be (Mumford, 1992). One hypothesis of how we might process that information is predictive coding, which states that perception is formed by combining inputs from the environment with predictions on that input (Rao and Ballard, 1999; Friston, 2005). Rather than a pictorial representation based on current sensorial information, perception rests upon recursive input-match-prediction loops. An analogy would be seeing an impressionistic picture, gradually extracting a realistic portrait and finally seeing its full detail. The brain's capacity to infer or fill in gaps is an important faculty when faced with ambiguous information, in novel environments and for error detection (Yuille and Kersten, 2006).

Imaging and electrophysiological techniques have played a fundamental role in understanding the human brain. In general, functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) have been utilised to identify which areas of the brain are active during any given process. These methods provide indirect measures of neuronal activity and have high spatial resolution but low temporal accuracy. In contrast, electroencephalography (EEG) and magnetoencephalography (MEG) provide direct measures of neuronal activity and have excellent time accuracy but relatively poor spatial resolution. In short, all these techniques are important for understanding *where* and *when* in the brain a neuronal process occurs, but it is the study of connectivity, such as described in this thesis, that is crucial for understanding *how* this processing is coordinated.

This thesis addresses two main points. Firstly, it focuses on the validation of dynamic causal modelling (DCM) for evoked responses measured with EEG. DCM is a generative or forward model which assumes that changes in cortico-cortical coupling are responsible for event-related potential (ERP) genesis. Secondly, it addresses construction of plausible connectivity models for hierarchical cortical organization motivated by predictive coding ideas. Critically, these models embody bottom-up and top-down connections among distant regions. The Mismatch Negativity (MMN), a response to a violation in the regularity of a structured auditory sequence, is the exemplar response chosen for study in this thesis.

1.1 Hierarchical Organisation and Predictive Coding in the brain

The notion that the brain is highly interconnected is critical for understanding brain function. Rather than studying brain areas in isolation, the perspective taken in this thesis is to look at the brain as a hierarchically organised system, in which active areas communicate with each other through synapses or changes in synaptic strength. The next section, on connectivity, will be important for the notion of predictive coding and Bayesian inference in the brain, as well as motivating the form of DCMs used later.

1.1.1 The connected brain

The neocortex corresponds to 94% of the total cerebral cortex, is made up of six layers (layers I-VI) and is 2-4 mm thick. More than half of the neocortex is dedicated to visual processing (55%), about 11% is devoted to somatosensory processing while 8% and 3% are involved in motor and auditory processing, respectively (Felleman and Van Essen, 1991). The idea that the brain might be hierarchically organised was put forward by Hubel and Wiesel (1962) in the visual domain. The notion that the brain is hierarchically organised also extends to other sensory modalities and integration between different modalities. Large-scale networks seem to be the best candidate solution for the binding problem; or in other words, for how we integrate the "symphony of emotions, perceptions, thoughts and actions" (Varela et al., 2001). A general principle of cortico-cortical connections is *reciprocity*: when two areas are linked through anti-parallel or bidirectional pathways. There are, however, a few exceptions to this rule, for example V1 has projections to V4 but V4 does not have projections to V1. Felleman and Van Essen (1991) described a set of connectivity rules previously noted by Rockland and Pandya (1979) and provided a critical assessment of the principle of hierarchical organization in the light of available data. These rules are formulated for extrinsic connections, i.e., excitatory connections that cross the white matter. In contrast, intrinsic connections are confined to an area within the cortical sheet and can be either excitatory or inhibitory in nature (see **Figure 1.1**). According to these rules, *forward* or ascending connections originate in agranular layers (I-III, V and VI) and terminate in the granular layer (layer IV). *Backward* or descending connections link agranular layers and *lateral* connections originate in agranular layers and target all layers. Forward and backward connections mediate bottom-up and top-down processing, respectively.

Figure 1.1 Connectivity rules in the brain. Forward or bottom-up connections originate in agranular layers and terminate in the granular layer. Backward or top-down connections link agranular layers. Lateral connections originate in agranular layers and target all layers. (Adapted from Felleman and Van Essen, 1991; David et al., 2005)

1.1.2 The dynamic brain, bottom-up and top-down processing

Most approaches that study brain function focus on "where in the brain" and "when in the brain" a given process is taking place. Implicitly such ideas treat brain areas as independent. In reality these areas are connected and the activity of one area is dependent on activity in other areas. Rather than inquiring about which areas are active at a particular time, the questions in this thesis will address how these areas communicate with each other, and which pathways or networks are active simultaneously. Mumford (1991, 1992), amongst others, has put forward ideas about the computational architecture of the neocortex that are in agreement with the view that the brain is hierarchically organised. In his work he discusses the notion of templates and residuals in thalamo-cortical and cortico-cortical loops. These ideas are based on the concept that each level takes its own part in the computation performed by the cortex. Namely, for two reciprocally connected areas, the "lower" area deals with more sensory or concrete information, whereas the "higher" area is concerned with more abstract information. The descending pathways carry templates which try to fit the information arriving via the senses. If the fit is not perfect the residuals would be sent upstream until top-down predictions and bottom-up constructions reach convergence. In his own words, "In the ultimate stable state, the deep pyramidal [cells] would send a signal that perfectly predicts what each lower area is sensing, up to expected levels of noise, and the superficial pyramidals wouldn't fire at all. [...] The brain would operate by a relaxation algorithm, in which the loop is repeated until it stabilizes [...] no more residuals to send upstream" (Mumford, 1992).

1.1.3 The brain as an empirical Bayesian device

Predictive coding ideas gained shape with the work by Rao and Ballard (1999) which demonstrated that an efficient strategy for encoding of natural images results from cortico-cortical feedback. Predictive coding formulations have been put together with notions of empirical Bayes, and statistical physics for solving the problem of perceptual learning and inference in the brain (Friston, 2003; 2005). Bayesian models have been employed in neuroscience and cognition by several research groups interested in the understanding of topics such as sensorimotor integration (Wolpert et al., 2005), sensorimotor decisions and learning (Körding and Wolpert, 2004; 2006), perception of tactile stimulation (Bays et al., 2006) and multisensory integration (for a review see Deneve and Pouget, 2004). Others employ the same ideas in visual perception (for a review see Kersten et al., 2004; Yuille and Kersten, 2006), in semantic memory (Steyvers et al., 2006), in understanding reasoning (Tenenbaum et al., 2006; Chater et al, 2006), and causality (Griffiths and Tenenbaum, 2005), decision-making (Behrens et al., 2007) and social interactions (Wolpert et al., 2003; Kilner et al., 2007). Hierarchical Bayes offers a good analogy for what may be happening in the brain: the prior probability of the causes, $p(\theta)$, formulated in higher hierarchical levels flow down to be combined with the likelihoods of the data given the causes, $p(y \mid \theta)$, in the lower hierarchical levels to compute the posterior probability, $p(\theta \mid y)$ which is then passed upstream to enter the following loop. According to the Bayes rule:

$$p(\theta \mid y) \propto p(\theta) p(y \mid \theta) \qquad (1.1)$$

This is a recursive process that stops once we reach reconciliation, i.e., until the inputs no longer cause updates to the posteriors of the generative model; or our recreation of what caused sensory input. In essence, this adds a probabilistic flavour to the template notion of pattern recognition discussed in Mumford (1992). Moreover, it makes the link to the neurobiological substrate underlying this computation, i.e., message-passing in cortical hierarchies and learning through changes in synaptic efficacy or connection strength. In brief, the brain tries to infer the causes of sensory input (i.e. builds a generative model) and to learn the relationship between input and cause (i.e. builds recognition models) by adjusting the synaptic efficacy so that the free energy is minimized (Friston et al., 2005). The concept of free energy refers to the difference between the true probability distribution over the causes and our guess of what it might be. Perceptual inference in the brain involves re-entrant dynamics that self-organise in order to suppress the free energy or prediction error. Prediction error corresponds to the mismatch between the predicted state of the world, at any level, and that predicted on the basis of the state in the level above. During suppression or minimization of the free energy, the brain changes its configuration by adjusting its parameters. This is done so that the brain's internal representations of the world match those furnished by the external inputs. From an empirical Bayesian perspective, perceptual inference at any level in the brain rests upon a balance between top-down priors from the level above via backward connections, and a bottom-up likelihood term encoded by forward connections from the level below (Friston et al., 2006a).

The following section provides a review of the MMN literature, especially focussing on the underlying cortical mechanisms behind the MMN. The hypothetical framework for this specific brain response is described at the end of this chapter. In brief, it is proposed that the MMN can be understood in the light of predictive coding and empirical Bayes (see **Figure 1.5**). In this view, evoked responses (ER) correspond to prediction error; in other words, ERs are an expression of unpredictable events. In this thesis, the MMN was chosen as an exemplar of an ER that corresponds to prediction error, but this can presumably be extended for ERs in general. Predictive coding states that prediction error is conveyed to higher cortical areas via forward connections, where predictions are updated in the light of new available data. These predictions, the posteriors or empirical priors in the subsequent loop, are then send back to the lower cortical areas via backward connections. This is repeated at all levels, so higher levels provide guidance to lower levels. Hence, this recurrent process ceases when reconciliation between predictions and sensory input is reached. The research performed in this thesis provides experimental evidence that the MMN is an ER that can be interpreted in the light of predictive coding (Friston, 2005). In summary, the predictive coding framework postulates that evoked responses correspond to prediction error that is explained away during perception and is suppressed by changes in synaptic efficacy during perceptual learning. In this context, the MMN would be the result of prediction error, which is due to an unexpected deviant, or oddball, embedded in learnt sequences of standard events. The MMN would arise when there is a mismatch between the current stimulus input (unpredictable deviants) and a memory trace of previous input (predictable standards). This is the working hypothesis that inspired the research described below.

1.2 The mismatch negativity: a review of the underlying mechanisms

The MMN is a brain response to violations of a rule established by a sequence of sensory stimuli (typically in the auditory domain) (Näätänen, 1992). The MMN reflects the brain's ability to perform automatic comparisons between consecutive stimuli and provides an electrophysiological index of sensory learning and perceptual accuracy. Although the MMN has been studied extensively, the neurophysiological mechanisms underlying the MMN are not well understood. Several hypotheses have been put forward to explain the generation of the MMN; amongst these accounts, the "*adaptation hypothesis*" and the "*model adjustment hypothesis*" have received most attention. This section presents a review of studies that focus on neuronal mechanisms underlying the MMN generation, discusses the two major explanative hypotheses, and proposes predictive coding as a general framework that attempts to unify both.

1.2.1 The MMN: a brief introduction

Small changes in the acoustic environment engage an automatic auditory change detection mechanism reflected in the MMN. The presentation of an *oddball* or *deviant* event, embedded in a stream of repeated or familiar events, the *standards*, results in an evoked response that can be recorded non-invasively with electrophysiological techniques such as EEG and MEG. The MMN is the negative component of the waveform obtained by subtracting the event-related response to the *standard* event from the response to the *deviant* event. This brain response is measured with EEG and has a magnetic counterpart called MMNm. The MMN is elicited by sudden changes in stimulation; peaks at about 100-250 ms from change onset and exhibits the strongest intensity in temporal and frontal areas of topographic scalp maps (Sams et al., 1985). Given its automatic nature, the MMN might be associated with pre-attentive cognitive operations in audition and, for this reason, it has been suggested that it reflects 'primitive intelligence' in the auditory cortex (Näätänen et al., 2001). Here this notion is finessed and it is suggested that the mechanisms behind the generation of the MMN can be understood within a predictive coding framework that appeals to empirical Bayes.

While the MMN has been studied intensively in the auditory modality, some studies show evidence for the existence of a visual MMN counterpart (Astikainen et al. 2004; Czigler et al. 2004; see Pazo-Alvarez et al., 2003 for review). Omitted stimuli or deviances such as direction of movement, form, orientation, location, contrast, size, spatial frequency and colour, elicit a negative component in the N2 latency range (250 – 450 ms). Nevertheless, there is controversy as to whether these N2-like waves elicited by a visual stimulus change reveal the same degree of automaticity as in the auditory MMN, and whether the emergence of this component is really based on a memory comparison process. A potential analogue to the MMN has also been reported in the somatosensory system, which seems to be generated in fine discrimination tasks (Kekoni et al., 1997; Akatsuka et al., 2005). Numerous studies have focused on ERP scalp maps, especially in clinical applications, when comparing, for instance, schizophrenic patients (Umbricht et al., 2003) or dyslexic subjects (Baldeweg et al., 1999) with normal controls. The MMN has also been proved useful in understanding auditory perception and formation of sensory memory representations (Atienza et al. 2002, van Zuijen et al., 2006).

A major area of MMN research is concerned with the underlying neuronal mechanisms of its generation. Several competing hypotheses have been put forward, based on experimental results obtained with ERPs, MEG and fMRI. The most common interpretation is that the MMN arises whenever there is a break of regularity in a structured auditory sequence (Näätänen, 1992), and that a temporo-prefrontal network, comparing the current sensory input with a memory trace of previous stimuli, is responsible for generating the MMN at the scalp level (Giard et al., 1990; Rinne et al., 2000; Opitz et al., 2002; Doeller et al., 2003). From this perspective, the MMN is assumed to reflect an automatic auditory change detection process that triggers a switch in the focus of attention (Escera et al., 1998; 2003). However, this notion has been challenged recently by claims that the MMN rests on a much simpler mechanism, namely *neuronal adaptation* in the auditory cortex. The adaptation hypothesis proposes that the apparent MMN results from the subtraction of a N1 response to a novel sound, from the N1 response to a non-novel or repeated sound; where the N1 to a repeated sound is delayed and suppressed, as novelty decreases (Jääskeläinen et al., 2004).

The following section reviews a variety of studies that have focused on a mechanistic understanding of how the auditory MMN is generated, discusses the major hypotheses, and suggests a general and unifying framework, predictive coding, for understanding the MMN. This point will be discussed in further detail in the final chapter of this thesis (see General Discussion and Conclusion, **Chapter 7**). As defined in the previous section, predictive coding is a general theory of perceptual inference. Under predictive coding the brain is regarded as a hierarchically organized cortical system, in which each level strives to attain a compromise between bottom-up information about sensory inputs provided by the level below and top-down predictions (or priors) provided by the level above (Mumford 1992; Rao and Ballard 1999; Friston, 2003). Within this framework the MMN would result from a failure to predict bottom-up input and consequently to suppress prediction error (Friston, 2005; Baldeweg, 2006, Garrido et al., 2007a; see also **Chapter 3**). The predictive coding account of the MMN unifies the competing hypotheses of *neuronal adaptation* and *model adjustment* (Garrido et al., in submission). This will be discussed further in **Chapter 5**.

1.2.2 General characteristics of the MMN

1.2.2.1 Scalp topography

The MMN is the negative component of a difference wave between responses to standard and deviant events embedded in an *oddball* paradigm. This negative response, of about 5μV maximum peak, is distributed over auditory and frontal areas, with prominence in frontal regions with a reversed polarity at mastoid sites (see **Figure. 1.2**).

The MMN peaks at about 100 to 250 ms after change onset but this latency varies slightly according to the specific paradigm or the type of regularity that is violated: frequency, duration, intensity, or the inter-stimulus interval (Näätänen et al., 2004) (see **Figure. 1.2**). In more complex paradigms an abstract rule is broken, such as inter-stimulus relationships (Tervaniemi et al., 1994; Paavilainen et al., 2001; Vuust et al., 2005) or phoneme regularity (Näätänen, 1997). Barely discriminable tones elicit a later MMN peaking at about 200–300 ms (Näätänen and Alho, 1995).

Figure 1.2 Scalp topography and latency of the MMN. **a**: ERP responses to standard and deviant tones overlaid on a whole scalp map of 128 EEG electrodes. **b**: ERP responses to the standard and deviant tones at a fronto-central channel. **c**: MMN, difference wave obtained by subtracting ERP to standards from ERP to deviants. **d**: MMN response averaged over a time window of [100, 200] ms interpolated to give a 3D scalp topography. (From Garrido et al., 2007a)

1.2.2.2 MMN under different paradigms

The MMN is elicited in the presence of any discriminable change in some repetitive aspect of auditory stimulation. This discriminable change can be of different types: frequency, duration, intensity, perceived sound-source location, silent gap instead of a tone, or one phoneme replaced by another. In a recent study, Näätänen et al., (2004) proposed a new paradigm in which a standard alternates with one of five deviant types that differ in duration, location, intensity, gap and frequency. Because of its effectiveness, this paradigm is particularly useful in clinical research as it can be used to obtain five different types of MMN responses over the same experimental time, whereas only one type of MMN is obtained in traditional paradigms.

It is generally believed that the MMN is evoked by any violation of an acoustic regularity or pattern. Indeed, the MMN is elicited by violations of abstract rules established in a structured auditory sequence (Näätänen et al., 2001). For example, in complex auditory patterns, it has been found that an MMN is elicited by an occasional ascending tone or tone repetition in a sequence of regularly descending tone pairs (Tervaniemi et al., 1994); by changing the direction of within-pair frequency change, independently of their absolute frequencies (Saarinen et al., 1992); and by violations of the rule that the higher the frequency, the louder the intensity (Paavilainen et al., 2001). The MMN is also detected when the stimuli are spectrally rich. This type of paradigm facilitates attentive pitch discrimination in comparison to pure sinusoidal tones; in other words, the MMN is larger and has shorter latency (Tervaniemi et al., 2000a). Moreover, MMN responses are elicited by violating regularity in roving paradigms (Baldeweg et al. 2004, Haenschel et al., 2006, Garrido et al., in submission, see also **Chapter 6**), or in more sophisticated paradigms comprising irregularities in rhythms (Vuust et al, 2005), musical sequences (van Zuijen et al., 2004;), and violations in phoneme regularity (Näätänen, 1997).

1.2.2.3 An index of memory traces?

It is commonly accepted that the MMN rests on the relation between the present and previous stimuli, rather than on the stimulus alone. Hence, the MMN may depend on a memory trace formed by preceding stimuli at the beginning of a stimulus block; i.e., during the presentation of the *standard* events. If the *deviant*, or the *new* event, occurs while this memory trace is still active, automatic change-detection is activated, giving rise to a MMN response (Näätänen, 2000). The duration of this period, also called echoic memory, has been reported to last at least 10s in normal subjects (Bottcher-Gandor and Ullsperger, 1992).

1.2.2.4 Dependence on attention?

The MMN is one of the earliest ontological cognitive components that can be observed in an ERP trace (Alho et al., 1990). An important characteristic of the MMN response to an auditory *oddball* paradigm is the fact that it can be detected even when the subject is not paying attention. The MMN can be measured without any task requirements and is elicited even when the subject performs a task that is not related to the stimulus. The MMN can be elicited irrespective of attention, during non-attentive states such as sleep (Sallinen et al., 1994), or even in coma; where the presence of a MMN has been proposed as a predictor for recovery of consciousness (Kane et al., 1993). This demonstrates the brain's capacity to perform complex comparisons between consecutive sounds automatically (Näätänen et al., 2001). Although the MMN is seldom affected by attention, some studies suggest that the MMN is attenuated when the subject's attention is outside the focus of the auditory stimulus (Arnott and Alain, 2002; Müller et al., 2002). On the other hand, the degree to which the visual stimulus is attended does not seem to have an influence on the MMN (Otten et al., 2000). To avoid overlap with other ERP components some authors argue that the best condition to observe an MMN is when subject attention is directed away from the stimulus (Näätänen, 2000).

It has been reported that the generation of the MMN, in particularly the source over the frontal lobe, is associated with an involuntary attention switching process, an automatic orienting response to an acoustic change (Escera et al., 1998; 2003). In addition, it has been suggested that the frontal generator of the MMN is related to an involuntary amplification or contrast enhancement mechanism that tunes the auditory change detection system (Opitz et al., 2002).

1.2.3 The relevance of the MMN and its applications

The fact that MMN can be elicited without special task requirements, independently of the subject's motivation and in the absence of attention, during sleep, or even before coma recovery, makes it particularly suitable for testing different clinical populations, infants and newborns (see Kujala et al., 2007 for a recent review). The following two subsections present a brief review of recent studies that used the MMN to address important questions in cognitive processing and clinical neuroscience.

1.2.3.1 MMN in cognitive studies

The MMN is considered to represent the only objective marker for auditory sensory accuracy (Näätänen, 2000). MMN studies have made important contributions to our understanding of the formation of auditory perception and streaming (see Denham and Winkler, 2006; for a review), construction of sensory memory representations, as well as how these are influenced by attention (Sussman et al., 1998; Sussman and Steinschneider 2006). It has been shown that whereas attention is not always necessary for auditory stream segregation (Sussman et al., 2007), switches in attention are important for streaming reset (Cusack et al., 2004). Woldorff et al. (1993) have shown that focused auditory attention can modulate sensory processing as early as 20 ms. Others have used the MMN to characterise the mechanisms of involuntary attention switching (Escera et al., 1998; 2003).

Several studies have used the MMN to understand mechanisms of perceptual learning. Tremblay et al. (1998) showed that training-associated changes in neural activity, indicated by the MMN, precede behavioural discrimination of speech. The MMN was also found to correlate with gains in auditory discrimination after sleep (Atienza et al., 2002; 2005). Implicit, intuitive and explicit knowledge have been characterized in terms of the elicited responses; the MMN and P3, combined with behavioural measures (van Zuijen et al., 2006).

1.2.3.2 MMN in clinical neuroscience

The MMN has proved useful in various clinical contexts (see Näätänen, 2000; 2003 for reviews on clinical research and applications). The most promising clinical application of MMN is in schizophrenia research. More than 30 studies have found significant reductions of MMN amplitude in patients with schizophrenia, both for frequency and duration deviants (Umbricht & Krljes, 2005). Moreover, individual MMN amplitudes correlate with disease severity and cognitive dysfunction (Baldeweg et al 2004) and functional status (Light and Braff, 2005), although there are conflicting reports about its association with genetic risk for schizophrenia (Michie et al., 2002; Bramon et al., 2004). Two features make the MMN a particularly interesting paradigm for schizophrenia research (see Stephan et al., 2006 for a review). First, the MMN depends on intact NMDA receptor signalling: pharmacological blockage of NMDA receptors has been shown to significantly reduce the MMN, both in invasive recordings studies in monkeys (Javitt et al., 1996) and human EEG/MEG studies (Kreitschmann-Andermahr et al., 2001; Umbricht et al., 2000; 2002). This is important because the NMDA receptor has a critical role in the plasticity of glutamatergic synapses, which is at the core of current pathophysiological theories of schizophrenia (Friston and Frith, 1995; Harrison &

Weinberger, 2005; Javitt, 2004; Stephan et al., 2006). Second, clinical investigations of schizophrenic patients require very simple paradigms that are robust to changes in attention and performance. As discussed above, the MMN fulfils these requirements very well.

The MMN has proved useful for investigating several diseases in addition to schizophrenia. Another important application is in the field of dyslexia: dyslexic patients show a diminished MMN, albeit only for frequency deviants and not for duration. This suggests that dyslexia is associated with auditory frequency discrimination impairment (Baldeweg et al., 1999). A reduced MMN in children with learning disabilities suggested that the deficit originates in the auditory pathway at a processing stage prior to conscious perception (Kraus et al., 1996). This is in accord with Rinne et al. (1999) who showed that speech processing occurs at early pre-attentive stages in the left hemisphere (at about 100-150 ms after sound onset).

1.2.4 The mechanisms of MMN generation

Despite the vast literature on MMN research, the mechanisms that underlie its generation remain a matter of debate. Two major competing hypotheses have emerged, the *model adjustment hypothesis* and the *adaptation hypothesis*. The following subsections describe these two competing hypotheses and discusses the experimental evidence that favours one or the other. Finally, predictive coding is suggested as a general unifying framework that can accommodate both hypotheses. This idea is supported by the results from the connectivity modelling approach to the MMN described in **Chapters 5** and **6** (see also Garrido et al., in submission).

1.2.4.1 The model adjustment hypothesis

The MMN can be regarded as an index of automatic change-detection governed by a pre-attentive sensory memory mechanism (Tiitinen et al., 1994). Several studies have proposed mechanistic accounts of how the MMN might be generated. The most common interpretation is that the MMN is a marker for error detection caused by a break in a learned regularity or familiar auditory context. The MMN would then result from the difference, or mismatch, between the current and the preceding input. The early work by Näätänen and colleagues suggested that the MMN results from a comparison between the present auditory input and the memory trace of previous sounds (Näätänen 1992). In agreement with this theory, others (Winkler et al., 1996; Näätänen and Winkler, 1999; Sussman and Winkler, 2001) have postulated that the MMN could reflect on-line modifications of a perceptual model that is updated when the auditory input does not match its predictions. This is the so called *model-adjustment hypothesis*. In the context of the model adjustment hypothesis, the MMN is regarded as a marker for error detection, caused by a deviation from a learned regularity. In other words, the MMN results from a comparison between the auditory input and a memory trace of previous sounds, reflecting an on-line updating of the model for predicting auditory inputs (Winkler et al., 1996; Näätänen and Winkler, 1999). According to this hypothesis, the MMN is a response to an unexpected stimulus change. This hypothesis has been supported by Escera et al. (2003) who provided evidence for the involvement of the prefrontal cortex in providing top-down modulation of the deviance detection system in the temporal cortices. In the light of Näätänen's model, it has been claimed that the MMN is caused by two underlying functional processes, a sensory memory mechanism related to temporal generators and an automatic attention-switching process related to the frontal generators (Giard et al., 1990). The role of prefrontal generators is supported by studies of patients with prefrontal lesions who showed diminished temporal MMN amplitudes (Alain et al., 1998). Furthermore, it has been shown that the temporal and frontal MMN sources have separate temporal dynamics (Rinne et al., 2000) but interact with each other (Jemel et al., 2002). This notion is also compatible with strong and reciprocal anatomical connectivity between auditory and prefrontal areas that has been found in primate tract tracing studies (Romanski et al., 1999). According to source reconstruction studies, the generators of the MMN are located bilaterally in the temporal cortex (Hari et al., 1984; Giard et al., 1990; Alho, 1995). In addition, there is evidence for generators in the prefrontal cortex, often stronger and reported more consistently on the right hemisphere for tone paradigms (Levänen et al., 1996) and on the left hemisphere for language paradigms (Näätänen et al., 1997; Tervaniemi, 2000b; Pulvermüller, 2001). A sensory memory mechanism has been associated with the temporal generators, whereas a cognitive role, or comparator-based mechanism, has been assigned to the prefrontal generators (Giard et al., 1990; Gomot et al., 2000; Maess et al., 2007). Numerous studies have consistently reported evidence for multiple generators of the MMN in the primary auditory cortex (**A1**). This has been reproduced across different modalities such as EEG (Deouell et al., 1998; Jemel et al., 2002; Marco-Pallarés et al., 2005; Grau et al., 2007), MEG (see for example Tiitinen et al., 2006; or Hari et al., 1984) and combined EEG with MEG measures (Rinne et al., 2000). The latter study revealed that prefrontal generators are activated later than the generators in the auditory cortex; this supports the hypothesis of a change detection mechanism in the prefrontal cortex, which is triggered by the temporal cortex. This study found temporal sources with both M/EEG, whereas prefrontal sources were only found with EEG; possibly because the frontal sources have a radial orientation and the MEG sensors are blind to the fields generated by radial sources (see **Figure 1.3**). An alternative explanation for why source current distribution looks different in EEG and MEG is the higher level of spatial smearing in EEG source reconstruction compared to EEG.

Figure 1.3 MMN generators estimated from EEG and MEG data. The centre of gravity changes from temporal to frontal areas over time. Frontal sources were detected only with EEG, either due to their radial orientation or to the higher level of spatial smearing in EEG source reconstruction. These sources were determined with minimum norm estimates (MNE). (Adapted from Rinne et. al, 2000)

Functional MRI (Molholm et al., 2005; Rinne et al. 2005) and combined fMRI-EEG studies (Opitz et al., 2002; Doeller et al., 2003; Liebenthal et al., 2003) have reported findings that are consistent with the results of the source reconstruction studies described above. Some of the combined fMRI-EEG studies show a double peak over frontal scalp locations suggesting the existence of two subcomponents for the MMN. Dipole modelling was performed in two time windows to explain the scalp ERP distribution (Opitz et al., 2002 and Doeller et al., 2003). The early component is reported to peak in the time window of 90 – 120 ms and can be modelled with dipoles located bilaterally in the superior temporal gyrus (**STG**). ERPs within the late time window, 140 – 170 ms, can be modelled with dipoles placed in left and right inferior frontal gyrus (**IFG**) (see **Figure 1.4**). The sources in the temporal areas might be involved in processing changes of the sound's physical properties, whereas the sources on the frontal areas might reflect reorientation of attention. Recent work has linked the early component (in the range of about 100-140 ms) to a sensorial, or non-comparator account of the MMN, originated in the temporal cortex, and the later component (in the range of about 140-200 ms) to a comparator-based mechanism of the MMN, involving the prefrontal cortex (Maess et al., 2007). Although MMN sources are found consistently over temporal and pre-frontal regions, a few studies have reported sources at other locations such as right temporal and parietal lobes (Levänen et al., 1996). Thus, these studies suggest that the MMN is generated by a temporo-frontal network which supports the model adjustment hypothesis.

Figure 1.4 MMN underlying sources revealed by EEG and conjoint EEG and fMRI measures. **(a)** Dipoles indicated by red arrows at bilateral **STG** and **IFG** (adapted from Doeller et al., 2003). **(b)** Dipole locations at bilateral **STG** and right **IFG** and **(c)** significant fMRI activation for deviants (adapted from Opitz et al. 2002). **(d)** Most significant independent component (computed by ICA-LORETA analysis, adapted from Marco-Pallarés et al., 2005). This figure shows consistency for MMN sources across different modalities.

1.2.4.2 The adaptation hypothesis

A recent study (Jääskeläinen et al., 2004) has challenged the more common view that the MMN is generated by a temporal-frontal cortical network. Instead, they suggest that the MMN results from a much simpler mechanism of local neuronal *adaptation* at the level of the auditory cortex, causing attenuation and delay of the N1 response. The N1 response is the negative component peaking at about 100 ms after stimulus onset and is associated with early auditory processing at the level of A1. They propose that the N1 response to standard (or 'non-novel') sounds is delayed and suppressed (or *attenuated*) as a function of its similarity to the preceding auditory events, reflecting short-lived adaptation of auditory cortex neurons[1]. As a

[1] Neuronal adaptation, or spike-frequency adaptation, results from activation of calcium-dependent potassium channels that lead to slow after hyperpolarizing currents, decreasing neuronal excitability

consequence, the observed response would be erroneously interpreted as a separate component from the N1 wave. According to the *adaptation hypothesis*, the fact that the neuronal elements within the auditory cortex become less responsive during continuous stimulation is sufficient to explain the generation of an apparent MMN. With the generation of a delayed and suppressed N1 in the auditory cortex, the MMN would emerge as a product of an N1 differential wave when subtracting the ERP to the standards from the ERP to the deviant.

The adaptation hypothesis rests on previous MEG studies indicating the presence of two subcomponents of the N1 response: a posterior subcomponent, N1p, peaking at about 85 ms from stimulus onset, and an anterior subcomponent, N1a, peaking at about 150 ms (Loveless et al., 1996). The amplitude of the posterior component is strongly suppressed during the presentation of identical stimuli, whereas this adaptation effect is smaller for the anterior component. In contrast to previous studies showing that repetitive standard sounds constitute a prerequisite to MMN elicitation, Jääskeläinen et al. (2004) furnish evidence for robust MMN to infrequent (or 'novel') stimuli when preceded by a single standard stimulus. Consistent with the *adaptation* hypothesis, EEG measurements employing small deviances around a standard tone demonstrate that the smaller the frequency separation between the standard and the deviant, the more the amplitude to the deviants is attenuated (May et al., 1999).

Although the model-adjustment and the adaptation hypotheses come across as opponents, adaptation pertains to neurophysiological mechanisms (Jääskeläinen et al., 2004), whereas model-adjustment speaks to functional or perceptual mechanisms (Winkler et al., 1996; Näätänen and Winkler, 1999). While the latter allows for adaptation effects (which the authors refer to as refractoriness), the adaptation hypothesis precludes a prediction or model-dependent contribution to the MMN. The concept of refractoriness has been introduced by Näätänen (1992). In this view, the MMN occurs through the activation of non-refractory cells by 'fresh' afferents. When a standard stimulus activates neurons that respond to a given frequency within their receptive field, these neurons will temporarily become refractory to further

and firing rate (see Faber & Sah 2003 for review). Adaptation is thus a local phenomenon that is independent of pre-synaptic mechanisms and rests on changes in post-synaptic responsiveness.

stimulation. Therefore, further standard stimulation within this refractory period will evoke a smaller response. However, a deviant stimulus, i.e., with a different frequency, occurring within this period will activate other neurons. The response elicited by the deviant is therefore larger than that to the standard. (See Picton et al., 2000, for a review.)

Adaptation is a compelling hypothesis for the generation of the MMN that explains the experimental results mentioned above. However, there are other empirical observations that are not compatible with the *adaptation* hypothesis (see Näätänen et al. (2005) for a critical assessment on the *adaptation* view of Jääskeläinen et al. 2004). One of the points against *adaptation* is the fact that it predicts that the MMN duration and latency should match those of the N1, which has been shown not to be the case (Winkler et al., 1997). Secondly, *adaptation* does not explain why an MMN can be elicited in the absence of a N1 response, for example, during sleep (Atienza and Cantero, 2001) or when unexpectedly omitting a stimulus (Yabe et al., 1997). However, one potential defence in favour of the adaptation hypothesis rests on the notion that neuronal dynamics, induced by the rhythmic stimulation, continue to oscillate upon cessation or interruption of stimulation (May et al., 1999). A third point of controversy is that, as mentioned above, the violation of abstract rules or complex inter-stimulus relationships can also elicit an MMN. For instance, an ascending tone pair in a sequence of descending tone pairs elicits an MMN (Saarinen et al., 1992) even though there is no stimulus repetition that could cause adaptation of a frequency-specific neuronal population. Given the tonotopic structure of auditory cortex, MMNs of this sort cannot be explained by local adaptation but must result from more complex mechanisms involving more than one neuronal population. Moreover, the scalp distribution of the MMN is different from the N1 (Giard et al., 1990). While the N1 components are larger in amplitude over the contralateral hemisphere, the MMN response to changes in acoustical features is right-hemispheric dominant (Levänen et al., 1996) and left-hemispheric dominant for phoneme changes, irrespective of the ear stimulated (Näätänen et al., 1997). Another finding that can not be explained by *adaptation* alone is that equivalent current dipole (ECD) modelling reveals that the temporal source underlying the MMN is located anterior to the source underlying the N1 (Hari et al., 1992; Tiitinen et al., 1993). Furthermore, the MMN has a second source in the frontal lobe, which expresses temporal dynamics that are distinct from the N1 source (Opitz et al., 2002, Doeller et al., 2003; Molholm et al., 2005; Grau et al., 2007). Evidence for a frontal generator was also provided from direct intracranial recordings in human epilepsy patients (Rosburg et al., 2005). Finally, pharmacologic manipulations show that NMDA antagonists block the generation of MMN without affecting activity in the primary auditory cortex (Javitt et al., 1996), which suggests that the MMN and the N1 employ different neuronal populations and are expressions of separate cortical processes. Finally, if the MMN results from neuronal adaptation, one would predict changes in MMN following manipulations of serotoninergic and muscarinic receptors. This is because activation of these receptors is known to enhance neuronal adaptation (c.f. McCormick et al., 1998). As described above, however, there is only weak and contradictory empirical evidence for MMN modulation by serotoninergic and muscarinic agents.

1.2.4.3 The MMN from the perspective of predictive coding

Predictive coding (or, more generally, hierarchical inference in the brain) states that perception arises from integrating sensory information from the environment and our predictions based on a model of what caused that sensory information. Prediction error is minimised through recurrent interactions among levels of a cortical hierarchy in order to estimate the most likely cause of the input (Friston, 2003; 2005). The model adjustment hypothesis explains the MMN as a marker for error detection caused by a deviation from a learned regularity. The MMN would thus result from a comparison between the auditory input and a memory trace that is embodied in top-down predictions. The ensuing prediction error could then be used for on-line updating of a model for predicting auditory inputs (Winkler et al., 1996; Näätänen and Winkler 1999). This is consistent with the predictive coding framework, where current inputs are predicted from past inputs (see **Figure 1.5**). In the case of a prediction error, i.e. when there is a mismatch between the predicted and the actual sensory input, the neural system implementing the model must be adjusted (for example, by short-term synaptic plasticity). During the repetition of subsequent standards, that adjustment is reflected neurophysiologically in the suppression of prediction error and the disappearance of the MMN (Friston, 2005; Baldeweg, 2006). This view is identical to predictive coding models of vision, which postulate that perception relies on hierarchically organised neural systems in which each level compares predictions from higher-level areas to incoming sensory information from lower areas (Rao and Ballard, 1999; Yuille and Kersten, 2006): Using backward connections, higher cortical areas attempt to fit their abstractions, or learned reconstructions of the world, to the data received from lower cortical areas. The lower areas, in turn, attempt to reconcile the predictions from higher areas with the actual input, and return, by means of forward connections, a prediction error signal, i.e. information on the features not predicted by the higher areas (Mumford, 1992). Hence, lower and higher areas communicate via reciprocal pathways until reconciliation; in other words, until the prediction error is suppressed and the encoding of sensory causes at higher cortical areas is optimised (Friston, 2003).

Critically, hierarchical inference (e.g., predictive coding) also rests on optimising the relative influence of bottom-up prediction error and prediction error based on top-down prior expectations. This is involves optimising the efficacy of lateral interactions or intrinsic connections, within an area or source (Friston, 2003). Put simply, when a standard stimulus can be predicted more precisely by top-down afferents, less weight is assigned to bottom-up influences and the post-synaptic responsiveness to sensory inputs decreases. This is exactly what the adaptation hypothesis predicts. In short, hierarchical inference, using prediction error, provides a principled framework in which the model adjustment and adaptation heuristics become necessary for sensory inference.

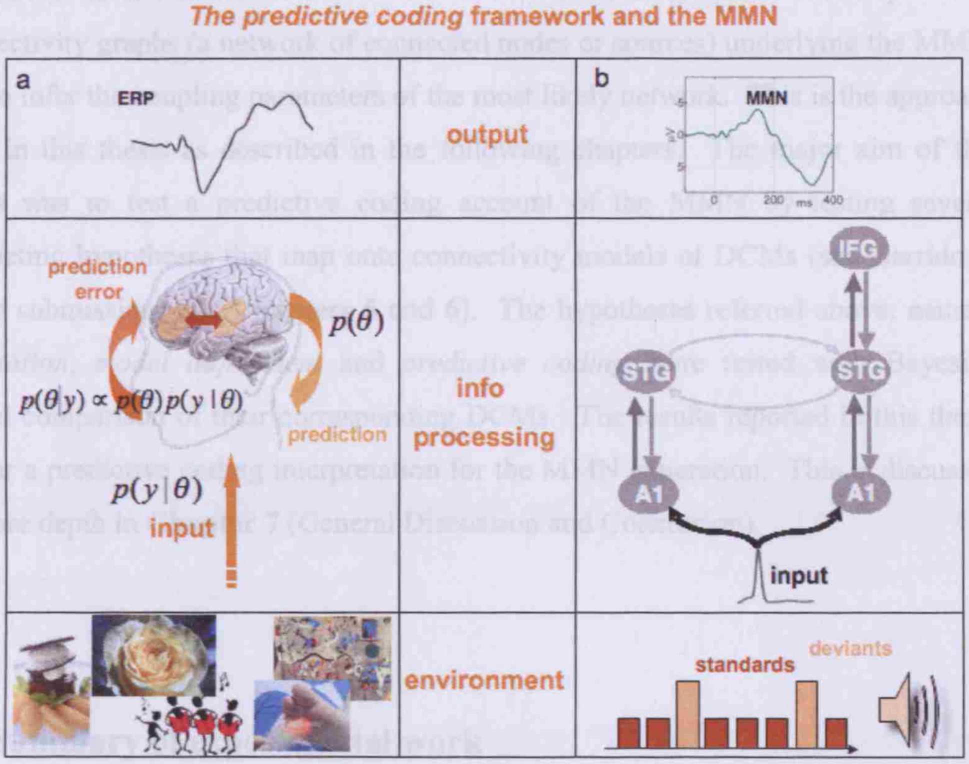


Figure 1.5 The MMN interpreted in terms of predictive coding. (a) illustrative schematic of the general framework of hierarchical Bayes and predictive coding as an explanation for ERPs. **(b)** the MMN, a concrete example and plausible underlying mechanisms.

Predictive coding formulations entail specific mechanisms that might underlie the MMN. A promising approach, to address these mechanisms, is to create biophysically realistic models that can represent competing hypotheses. These models can be tested empirically and provide evidence to disambiguate amongst competing theories. A pioneering study of this sort was performed by May et al. (1999) who constructed a model of tonotopically organised auditory cortex consisting of leaky integrate-and-fire neurons and compared its predictions to experimentally measured MEG/EEG data. Their question was whether the MMN could be explained by a local post-synaptic mechanism (i.e., neuronal adaptation) alone, or whether additional non-local synaptic mechanisms were required. They chose lateral inhibition (i.e., reciprocal inhibitory connections amongst neighbouring neuronal populations) as a candidate mechanism of the latter sort. They found that their experimental data could best be approximated by a model that combined adaptation and lateral inhibition.

Another class of models are those that use DCM to test the likelihood of plausible connectivity graphs (a network of connected nodes or sources) underlying the MMN, and to infer the coupling parameters of the most likely network. This is the approach used in this thesis as described in the following chapters. The major aim of this thesis was to test a predictive coding account of the MMN by testing several competing hypotheses that map onto connectivity models or DCMs (see Garrido et al., in submission; and **Chapters 5** and **6**). The hypotheses referred above, namely *adaptation*, *model adjustment* and *predictive coding* were tested with Bayesian model comparison of their corresponding DCMs. The results reported in this thesis favour a predictive coding interpretation for the MMN generation. This is discussed in more depth in **Chapter 7** (General Discussion and Conclusion).

1.3 Summary of experimental work

Chapters 3-6 describe the experimental work developed for this thesis which entailed testing a predictive coding account of the MMN generation by using DCMs. Having established the predictive validity of DCM for ERPs in **Chapter 3**, DCM is used to investigate the role of backward connections, or top-down processing in ERP generation and source activity as a function of time; this is described in **Chapter 4**. The selection of a DCM for the MMN, amongst plausible mechanistic hypotheses, namely *adaptation*, *model adjustment* and *predictive coding*, is presented in **Chapter 5**. Finally, the effects of learning (through repetition) on the connectivity parameters of the underlying causal model are explored in **Chapter 6**.

Chapter 2

Dynamic Causal Modelling of Event-Related Potentials: Methods

This chapter describes Dynamic Causal Modelling (DCM) of event-related potentials (ERPs). DCM is the method used in this thesis to test the alternative accounts of ERP generation, specifically for the MMN. This chapter provides an extended description of DCM. All results chapters in this thesis (**Chapters 3-6**) use this method. The specific methods for each study will be described in the corresponding chapter.

2.1 Dynamic Causal Modelling

DCM is a procedure that models interactions among cortical regions, allows one to make inferences about system parameters and investigate how these parameters are influenced by experimental factors. With DCM it is possible to assess how a given experimental manipulation activates a cortical pathway rather than a cortical area or source. This approach uses a biologically informed model that allows for inferences about the underlying neuronal networks generating evoked responses such as event-related potentials (ERPs) and fields (ERFs).

Most approaches to connectivity in the MEG/EEG literature use functional connectivity measures, such as phase-synchronisation, temporal correlations or coherence between the activities of two areas at the scalp or source level. Functional connectivity is used to establish statistical dependencies between time series and is useful because it does not require an underlying model or knowledge of the causal nature of the interactions. However, there are circumstances when the causal architecture of the interactions is the focus of interest. As opposed to functional connectivity, DCM uses the concept of effective connectivity, which refers explicitly to the influence one neuronal system exerts over another. This influence is parameterised in a causal or an explicit generative model, which can then be estimated using model inversion. In DCM, the brain is regarded as a deterministic nonlinear dynamic system that is subject to inputs and produces outputs (Friston et al., 2003). Effective connectivity is estimated by perturbing the system and measuring the response using Bayesian model inversion. Furthermore, by taking the marginal likelihood over the conditional density of the model parameters, one can estimate the probability of the data given a particular model. This is known as the marginal likelihood or evidence and can be used to compare different models, or mechanistic hypotheses underlying a specific perceptual or cognitive process.

As opposed to other connectivity tools such as structural equation modelling or autoregressive models, DCM treats inputs as known and determinist, whereas the former models treat inputs as stochastic. Another point of departure is that conventional connectivity approaches assume that the observed responses are driven by endogenous and random fluctuations. Furthermore, DCM can accommodate non-linear systems. These features make of DCM a more biologically realistic model, but also more dependent on biological constraints.

2.1.1 The forward model

An evoked potential is a perturbation in the spontaneous electroencephalographic activity resulting from sensory stimulation that causes subsequent activation. Activity measured with EEG mainly arises from extracellular current flow ensuing from postsynaptic potentials in synchronously activated pyramidal cells. Previous work has shown that ERP responses can be produced by perturbations in cortical networks (Jansen and Rit, 1995; David et al., 2005). DCM is a spatiotemporal model that includes an electromagnetic forward model that implicitly performs source reconstruction with Bayesian inversion. Hence, DCM provides a supplement to conventional electromagnetic forward models by positing further constraints on the neuronal dynamics in each source and how each source influences connected sources. This influence is parameterised by the connectivity parameters for forward, backward and lateral connections.

2.1.2 Hierarchical MEG/EEG neural mass models

DCMs for MEG/EEG use neural mass models (David and Friston, 2003) to explain source activity in terms of the ensemble dynamics of interacting inhibitory and excitatory subpopulations of neurons, based on the model of Jansen and Rit (1995). This model emulates the activity of a source using three neural subpopulations, each assigned to one of three cortical layers; an excitatory subpopulation in the granular layer, an inhibitory subpopulation in the supra-granular layer and a population of deep pyramidal cells in the infra-granular layer. The excitatory pyramidal cells receive excitatory and inhibitory input from local interneurons (via intrinsic connections, confined to the cortical sheet), and send excitatory outputs to remote cortical areas via extrinsic connections. David et al. (2005) describe a hierarchical model using extrinsic connections among multiple sources that conform to the connectivity rules described in Felleman and Van Essen (1991). These rules allow one to build a network of coupled sources linked by extrinsic connections. Within this model, bottom-up or forward connections originate in the infra-granular layers and terminate in the granular layer; top-down or backward connections link agranular layers and lateral connections originate in infra-granular layers and target all layers. All these extrinsic cortico-cortical connections are excitatory and are mediated through the axons of pyramidal cells. Exogenous inputs to the model have the same characteristics as forward connections and deliver exogenous (*i.e.*, sensory) input, u, to the granular layer (see **section 1.1.1** in **Chapter 1**, for details on connection rules).

The DCM is specified in terms of its state equations and an observation model or output equations. The set of state equations summarise the average synaptic dynamics in terms of spike-rate-dependent current and voltage changes, for each subpopulation

$$\dot{x} = f(x,u,\theta) \tag{2.1}$$

This means that the evolution of the neuronal state, x, is a function (parameterised by θ) of the state and the input u. The output equation couples the specific states (the average depolarisation of pyramidal cells in each source), x_0, to the MEG/EEG signals y using a conventional linear electromagnetic forward model.

$$y = L(\theta)x_0 + \varepsilon \tag{2.2}$$

The lead field matrix, $L(\theta)$, (i.e., the forward model) is parameterised in terms of the location and orientation of each source as described in Kiebel et al. (2006) and corresponds to the contribution of each source to the signal measured by the electrodes at the scalp level. Hence, this equation establishes the relationship between the neuronal states, x_0, and the EEG.

Equation (2.1) summarises the state equations, specifying the rate of change of the potentials as a function of the current and how currents change as a function of the currents and the potentials (see Jansen and Rit, 1995; David et al., 2005; 2006; and David and Friston, 2003 for further details). The state equations embody the connection rules described above, where θ includes extrinsic coupling parameters (forward, backward and lateral connections C^F, C^B, C^L, respectively), intrinsic coupling parameters ($\gamma_{i=1,\dots,4}$), synaptic parameters ($H_{e,i}$ and $\tau_{e,i}$), input parameters (η_1, η_2 and θ_1, θ_2) and conduction delays (Δ). **Equation (2.3)** is an expanded version of **Equation (2.1)** that shows the nine differential equations for a single source. See also **Figure 2.1** for a schematic description of how these equations are associated to the three cortical layers in a cortical source, and how these relate to forward, backward and lateral connectivity parameters in a DCM.

Figure 2.1 Schematic DCM for a single source. Each source is modelled with three subpopulations: inhibitory interneurons and pyramidal cells that correspond to the supra- and infra-granular layers and the spiny stellate cells that correspond to the granular layer. This summarises the state equations that are assigned to each subpopulation and the connection rules referred above. (Adapted from David et al., 2006)

These state equations are first-order differential equations and are derived from the behaviour of the three neuronal subpopulations, which operate as damped linear oscillations:

$$
\begin{aligned}
\dot{x}_7 &= x_8 \\
\dot{x}_8 &= \frac{H_e}{\tau_e}\left(\left(C^B + C^L + \gamma_3 I\right)S(x_0)\right) - \frac{2x_8}{\tau_e} - \frac{x_7}{\tau_e^2} \\
\dot{x}_1 &= x_4 \\
\dot{x}_4 &= \frac{H_e}{\tau_e}\left(\left(C^F + C^L + \gamma_1 I\right)S(x_0) + C^U u\right) - \frac{2x_4}{\tau_e} - \frac{x_1}{\tau_e^2} \\
\dot{x}_0 &= x_5 - x_6 \\
\dot{x}_2 &= x_5 \\
\dot{x}_5 &= \frac{H_e}{\tau_e}\left(\left(C^B + C^L\right)S(x_0) + \gamma_2 S(x_1)\right) - \frac{2x_5}{\tau_e} - \frac{x_2}{\tau_e^2} \\
\dot{x}_3 &= x_6 \\
\dot{x}_6 &= \frac{H_i}{\tau_i}\gamma_4 S(x_7) - \frac{2x_6}{\tau_i} - \frac{x_3}{\tau_i^2}
\end{aligned}
\tag{2.3}
$$

where $x_{i=1,\ldots,8}$ are the mean transmembrane potentials and currents of the three populations within a given source. In accordance with the connection rules described above, it can be seen that the state dynamics are mediated by lateral connectivity in all three layers of the single source model (see $\dot{x}_8$, $\dot{x}_5$ and $\dot{x}_4$) whereas backward connections mediate the state dynamics in the agranular layers (see $\dot{x}_8$ and $\dot{x}_5$) and forward connections mediate state dynamics in the granular layer (see $\dot{x}_4$).

The integration of this dynamic model to form predicted responses, as defined by the differential equations pertaining to each subpopulation, rests on two operators (**Equations 2.4** and **2.5**). This integration can be expressed as a convolution $p(t)$ that transforms the average density of its pre-synaptic inputs into an average postsynaptic membrane potential (David and Friston, 2003). The convolution kernel is given by:

$$
p(t) = \begin{cases} \dfrac{H_{e,i}}{\tau_{e,i}} t \exp\left(-t/\tau_{e,i}\right) & t \geq 0 \\ 0 & t < 0 \end{cases}
\tag{2.4}
$$

where H_e and H_i control the maximum postsynaptic potentials and τ_e and τ_i represent the rate constants for excitatory and inhibitory synapses, respectively. The remaining operator, S, transforms the average postsynaptic membrane potential of each subpopulation, x, into firing rate, which is the input to other subpopulations. This operator is assumed to be an instantaneous sigmoid nonlinearity.

$$S(x) = \frac{1}{1+\exp(-rx)} - \frac{1}{2} \tag{2.5}$$

where r is a fixed parameter that controls its curvature ($r = 0.56$). The intrinsic coupling parameters, γ, and the intrinsic conduction delays, Δ_{ii}, are fixed parameters. All the others are free parameters that have a prior Gaussian distribution, specified with a mean and variance. All parameters have weakly informative priors. Amongst the extrinsic coupling parameters, the prior expectations on the forward connections are weaker than the priors on the backward connections, which are in their turn weaker than the lateral connections. This accounts for the fact that forward connections exert stronger effects than backward and lateral connections.

The input that drives the system is fed through the input connections C^U which are conceptually equivalent to a forward connection that delivers the input, u, to the granular layer. This input function models the afferent activity relayed by subcortical structures and has the form of a burst with small fluctuations:

$$u(t) = \eta_2^{\eta_1} t^{\eta_1 - 1} \exp(-\eta_2 t)/\Gamma(\eta_1) + \sum \theta_i^c \cos(2\pi(i-1)t) \tag{2.6}$$

This function has two components. The first component is a gamma density function with shape and scale constants, η_1, η_2, which models an event-related burst of input that is delayed with respect to stimulus onset and dispersed by the subcortical synapses and axonal conduction. The second component is a discrete cosine set modelling systematic fluctuations in the input as a function of peri-stimulus time. See David et al. (2005; 2006), David and Friston (2003), Jansen and Rit (1995), and Friston et al. (2007) for further details.

It is important to emphasize that DCM, as specified here, depends on the assumptions made. In this thesis, DCM was used to model an ER that pertains to the auditory domain. However, DCM can be used, in principle, to model ER in the context of other sensory modalities; this is valid as long as the assumptions made hold true.

2.2 Bayesian Model Comparison

Inversion of a specific DCM model, m, corresponds to approximating the posterior probability on the parameters, $p(\theta \mid y, m)$. This is proportional to the probability of the data (the likelihood) conditioned upon the model and its parameters, $p(y \mid \theta, m)$, times the prior probability on the parameters, $p(\theta \mid m)$, according to the Bayes rule:

$$p(\theta \mid y, m) \propto p(y \mid \theta, m) p(\theta \mid m) \tag{2.7}$$

This approximation uses variational Bayes that is formally identical to Expectation-Maximisation (EM), as described in Friston (2002). The EM can be formulated, in analogy to statistical mechanics, as a co-ordinate descent on the free energy, F, of a system. The aim is to minimise the free energy with respect to a variational density $q(\theta)$. When the free energy is minimised $q(\theta) = p(\theta \mid y, m)$, the free energy $F = -\ln p(y \mid m)$ is the negative marginal log-likelihood or negative log-evidence. After convergence and minimisation of the free energy, the variational density is used as an approximation to the desired conditional density and the log-evidence is used for model comparison.

One often wants to compare different models and select the best before making statistical inferences on the basis of the conditional density. The best model, given the data, is the one with highest log-evidence $\ln p(y \mid m)$ (assuming a uniform prior over models). Given two models m_1 and m_2 one can compare them by computing their Bayes factor or, equivalently, the difference in their log-evidences:

$$\ln B_{1,2} = \ln p(y|m_1) - \ln p(y|m_2) \tag{2.8}$$

If this difference is greater than about three (*i.e.*, their relative likelihood is more then 20:1) then one asserts there is strong evidence in favour of the first model (see Penny et al., 2004 for details).

The formalism described above is suitable for comparing different models of a given data set, for instance data acquired from a single subject. However, one may wish to select the model that best explains multiple data sets, *i.e.*, the best model at the group level. Assuming each data set is independent of the others (*i.e.*, all subjects are independent of each other), we can simply multiply the marginal likelihoods, or, equivalently, add the log-evidences from each subject

$$\ln p(y_1,...,y_n \mid m_i) = \sum_{j=1}^{n} \ln p(y_j \mid m_i) \tag{2.9}$$

to obtain the log-evidence for the i-th model across all n subjects. Log-evidences will be used extensively in subsequent chapters to compare competing models or hypotheses about the MMN.

Chapter 3

Dynamic Causal Modelling of Evoked Potentials: A Reproducibility Study

The previous chapter presented DCM as a novel tool for modelling and analysis of connectivity in the brain. This chapter addresses the validity of DCM by appraising its reproducibility across a multi-subject data set and shows that it can be used in realistic context for hypothesis testing. Three different connectivity models were considered to explain ERPs elicited in an oddball paradigm. Bayesian model comparison was used to select the best model within and between subjects. The rationale is that if DCM is a robust method, then the results of model comparison in one subject should predict the results in another (i.e., they should be consistent over subjects).

3.1 Introduction

In dynamical causal modelling, one views the brain as a dynamic network of interacting sources that produces observable responses. This perspective furnishes spatiotemporal, generative or forward models for evoked responses as measured with EEG/MEG (David et al., 2005; Friston et al., 2003). In brief, DCM entails specification of a plausible model of electrodynamic responses. This model is inverted by optimising a variational free-energy bound on the model-evidence to provide the conditional density of the model parameters and the model evidence for model comparison. This is an important advance over conventional analyses of evoked responses because it places natural constraints on the inversion; namely, activity in one source has to be caused by activity in another. Face-validity of DCM for ERPs has been previously established (see David et al., 2005; 2006; Kiebel et al., 2006). In this chapter, predictive-validity is addressed in terms of the reproducibility of DCM over subjects. In brief, if DCM discloses a likely underlying brain network, then the results of model comparison in one subject should predict the results in another (*i.e.*, they should be consistent over subjects). To ensure the same network was engaged in all subjects, a MMN paradigm was used. This represents one of the best studied and reproducible phenomena in ERP research (Näätänen et al., 2001).

3.1.1 Dynamic Causal Modelling

Previous work suggests that event-related responses can be modelled as perturbations of cortical networks (David et al., 2005; Jansen and Rit, 1995). In particular, it has been shown that dynamic causal models (DCMs) can explain event-related potentials (ERPs) and fields (ERFs) measured with EEG and MEG, respectively. Furthermore, David et al. (2006) showed that differences in evoked responses can be explained by changes in effective connectivity or coupling among neuronal sources. DCM can be regarded as an elaboration of conventional spatial forward models of EEG/MEG data, in which the sources are coupled according to biological constraints. The inversion of a DCM provides information about the underlying cortical pathways and their causal architecture. In this chapter, the reproducibility of DCM was assessed by testing hypotheses about cortical networks suggested by a predictive coding view of novelty or mismatch responses. As discussed in **Chapter 1**, predictive coding is a formulation of perceptual learning and inference that rests on hierarchical Bayes (Rao and Ballard, 1999; Friston, 2003). See **section 1.1** of **Chapter 1** for full description of the conceptual framework. Critically, by formulating different implementations of perceptual learning in terms of different DCMs it was possible to adjudicate among competing hypotheses using Bayesian model comparison. Moreover it was possible to assess the reproducibility of model comparison over subjects.

Bayesian model comparison (Penny et al., 2004) selects the model, among competing models, that best explains the data. Given equal prior probabilities for the models considered, they are compared using their marginal likelihood or evidence for each model. In this work, the predictive validity of DCM was assessed by testing its reproducibility in a multi-subject study, under an auditory oddball paradigm. Specifically, reproducibility of DCM was evaluated in terms of model comparison, from subject to subject. The choice of the oddball paradigm was motivated by the large body of work in this field and current interest in the mechanisms underlying the generation of the MMN; in particular, the specific hypotheses about these mechanisms furnished by theoretical treatments (Winkler et al., 1996; Friston, 2005).

3.1.2 Mismatch responses

Small changes in the acoustic environment initiate involuntary attentional capture, which may be engaged automatically by an auditory change detection mechanism indexed by the MMN. It has been proposed that a temporo-prefrontal network generates the MMN by comparing sensory input with the memory trace of previous stimuli. Novel events (oddballs) embedded in a stream of repeated or familiar events (standards) produce a distinct response that can be recorded non-invasively with electrophysiological techniques such as EEG and MEG. The MMN is the negative component of the waveform obtained by subtracting the event-related response to a standard from the response to an oddball, also called a deviant in the literature. This pre-attentive response, to sudden changes in stimulation, peaks at about 100-250 ms from change onset; with the greatest intensity located over the frontal region (usually reported for channels Fz and FCz). Given its automatic nature, the MMN has been associated with pre-attentive cognitive operations in audition and has, for this reason, been proposed as a reflection of 'primitive intelligence' in the auditory cortex (Näätänen et al., 2001).

As discussed in **Chapter 1 (section 1.2)**, there have been several compelling mechanistic accounts of how the MMN might arise, many focussing on plasticity or adaptation (e.g. May et al., 1999). Winkler et al. (1996) suggested that the MMN might arise from a model-adjustment process, whereby the auditory system adjusts its perceptual model to adapt to the stimuli encountered. A similar conclusion was reached from a theoretical treatment of perceptual inference and learning based on hierarchical Bayes (Friston, 2005). In this framework, evoked responses correspond to prediction error. Changes in synaptic connections, during the repeated presentation of standards may render suppression of prediction error more efficient. This would lead to a reduction in evoked responses and the emergence of a mismatch response, when an unlearned stimulus is presented. This differential response is mediated by differences in effective connectivity or coupling. In short, mechanistic theories about the MMN posit changes in synaptic efficacy between the cortical levels of a hierarchical setting. In what follows, model comparison is used to assess the reproducibility across subjects in terms of DCMs that best explains differences in ERPs to frequent (standards) and rare events (oddballs). Three models were tested to ask whether mismatch responses are better explained by changes in forward connections, backward connections, or both, and whether the conclusions generalise over subjects, whose data were acquired independently.

3.2 Methods and Statistical analysis

This section describes the methodology used for modelling the ERPs and the statistical analysis performed which enabled inferences on: *what is the best model*; and *what are the model parameters* at the single subject and group levels.

3.2.1 Experimental design

3.2.1.1 Subjects

This study involved a group of thirteen healthy volunteers aged 24–35 (5 female). Each subject gave signed informed consent before the study, which proceeded under local ethical committee guidelines.

3.2.1.2 Task

Subjects sat on a comfortable chair in front of a desk in a dimly illuminated room. Electroencephalographic activity was measured during an auditory 'oddball' paradigm, in which subjects were presented with "standard" (1000 Hz) and "deviant" tones (2000 Hz), occurring 80% (480 trials) and 20% (120 trials) of the time, respectively, in a pseudo-random sequence. The stimuli were presented binaurally via headphones for 15 minutes every 2 seconds. The duration of each tone was 70 ms with 5 ms rise and fall times. The subjects were instructed not to move, to keep their eyes closed and to count the deviant tones.

3.2.2 Data acquisition and pre-processing

EEG was recorded with a *Biosemi* system with 128 scalp electrodes. Data were recorded at a sampling rate of 512 Hz with a fixed 1st order analogue anti aliasing filter of -3dB at 3.6 kHz. Vertical and horizontal eye movements were monitored using three EOG (electro-oculogram) electrodes. The horizontal eye movements were monitored by two EOGs placed on the inner and outer corners of the subjects' left eyebrow. The vertical eye movements were monitored by the latter EOG and another one placed about 4 to 5 cm below. The pre-processing and data analysis steps were performed offline with SPM5 (http://www.fil.ion.ucl.ac.uk/spm/) in the following order: epoching with a peri-stimulus window of -100 to 400 ms, down-sampling to 200 Hz, band-pass filtering (Butterworth) between 0.5 – 40 Hz and re-referencing to the average of the right and left ear lobes. A 40 Hz cut-off was used in order to clean the data from high frequency noise (including the 50 Hz in the electric power). The use of 40 Hz does not compromise the components of interest as their frequency is well below this threshold. Ideally, filtering should have been the first of the pre-processing steps so that aliasing was minimized. Here, however, this has been comprised for the sake of computational expediency. Trials in which the absolute amplitude of the signal exceeded $100\,\mu V$ were excluded. Two subjects were eliminated from further analysis due to excessive trials containing artefacts. In the remaining subjects, an average 18% of trials were excluded. For computational expediency the dimensionality of the data was reduced to three channel mixtures or spatial modes. These were the principal modes of a singular value decomposition of the channel data between 0 and 250 ms, from both trial types. The use of three principal eigenvariates preserved more than 69% of the variance in all subjects (see **Figure 3.1**).

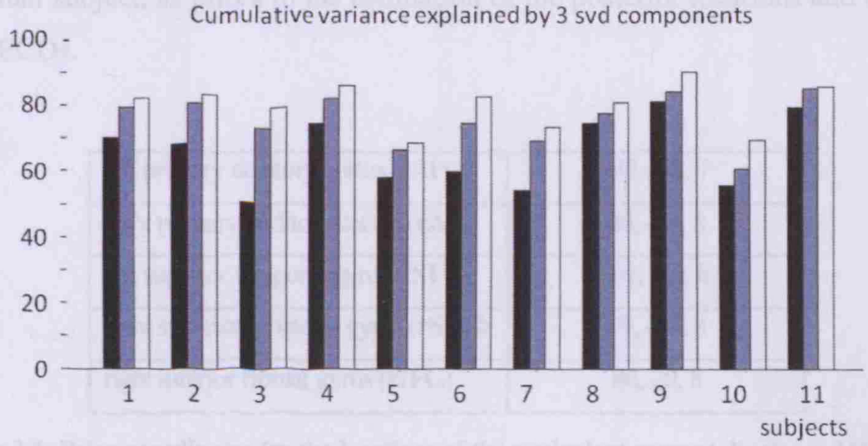
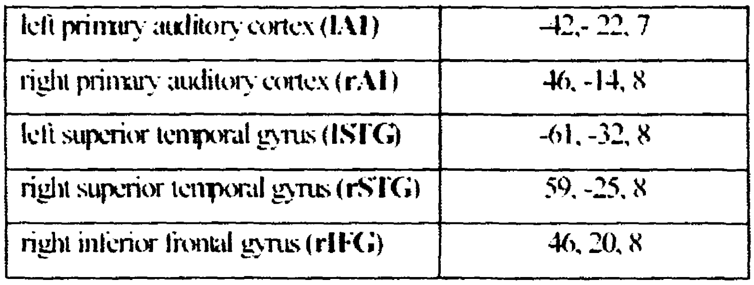


Figure 3.1 Cumulative variance explained by 3 svd components.

3.2.3 DCM specification

This section describes three plausible models defined by a given architecture and dynamics, which were used here to test the validity of DCM. See **Chapter 2** for a full description of DCM.

The network architecture was motivated by recent electrophysiological and neuroimaging studies looking at the sources underlying the MMN (Opitz et al., 2002; Doeller et al., 2003). Five sources were modelled as ECDs, over left and right **A1**, left and right **STG** and right **IFG**. The mechanistic model attempts to explain the generation of each individual response (i.e., responses to standards and responses to deviants). Therefore, left and right **A1** were chosen as cortical input stations for processing auditory information. The subcortical input function was parameterised as a mixture of a Gaussian bump function (with unknown latency and dispersion) and a background alpha rhythm (with unknown phase and amplitude), that attempted to model residual alpha ringing. Opitz et al. (2002) identified sources for the differential response, with fMRI and EEG measures, in both left and right **STG,** and right **IFG.** Here the coordinates reported by Opitz et al., (2002) (for left and right **STG** and right **IFG**) and Rademacher et al. (2001) (for left and right **A1**) are employed as prior source location means, with a prior variance of 32 mm^2. These coordinates, given in the literature in Talairach space, were converted to MNI space using the algorithm described in (http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach). The moment parameters had prior mean of 0 and a variance of 8 in each direction. These parameters were used, for each individual subject, as priors in the estimation of the posterior locations and moments of the ECDs.



left primary auditory cortex (**lA1**)	-42,- 22, 7
right primary auditory cortex (**rA1**)	46, -14, 8
left superior temporal gyrus (**lSTG**)	-61, -32, 8
right superior temporal gyrus (**rSTG**)	59, -25, 8
right inferior frontal gyrus (**rIFG**)	46, 20, 8

Table 3.1: Prior coordinates for the locations of the equivalent current dipoles in MNI space (mm).

Using these sources and prior knowledge about the functional anatomy the following DCM was constructed: An extrinsic input entered bilaterally to **A1**, which were connected to their ipsilateral **STG**. Right **STG** was connected with the right **IFG**. Inter-hemispheric (lateral) connections were placed between left and right **STG**. All connections were reciprocal (i.e., connected with forward and backward connections or with bilateral connections).

Given this connectivity graph, specified in terms of its nodes and connections, three models were tested. These models differed in the connections which could show putative learning-related changes, i.e., differences between listening to standard or deviant tones. Models **F**, **B** and **FB** allowed changes in forward, backward and both forward and backward connections, respectively (**Figure 3.2**). All three models were compared against a baseline or null model. The **null** model had the same architecture described above but precluded any coupling changes between standard and deviant trials.

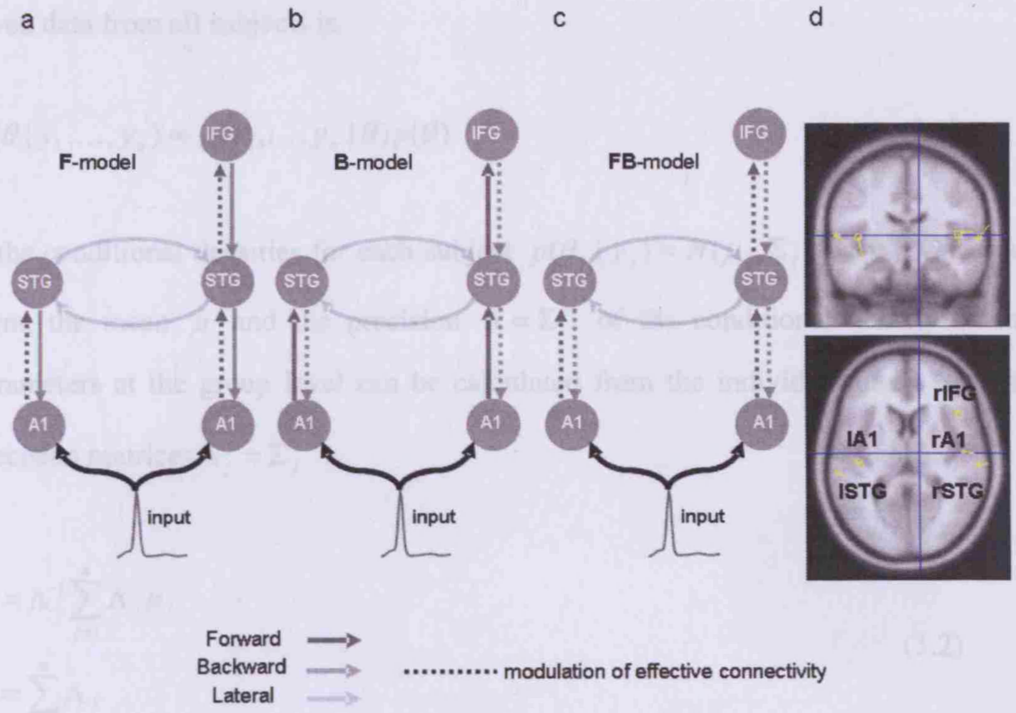


Figure 3.2 Model specification. The sources comprising the network are connected with forward (dark grey), backward (grey) or lateral (light grey) connections as shown. **A1**: primary auditory cortex, **STG**: superior temporal gyrus, **IFG**: inferior temporal gyrus. Three different models were tested within the same architecture **(a-c)**, allowing for learning-related changes in forward **F**, backward **B** and forward and backward **FB** connections, respectively. The broken lines indicate the connections that were allowed to change. Sources of activity, modelled as dipoles (estimated posterior moments and locations), are superimposed in an MRI of a standard brain in MNI space **(d)**.

3.2.4 Statistical analysis: Bayesian model comparison

The statistical analysis in this chapter uses the procedures described in **Chapter 2**, for Bayesian model comparison at both single-subject and group levels. See **section 2.2** for further details. The next subsection describes the inference made on the parameters of the best model pooled over subjects, i.e., at the group level.

3.2.4.1 Parameter estimation at the group level

Bayes' theorem updates our belief about a parameter in the light of new evidence from the data. Bayesian inference can be particularly useful at the second (between-subject) level of statistical analysis (Neumann and Lohmann, 2003). In particular, it is easy to combine the conditional densities from several subjects to obtain a single conditional density for the group. The conditional probability of the parameters given data from all subjects is:

$$p(\theta \mid y_1,\dots,y_n) \propto p(y_1,\dots,y_n \mid \theta)p(\theta) \tag{3.1}$$

If the conditional densities for each subject $p(\theta_j \mid y_j) = N(\mu_j, \Sigma_j)$ have a Gaussian form, the mean μ and the precision $\Lambda = \Sigma^{-1}$ of the conditional density of the parameters at the group level can be calculated from the individual mean μ_j and precision matrices $\Lambda_j = \Sigma_j^{-1}$.

$$\begin{aligned} \mu &= \Lambda^{-1}\sum_{j=1}^{n}\Lambda_j\mu_j \\ \Lambda &= \sum_{j=1}^{n}\Lambda_j \end{aligned} \tag{3.2}$$

This provides a useful way to summarise the results of several DCMs from different subjects.

3.3 Results

3.3.1 Event-related potentials

The difference between the ERPs evoked by the standard and deviant tones revealed a standard MMN (**Figure 3.3**). This negativity was present from 90–190 ms (**Figure 3.3a-c**) and had a broad spatial pattern, encompassing electrodes previously associated with auditory and frontal areas (**Figure 3.3d**).

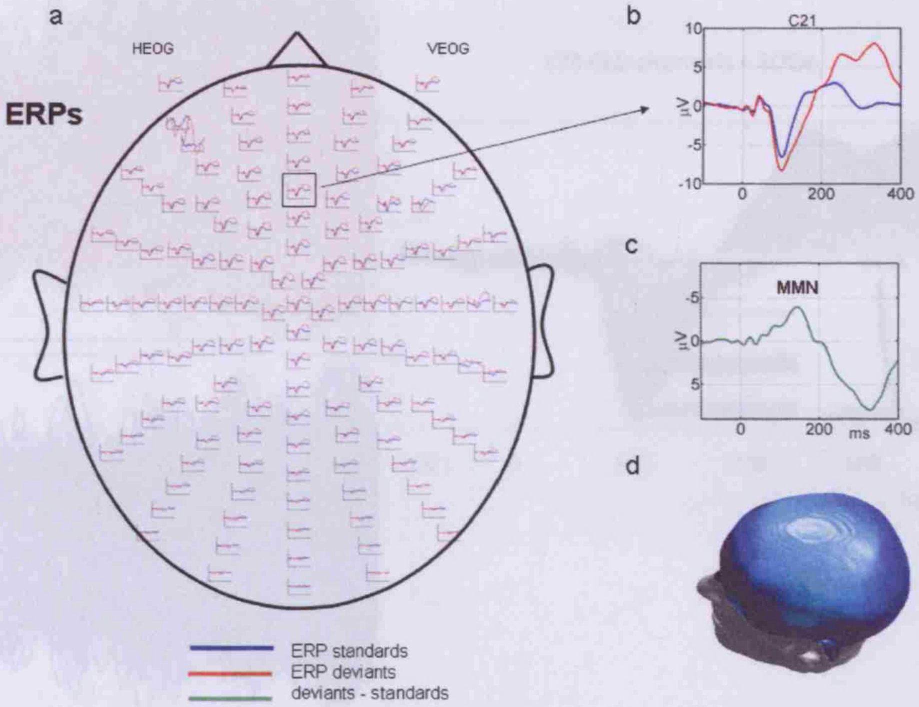


Figure 3.3 Grand mean ERPs, i.e., averaged over all subjects. **(a)** ERP responses to the standard and deviant tones overlaid on a whole scalp map of 128 EEG electrodes. **(b)** ERP responses to the standard and deviant tones at channel C21 (fronto-central). **(c)** MMN, the difference wave obtained by subtracting the grand-average ERP to standards from the ERP to deviants, at channel C21. **(d)** grand mean MMN response averaged across subjects and over a time window of [100, 200] ms interpolated to give a 3D scalp topography.

Figure 3.3a shows that the MMN is also present on the EOG channels, which is due to contamination by the brain evoked responses. This is demonstrated in **Figure 3.4,** which shows that there is no eye movement artefact in the VEOG and the HEOG that could be causing an apparent MMN. Moreover, this figure also shows the gradient of the peak oddball response (at about 100 ms) from the back to the front electrodes, as well as the EOG signals. It can be seen that the peak amplitude increases from the back to the central electrodes, and then decreases from there to the EOGs. This reveals how that the activity on the HEOG and VEOG channels is as a consequence of the MMN and not visa versa.

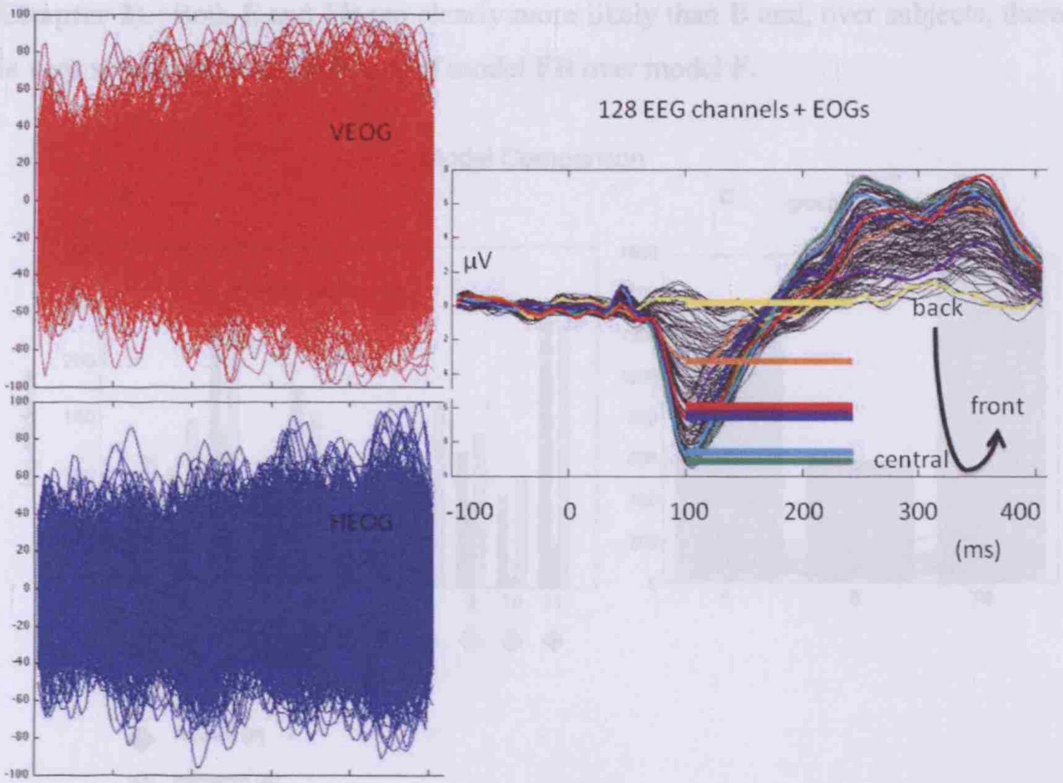


Figure 3.4 Contamination of the VEOG and HEOG by the brain evoked responses. (Left) VEOG and HEOG signals from all subjects and all trials. (Right) Grand mean oddball responses for from the 128 EEG, the VEOG, and the HEOG channels.

3.3.2 Model selection

Four different DCMs, forward only (**F**-model), backward only (**B**-model), forward and backward (**FB**-model) and the **null** were inverted for each subject. **Figure 3.6** illustrates the model comparison based on the increase in log-evidence over the **null** model, for all subjects. **Figure 3.5a** shows the log-evidence for the three models, relative to the null model, for each subject, revealing that the three models were significantly better than the null in all subjects. The diamond attributed to each subject identifies the best model on the basis of the highest log-evidence. The **FB**-model was significantly better in seven out of eleven subjects. The **F**-model was better in four subjects but only significantly so in three (for one of these subjects [subject 6], model comparison revealed only weak evidence in favour of the **F**-model over the **FB**-model, though still very strong evidence over the **B**-model. In all but one subject, the **F** and **FB**-models were better than the **B**-model. **Figure 3.5b** shows the log-evidences for the three models at the group level (i.e., using. **Eq. 2.9** – see

Chapter 2). Both **F** and **FB** are clearly more likely than **B** and, over subjects, there is very strong evidence in favour of model **FB** over model **F**.

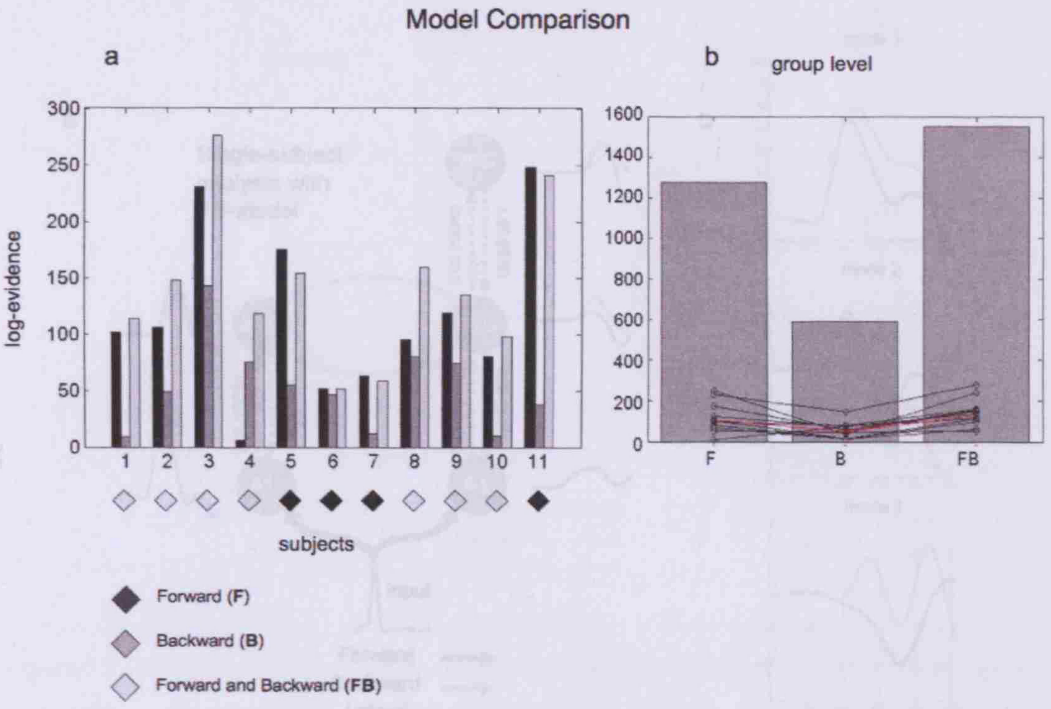


Figure 3.5 Bayesian model selection among DCMs for the three models, F, B and FB expressed relative to a DCM in which no connections were allowed to change (null model). The graphs show the free energy approximation to the log-evidence. **(a)** Log-evidence for models **F**, **B** and **FB** for each subject (relative to the null model). The diamond attributed to each subject identifies the best model on the basis of the subject's highest log-evidence. **(b)** Log-evidence at the group level, *i.e.*, pooled (summed) over subjects, for the three models. The red line represents the averaged log-evidence (across subjects) and the black points represent the log-evidence for each subject. This shows that there is not an outlier subject with a particularly strong effect driving the group log evidence.

Figure 3.6a shows, for the best model **FB**, the predicted responses at each node of the network for each trial type (*i.e.*, standard or deviant) for a single subject (subject 9). The coupling gains and the conditional probability of the gains being greater or smaller than one are shown for each connection in the network. The values on each connection represent a scaling effect, for example, a coupling change of 2.04 from **lA1** to **lSTG** means that the effective connectivity increased by 104% for rare events relative to frequent events. The response, in measurement space, of the three principal spatial modes is shown on the right (**Figure 3.6b**). This figure shows a remarkable agreement between predicted (solid) and observed (dotted) responses.

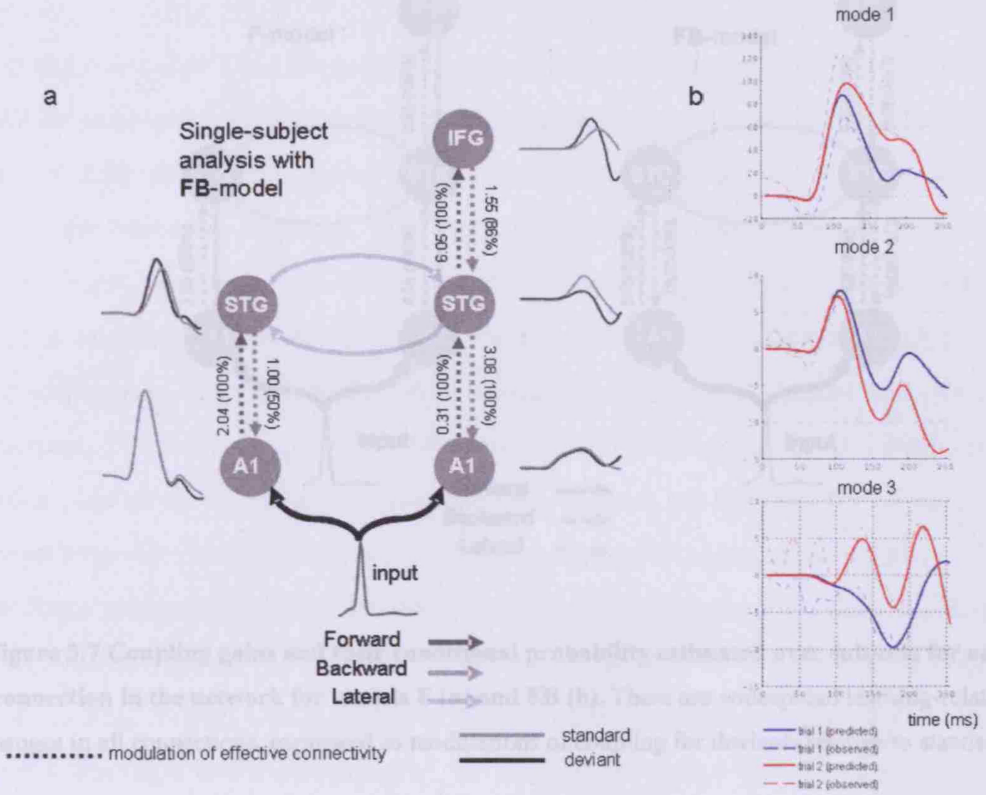


Figure 3.6 DCM results for a single subject [subject 9] (FB model). (a) Reconstructed responses for each source and changes in coupling adjacent to the connections during oddball processing relative to standards. The mismatch response is expressed in nearly every source. **(b)** Predicted (solid) and observed (broken) responses in measurement space, which result from a projection of the scalp data onto their first three spatial modes.

Figure 3.7 summarises the conditional densities of the coupling parameters for the F-model (**Figure 3.7a**) and **FB**-model (**Figure 3.7b**). The coupling gains and the conditional probability of the gains being greater or smaller than one, pooled over subjects, are shown for each connection in the network (*i.e.,* using **Eqs. 3.1** and **3.2**). For the **F**-model the effective connectivity has increased in all connections with a conditional probability of 100%. For the **FB**-model the effective connectivity has changed in all forward and backward connections with a probability of 100%. Equivalently, and in accord with theoretical predictions, all extrinsic connections (*i.e.*, influences) were modulated for rare events as compared to frequent events.

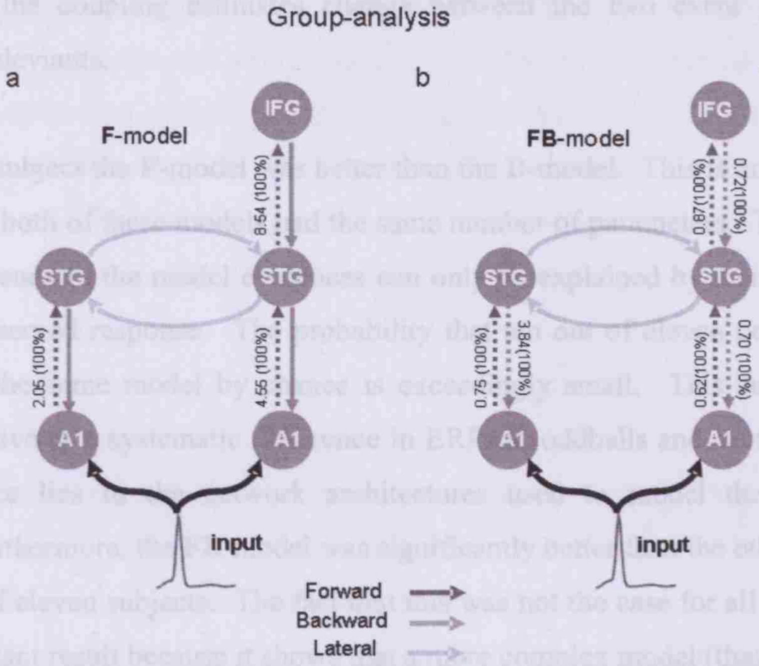


Figure 3.7 Coupling gains and their conditional probability estimated over subjects for each connection in the network for models F (a) and FB (b). There are widespread learning-related changes in all connections, expressed as modulations of coupling for deviants relative to standards.

3.4 Discussion

This study evaluated the predictive validity of DCM by looking at its reproducibility over subjects in the context of the MMN. To this end DCM was used to explain differences in ERPs in terms of changes in effective connectivity. Three models of connectivity were tested, for each subject, within the same underlying cortical architecture but differing in modulations of specific types of connections. There was a very reproducible pattern of results across subjects. Model comparison revealed that conjoint changes in forward and backward connections (**FB**-model), relative to changes in forward (**F**-model) or backward connections alone (**B**-model) are consistently better across subjects. In every subject, these models were better than a **null** model that precluded any coupling changes. The evidences for models **F** and **FB** were, overall, much bigger than the model evidence for **B.** A similar consistency was observed quantitatively, in terms of the conditional densities of each connection; as predicted, the coupling estimates change between the two event types, i.e., standards and deviants.

In all but one subject the **F**-model was better than the **B**-model. This is an important result because both of these models had the same number of parameters. This means that any difference in the model evidences can only be explained by their ability to predict the observed response. The probability that ten out of eleven comparisons would select the same model by chance is exceedingly small. This suggests the DCM is sensitive to a systematic difference in ERPs to oddballs and standards; and their difference lies in the network architectures used to model the observed responses. Furthermore, the **FB**-model was significantly better than the other models in seven out of eleven subjects. The fact that this was not the case for all subjects is another important result because it shows that a more complex model (that can fit the data more accurately) is not necessarily the most likely model. This slight inter-subject variability for the winning model might be due to differences in subjects' attention state, which would probably be expressed in the importance that backward connections have in the dynamics of the cortical network. For example, model FB would presumably be better explaining the data recorded from a more attentive subject. Another possibility is that different subjects use different strategies for counting oddball events, which would be also expressed in the relative importance of forward and backward connections in the model.

3.4.1 Choice of paradigm

This paradigm was deliberately chosen such that it evoked two responses exhibiting a large difference over peri-stimulus time. Note that this is not a classical oddball paradigm, as employed in the MMN literature. The large pitch difference between responses to standards and deviants elicits large differences for both N1 and MMN components, which cannot be disentangled. In this sense, this paradigm is not necessarily appropriate for modelling the MMN alone, but suitable for assessing the reproducibility of DCM: DCM is a model of dynamic responses or transients that are continuous in time. This means that the DCM is not an explanation for a particular response component (e.g., the MMN) but the compound response over all peri-stimulus times; it is likely that this paradigm induced N1 effects (due to the large different in standard and oddball tones) and, at least phenomenologically, a P300-like component (although the analysis presented here only went up to 250ms). However, all these components could be explained by differences in a simple network model of interacting neuronal populations. It would be interesting to assess the ontological status of response components (e.g., N1, MMN, and P300) in the light of mechanistic models like DCM (see below).

3.4.2 Choice of model

DCM is not an exploratory technique; it does not explore all possible models: DCM tests specific models of connectivity and, through model selection, can provide evidence in favour of one model relative to others. The results of a DCM analysis depend explicitly upon the models evaluated, which are generally motivated by mechanistic hypotheses. This means there may exist other equally or more plausible models with different architectures (in respect of the areas and connections involved). The network chosen for the DCMs was motivated by the results of previous MMN studies (Opitz et al., 2002; Doeller et al., 2003). It has been shown that the generators of the MMN lie bilaterally on the temporal cortex. In addition, some studies have shown bilateral generators on the frontal cortex, which are activated later than the auditory cortex generators (Rinne et al., 2000). Recent studies showed a double peak over frontal scalp locations suggesting the existence of two sub-components for the MMN (Opitz et al., 2002 and Doeller et al., 2003). The early component is reported to peak around 90 – 120 ms and can be modelled with sources located bilaterally in the STG. Components peaking around 140 – 170 ms have been shown to be modelled effectively with dipoles in both left and right IFG that are usually stronger in the right hemisphere. Moreover, the right IFG is reported more consistently as in the literature than the homologous source on the left. This is why a unilateral right IFG has been used. The inclusion of both left and right primary auditory cortices was necessary because DCM attempts to explain each ERP individually and any differences (in this case, responses to standards and deviants and the implicit mismatch). Therefore, left and right **A1** were chosen as the cortical targets of thalamic input, for processing auditory information.

Indeed there may be other plausible models. *So, why not include more sources in the model, for example left IFG?* This is an important question that has a principled answer. Each model is defined by its number of sources and their connectivity. Given a particular model, DCM will optimize the parameters of that model. To explore the space of models, one uses the marginal likelihood or log-evidence to compare one model with another. This enables one to adjudicate between different models with different sources. This is potentially an important application of DCM in the context of model selection. This chapter focuses on the reproducibility of different connectivity architectures. In the following chapters an exploration of model space will be presented (see **Chapters 5 and 6**).

3.4.3 Mechanisms of ERP generation

As detailed in **Chapter 1**, the mechanism of MMN generation is unknown. Traditionally, this response is thought to be associated with an automatic cortical change-detection process, which detects a difference between the current and the preceding input. It has been proposed that the MMN would reflect modifications to existing parts of a model of the acoustic environment. This model adjustment hypothesis discusses the existence of a dynamic system of change detection which updates its model of sensory input as the changes occur (Winkler et al. 1996; Sussman and Winkler, 2001). Doeller et al. (2003) suggested that the prefrontal cortex is involved in a top-down modulation of the deviance detection system in the temporal cortices. The traditional interpretation has been criticised by Jääskeläinen et al. (2004) who proposed that the MMN results from an *adaptation* mechanism (see also May et al., 1999) and is erroneously interpreted as a separate component generated by change-specific neurons. The N1 response to standard (or 'non-novel') sounds is thought to be delayed and suppressed (or *attenuated*) as a function of its similarity to the preceding auditory events, reflecting short-lived adaptation of auditory cortex neurons. Under this hypothesis, the response termed as MMN, would therefore be a product of an N1 differential wave emerging in the subtraction of the standards from the deviant's ERP. Recently, the MMN has been framed within a predictive coding scheme; where it is interpreted as a failure to suppress prediction error, which can be explained quantitatively in terms of coupling changes among cortical regions (Friston, 2005). This predicts the adjustment of a generative model of current stimulus trains (*cf.*, the model adjustment hypothesis) using plastic changes in synaptic connections (*cf.* the adaptation hypothesis). This point will be discussed in more depth in **Chapter 5**.

3.4.4 Consistency over model parameters

How does one reconcile different results from different DCMs? This chapter looked at the reproducibility of DCMs in terms of consistency over subjects in model space. Although not the focus of this study, it is also interesting to address the consistency of model parameters. Inference on the parameters of a particular DCM is conditional upon the specific model and data used. Clearly, one could get very different conclusions by changing the model (e.g., **Figure 3.7**). One may be very confident about contradictory results from two different models. This contradiction is resolved by considering the relative likelihoods of the two models. Generally, one only considers inference on the parameters of the best model. It is possible to use Bayesian model averaging; however, this is most useful when the models have roughly the same probability. If the results change under the *same* model with *different* data sets, then this speaks to inter-subject variability that may be real. Clearly, it would be re-assuring to see that the parameter estimates from each model were also consistent over subjects. This can be addressed with a second, between-subject analysis of the parameter estimates. The consistency of DCM inversion in terms of model parameters can be illustrated with a classical multivariate statistical analysis (MANOVA) of subject-specific parameter estimates, from the best model (**FB**); DCM parameters were summarised using the average coupling changes for the forward and backward connections. Using just two dependent summary variables ensured sufficient (*i.e.*, nine) degrees of freedom to infer the group averages were significantly different from zero. The resulting F-value (based on Wilk's lambda; d.f. 2,9) was 4.725 corresponding to a *p*-value of 0.04. This implies a systematic and consistent pattern of changes over forward and backward connections, over subjects. The average change in forward connections was 80% and the average change in backward connections was 57%. This is consistent with the average change in forward and backward connections for the DCM analysis of grand-mean data (42% and 52% respectively). A more informed way of addressing this issue would involve a hierarchical Bayesian model that includes random effects from each subject. This is a potential direction for future work (see **Chapter 7**).

3.4.5 Technical issues

This section attempts to clarify few technical issues that are relevant to both this chapter and the experimental chapters that follow.

What is the relationship between "inverting" a DCM and inverting a classical electromagnetic forward model to locate intra-cranial source locations?
Model inversion is used in exactly the same way in both contexts. In fact, inverting a DCM subsumes the inversion of a conventional forward model. This is because DCM has two components; a neural-mass model of the interactions among a small number of dipole sources and a classical electromagnetic forward model that links these sources to extra-cranial measurements. Inverting the DCM implicitly optimizes the location and orientation of the sources. Indeed, if the neuronal part of the model is removed, DCM would reduce to a conventional forward model. The current implementation of DCM for ERPs uses the electromagnetic forward model solutions encoded in the fieldtrip software (http://www2.ru.nl/fcdonders/fieldtrip).

Informative priors have been used on source locations. H*ow important are prior source locations in DCM?* Not very important; in other words, changing the location priors would not change the results very much. This is because there is very little information in the EEG or MEG measurements about the spatial location of sources. In contrast, there is an enormous amount of information about their orientation. A usual heuristic is to imagine that one is trying to infer the location and orientation of a pen-torch that is illuminating the inside of a balloon. When looking at the balloon from outside, a small change in the position of the torch will have only a small effect on the pattern of illumination. Conversely, changing its orientation will have a large impact. In brief, location is not an important quantity, whereas orientation is. For this reason, relatively informative priors have been used on the location whereas the orientation parameters were left free. As noted above, it would be relatively straightforward to assess the impact of different location priors using model selection.

What does DCM actually operate on, the channel data or some estimate of source dynamics? DCM operates on the channel data in the same way as sources are reconstructed in a conventional setting. In both cases, one needs to specify the number of sources and how their activity is expressed in sensor space. The only difference is that, with DCM, one also has to specify the connections among the sources and which sources receive sensorimotor input.

How is the significance of the difference between models evaluated at the between-subject level?

For a single subject, a difference in log-evidence of about three is considered strong evidence in favour of the more likely model. This is because a difference of three means the evidence for the more likely model is about twenty times the evidence for the other. This can be compared to the use of $p = 0.05 = 1/20$ used in classical inference. When pooling log-evidences from several subjects, one adds the evidences and, implicitly, the differences. This is because adding logs is the same as multiplying probabilities and the probability of getting two independent sets of data is simply the product of the probabilities of getting each alone. It can be seen that one will quickly reach a threshold of three if, and only if, the difference in log-evidences is consistent over subjects.

3.5 Conclusion

This chapter studied possible mechanisms underlying the generation of the MMN at the level of coupling among sources, and the extent to which the ensuing inferences generalise over subjects. Here the predictive validity of DCM has been established in the context of testing mechanistic hypotheses underlying a specific brain response. The model comparison addressed hierarchical implementations of predictive coding, in terms of extrinsic forward and backward connections. Testing the adaptation hypothesis would require modulations of intrinsic connections (i.e. among neural subpopulations within cortical units). This issue will be addressed in **Chapter 5** where the relative roles of extrinsic and intrinsic connectivity changes are tested (see also Kiebel et al., 2007 for details of the methodological advances).

In brief, this study suggests that the emergence of the MMN can be explained by repetition-dependent changes in forward and backward connections. This was found consistently across the group of subjects studied. Changes in forward connections might be associated with the construction of higher-level representations of sensory data. This is consistent with the conjecture that the MMN reflects perceptual learning of standards, using predictive coding. This presents the first attempt to invert a mechanistic model for the MMN using empirical data, and demonstrates the robustness of the method in group studies.

3.6 Summary

The aim of this work was to establish the predictive validity of DCM by assessing its reproducibility across subjects. An oddball paradigm was used to elicit mismatch responses. Sources of cortical activity were modelled as equivalent current dipoles, using a biophysical informed spatiotemporal forward model that included connections among neuronal subpopulations in each source. Bayesian inversion provided estimates of changes in coupling among sources and the marginal likelihood of each model. By specifying different connectivity models it was possible to evaluate three different hypotheses: differences in the ERPs to rare and frequent events are mediated by changes in forward connections (**F**-model), backward connections (**B**-model) or both (**FB**-model).

The results were remarkably consistent over subjects. In all but one subject, the forward model (**F**-model) was better than the backward model (**B**-model). This is an important result because these models have the same number of parameters (i.e., the complexity). Furthermore, the **FB**-model was significantly better than both, in seven out of eleven subjects. The fact that this was not the case for all subjects shows that a more complex model (that can fit the data more accurately) is not necessarily the most likely model. At the group level the **FB**-model supervened. These findings demonstrate the validity and usefulness of DCM in characterising EEG/MEG data and its ability to model ERPs in a mechanistic fashion.

Chapter 4

Evoked brain responses are generated by feedback loops

The previous chapter established the predictive validity of DCM. Importantly, it demonstrated that DCM can be used to answer interesting questions about cortical organization and function. This chapter presents the first application of DCM to ERPs. The aim of this study was to investigate the role of backward connections on the generation of long-latency ERP responses. Neuronal responses to stimuli, measured electrophysiologically, unfold over several hundred milliseconds. Typically, they show characteristic waveforms with early and late components. It is thought that early or exogenous components reflect a perturbation of neuronal dynamics by sensory inputs – bottom-up processing. Conversely, later, endogenous components have been ascribed to recurrent dynamics among hierarchically disposed cortical processing levels – top-down effects. The work reported in this chapter tests these theoretical formulations by comparing DCMs with and without backward connections. Furthermore, the contribution of backward connections to evoked responses was quantified as a function of peri-stimulus time.

CHAPTER 4. EVOKED BRAIN RESPONSES ARE GENERATED BY FEEDBACK LOOPS

4.1 Introduction

Event-related potentials or fields in electroencephalography and magneto-encephalography are one of the mainstays of non-invasive neuroscience. Typically, the response evoked by a stimulus evolves in a systematic way showing a series of waves or components. Many of these components are elicited so reliably that they are studied in their own right. These include very early sensory-evoked potentials, observed within a few milliseconds, early cortical responses such as the N1 and P2 components and later components expressed several hundred milliseconds afterwards. Broadly speaking, ERP components can be divided into early and late (Syndulko et al., 1982). Early or short-latency stimulus-dependent (exogenous) components reflect the integrity of primary afferent pathways. Late stimulus-independent (endogenous) components entail long-latency (more than 100ms) responses that are thought to reflect cognitive processes (Syndulko et al., 1982; Gaillard et al., 1988). Early components have been associated with exogenous bottom-up stimulus-bound effects, whereas late components have been ascribed to endogenous dynamics involving top-down processes such as attention. Indeed, the amplitude and latency of early (e.g., P1 and N1) and late (e.g., N2pc) components have been used as explicit indices of bottom-up and top-down processing respectively (Schiff et al., 2006). This chapter demonstrates that late components are mediated by recurrent interactions among remote cortical regions; specifically it is shown that late components rest upon backward extrinsic cortico-cortical connections that enable recurrent or reentrant dynamics.

Other possible mechanisms for ERP generation have been suggested such as phase-resetting (Makeig et al., 2002) and baseline shift associated with alpha oscillations (Nikulin et al., 2007). Makeig et al. (2002) use independent component analysis (ICA) in single-trial data to show that in a human visual selective attention task, ERPs are mainly generated by stimulus induced phase resetting of electroencephalographic processes. Mazaheri and Jensen (2006) have argued against this view, by showing that in single trials the α oscillations, after visual stimuli, preserve their phase relationship with respect to the phase before the stimuli. Moreover, they have showed that the ERF can be explained by stimulus-locked activity in the β band that is absent before the stimulus. Nikulin et al. (2007) have proposed an alternative mechanism that explains generation of evoked responses with a baseline shifting due to modulations in the amplitude of α oscillations after stimuli.

The aim of this chapter is to investigate the role of backward connections in the elaboration of long-latency ERP responses. This enquiry has been enabled by recent advances in the analysis of EEG data; specifically, DCM. A detailed description of DCM can be found elsewhere (David and Friston, 2003; David et al., 2005; 2006; Kiebel et al., 2006; Garrido et al., 2007a) and in the previous chapters (**Chapter 2 and 3**). Models with and without backward connections were compared against each other. The contribution of forward and backward connections to predicted responses at the source level was quantified as a function of peri-stimulus time. In brief, EEG was recorded from healthy subjects whilst listening to a stream of auditory tones embedded in an oddball paradigm, as described in the previous chapter. The same data was used here, however, only the ERP elicited by the deviant stimulus was analysed. The results are reported at both the subject and group level; using the ERP averaged over trials within-subject and the ERP averaged over all subjects. In these analyses only two models were compared; both had the same source architecture (same sources) but were distinguished by the presence of backward connections among sources (see **Figure 4.2** in the results section). As opposed to the analyses performed in the previous chapter, these models attempted to explain a single ERP as a function of peri-stimulus time, instead of the difference between deviants and standards during a fixed time window. Moreover, these models tested for how relevant the presence of forward and backward connections is in explaining late ERP components, whereas the previous analysis focussed on the importance of modulations of these connections in explaining mismatch responses.

4.2 Methods and Statistical analysis

4.2.1 Experimental design

Evoked responses elicited within a pure-tone oddball paradigm were chosen for this study because they are known to comprise substantial late components. The experimental design used in this chapter is identical to the one used in the previous chapter (see **Chapter 3**). In this chapter, however, only the responses to the oddball tones were analysed, as opposed to the previous analysis where both deviant and standard trials were modelled.

4.2.2 Data acquisition and pre-processing

Data acquisition and pre-processing procedures were very similar to those adopted in the previous chapter (**Chapter 3**). In this chapter, however, the dimensionality of the data was reduced to eight channel mixtures or spatial modes, instead of three. These were the eight principal modes of a singular value decomposition of the channel data from trials with responses to deviant tones only. The use of eight principal eigenvariates preserved more than 97% of the variance in every subject (see **Figure 4.1**).

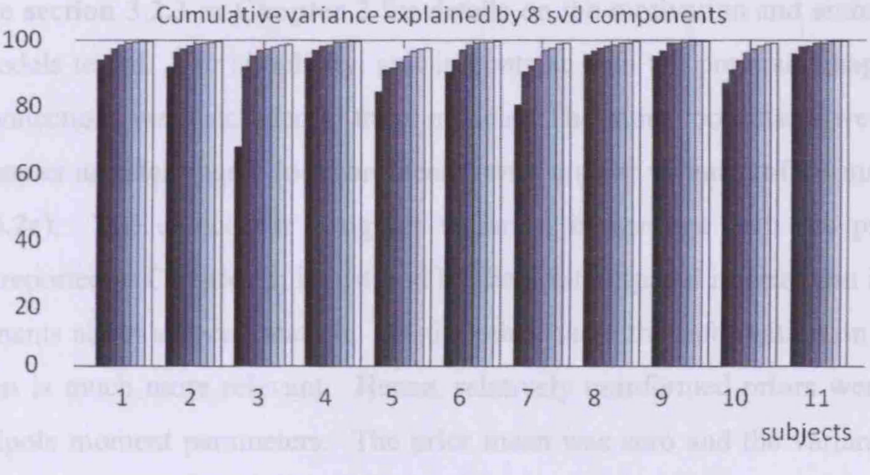


Figure 4.1 Cumulative variance explained by 8 svd components.

4.2.3 Statistical analysis: Bayesian model comparison

The statistical analysis in this chapter uses the procedures described in **Chapter 2**, for Bayesian model comparison at both single-subject and group levels. See **section 2.2** for further details.

4.3 Results

4.3.1 Dynamic Causal Models specification

This section describes two DCMs defined by a given architecture and dynamics. These two models were tested against each other in order to investigate the role of backward connections in the generation of long-latency ERP responses. See **Chapter 2** for a full description of DCM.

The network architecture was motivated by recent electrophysiological and neuroimaging studies looking at the sources underlying the MMN (Opitz et al., 2002; Doeller et al., 2003). Using prior knowledge about the functional anatomy the following DCM was constructed: An extrinsic input entered bilaterally to **A1**, which were connected to their ipsilateral **STG**. Right **STG** was connected with the right **IFG**. See **section 3.2.3** in **Chapter 3** for details on the motivation and architecture of the models tested. For simplicity, and in contrast with the previous chapter, no lateral connections were included in these models. The same coordinates were used in this chapter as prior source location means, with a prior variance of 16 mm^2 (see **Figure 4.2c**). The choice for a tighter variance, in contrast with the previous analysis, reported in **Chapter 1**, is justified by the limited spatial information in EEG measurements about source location. On the other hand, the information on source orientation is much more relevant. Hence, relatively uninformed priors were used for the dipole moment parameters. The prior mean was zero and the variance was 256 mm^2 in each direction.

Given this connectivity graph, specified in terms of its nodes and connections, two models were tested. These differed in terms of the presence of reciprocal or recurrent connections: model **FB** had reciprocal, i.e., forward and backward connections and model **F** lacked backward connections, having forward connections only (**Figure 4.2a, b**). In other words, model **FB** resembles recurrent dynamics, or parallel bottom-up and top-down processing, whereas model **F** emulates simple bottom-up mechanism. Note that these models, **F** and **FB,** are graphically very similar to the models tested in **Chapter 3**. However, here we are interested in the existence of connections *per se*; not changes in connections inducing mismatch responses. For example, the **F** model tested in this chapter has forward connections only; there are no backward or lateral connections, as before.

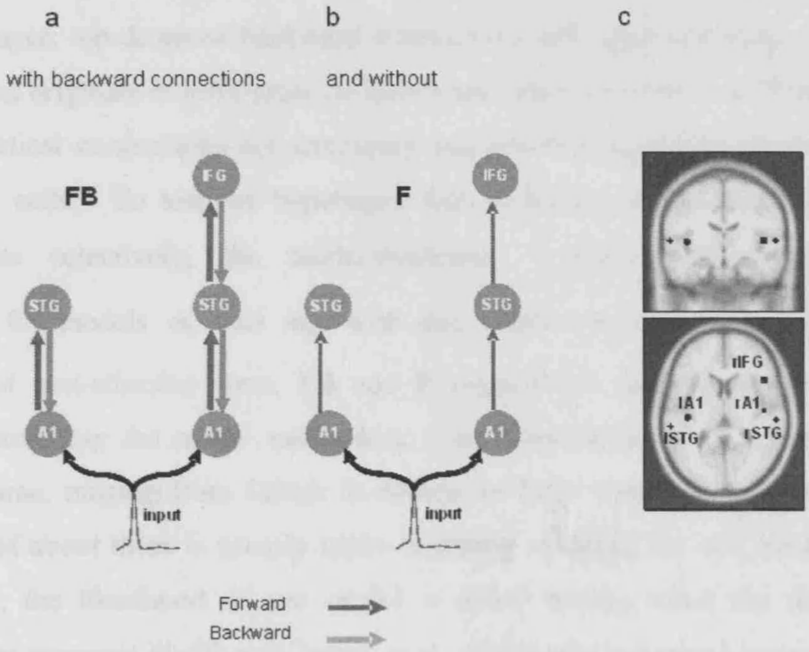


Figure 4.2 Model specification. The sources comprising the network are connected with backward (grey) and/or forward (dark grey) connections as shown. **A1**: primary auditory cortex, **STG**: superior temporal gyrus, **IFG**: inferior temporal gyrus. Two different models were tested within the same architecture, with and without backward connections; **(a)** and **(b)** respectively. **(c)** Sources of activity, modelled as equivalent dipoles (estimated posterior moments and locations), are superimposed in an MRI of a standard brain in MNI space and their prior mean locations are: **lA1** [-42,-22, 7], **rA1** [46, -14, 8], **lSTG** [-61, -32, 8], **rSTG** [59, -25, 8], **rIFG** [46, 20, 8] in mm.

4.3.2 ERPs and Bayesian model comparison

The ERPs elicited by the deviants were modelled using a DCM with these five sources (*c.f.*, equivalent current dipoles) as in previous analyses of mismatch differences (Kiebel et al., 2006; Garrido et al., 2007a). The evoked response to the deviant peaked on average at about 100 ms and had a wide spread negative topography at the frontal electrodes, in agreement with previous studies (**Figure 4.3a**). The free parameters of these models included the location and orientation of each dipole, excitatory and inhibitory rate constants and post-synaptic densities for the three sub-populations of neurons within each source, intrinsic (within-source) coupling strengths and extrinsic (between-sources) coupling. Extrinsic connections were divided into forward and backward and conformed to the connectivity rules described in Felleman and Van Essen (1991) and David et al. (2005). Bottom-up or forward connections originate in the infra-granular layers and terminate in the granular layer; top-down or backward connections link agranular layers and lateral connections originate in infra-granular layers and target all layers. All these extrinsic cortico-cortical connections are excitatory and are mediated through the axons of pyramidal cells. To test the hypothesis that backward connections mediate late components selectively, the model-evidence, $L_i = p(y \mid \theta, m_i) \approx \exp(F_i)$, was evaluated for models m_1 and m_0, with and without backward connections as a function of peri-stimulus time; **FB** and **F** respectively (see **Figure 4.3c**). This involved inverting the model using data, from stimulus onset to a variable post-stimulus time, ranging from 120ms to 400ms, in 10ms steps. A difference in log-evidence of about three is usually taken as strong evidence for one model over the other (*i.e.*, the likelihood of one model is about twenty times the other). The evidence or marginal likelihood (Penny et al., 2004) was compared between the two models as a function of increasing peri-stimulus time windows, for both the grand average ERP across subjects and for each subject individually (**Figures 4.2c** and **4.4** respectively). Both analyses revealed the same result. The longer evoked responses evolve, the more likely backward connections appear. For the group data this is evident in **Figure 4.3c**, which shows that the model with backward connections (**FB**) supervenes over the model without (**F**). This is particularly clear later in peri-stimulus time (220 ms post-stimulus or later).

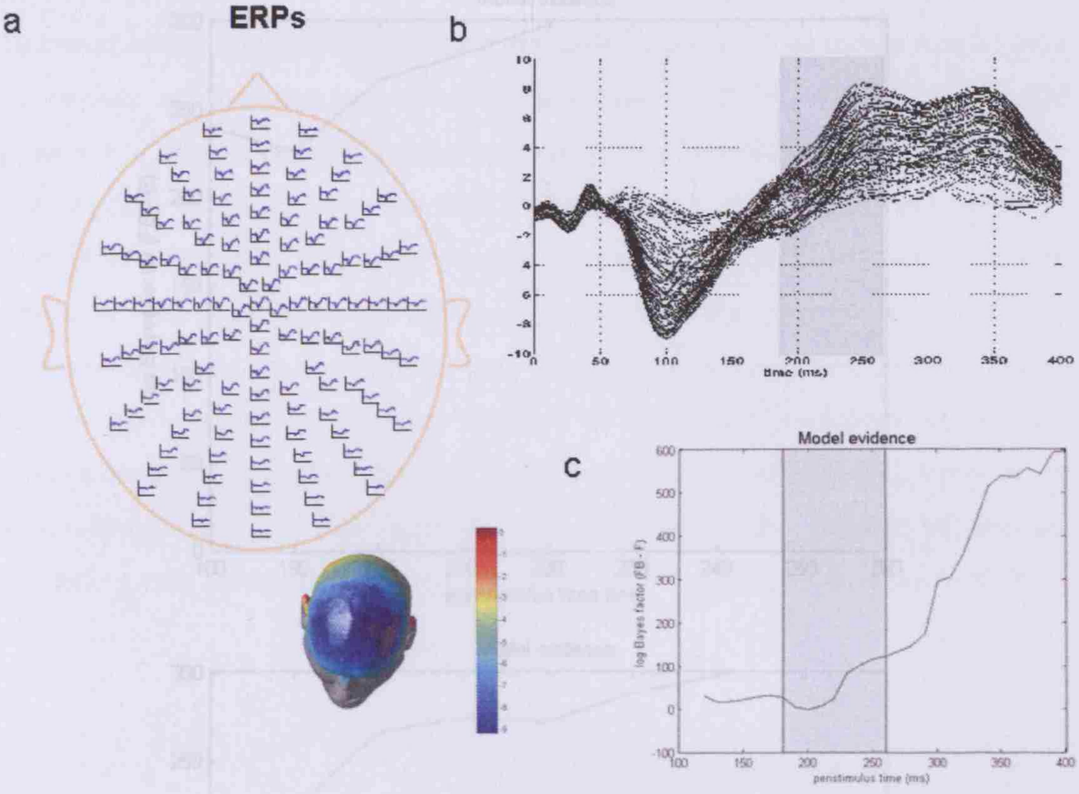


Figure 4.3 Bayesian model comparison among DCMs of Grand mean ERPs. (a) Grand mean ERP responses, *i.e.*, averaged over all subjects, to the deviant tone overlaid on a whole scalp map of 128 EEG electrodes. The artefact in the left frontal electrode site as shown in **Figure 3.3a** is no longer present. This was caused by a bad channel (in one of the subjects) that has been excluded from this analysis. **(b)** Overlapped ERP responses to deviant tones from all 128 sensors over the peri-stimulus interval [0, 400] (ms). **(c)** Differences in negative free-energy or log-evidence comparing the model with backward connections (**FB**) against the model without (**F**). The gray patch indicates the interval chosen to model the ERPs for each individual subject (see also **Figure 4.5**).

For completeness and validation purposes, we also evaluated a backward connection-only model. This was done using the grand mean data (as in **Figure 4.3** and **4.4**). The two graphs below show, as expected, a poor performance of the **B** model as compared to models **FB** and **F** (with and without backward connections, respectively). In both cases, models **FB** and **F** perform much better than model **B** (see **Figure 4.4**). As anticipated, this model had much lower evidence than either the forward or forward and backward models considered here (over all peri-stimulus times examined).

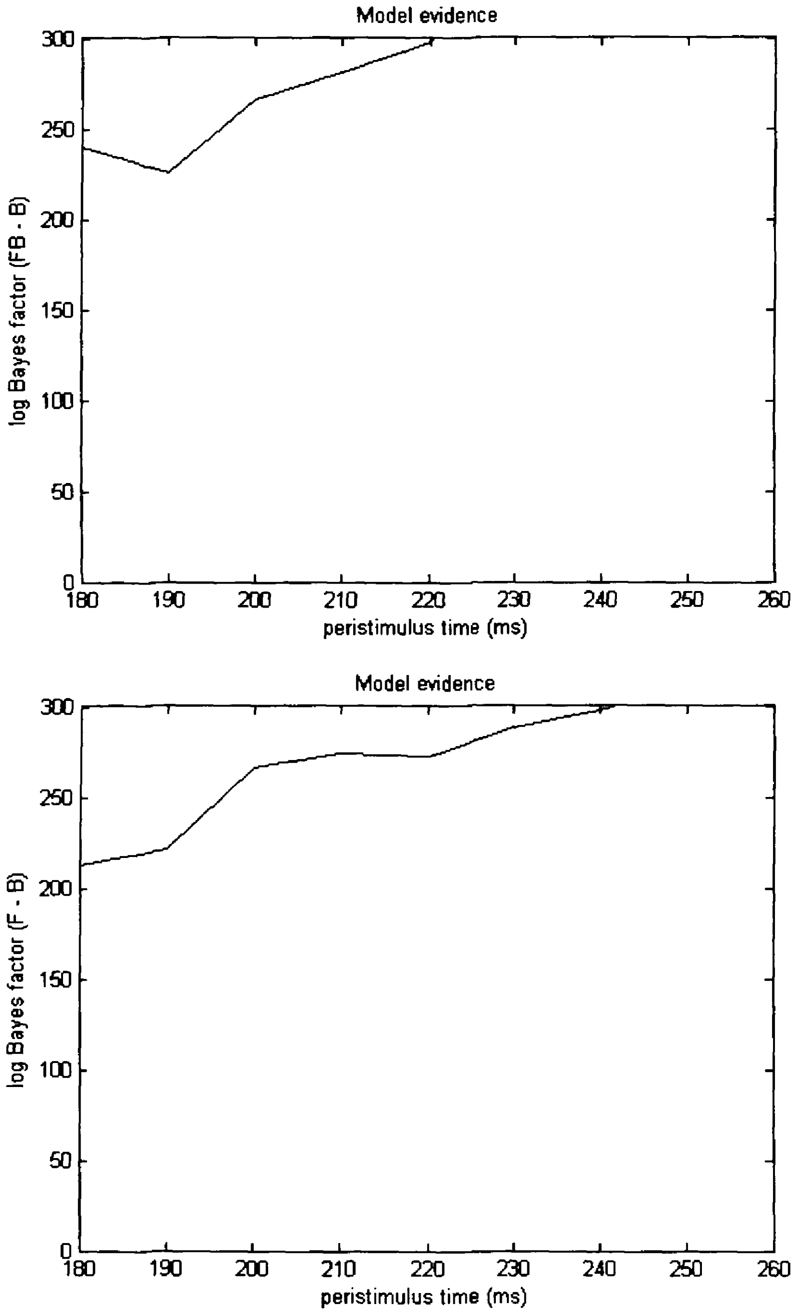


Figure 4.4 Evaluation of the backward connection-only model. These two graphs show, as expected, a poor performance of the **B** model as compared to models **FB** and **F** (with and without backward connections, respectively).

Motivated by these results, a window of interest was selected; 180 to 260 ms for expediency, to perform an identical analysis for each subject. The results for individual subjects recapitulated the group analysis (**Figure 4.5**). For the majority of subjects (8 out of 11), the forward model supervenes over the model with backward connections, when explaining the data in the first half of peri-stimulus time.

Conversely, in the second half, for most subjects (8 out of 11), the model with backward connections supervenes over the model without. This means that forward connections are sufficient to explain ERP generation in early periods, but backward connections become essential in later periods. This effect occurs after 220 ms and is more evident for longer latencies. In short, backwards connections are not necessary to explain early data and only incur a complexity penalty, without increasing accuracy. This does not mean backward connections are 'switched off'; it simply means their effects are not manifest until later in peri-stimulus time, by which time activity has returned from higher levels. At this point, backward connections become necessary to explain the data. This can be seen quantitatively in a plot of the log-evidence over time and qualitatively, in terms of the number of subjects supporting each model, at larger peri-stimulus time (**Figure 4.5 a, b** respectively).

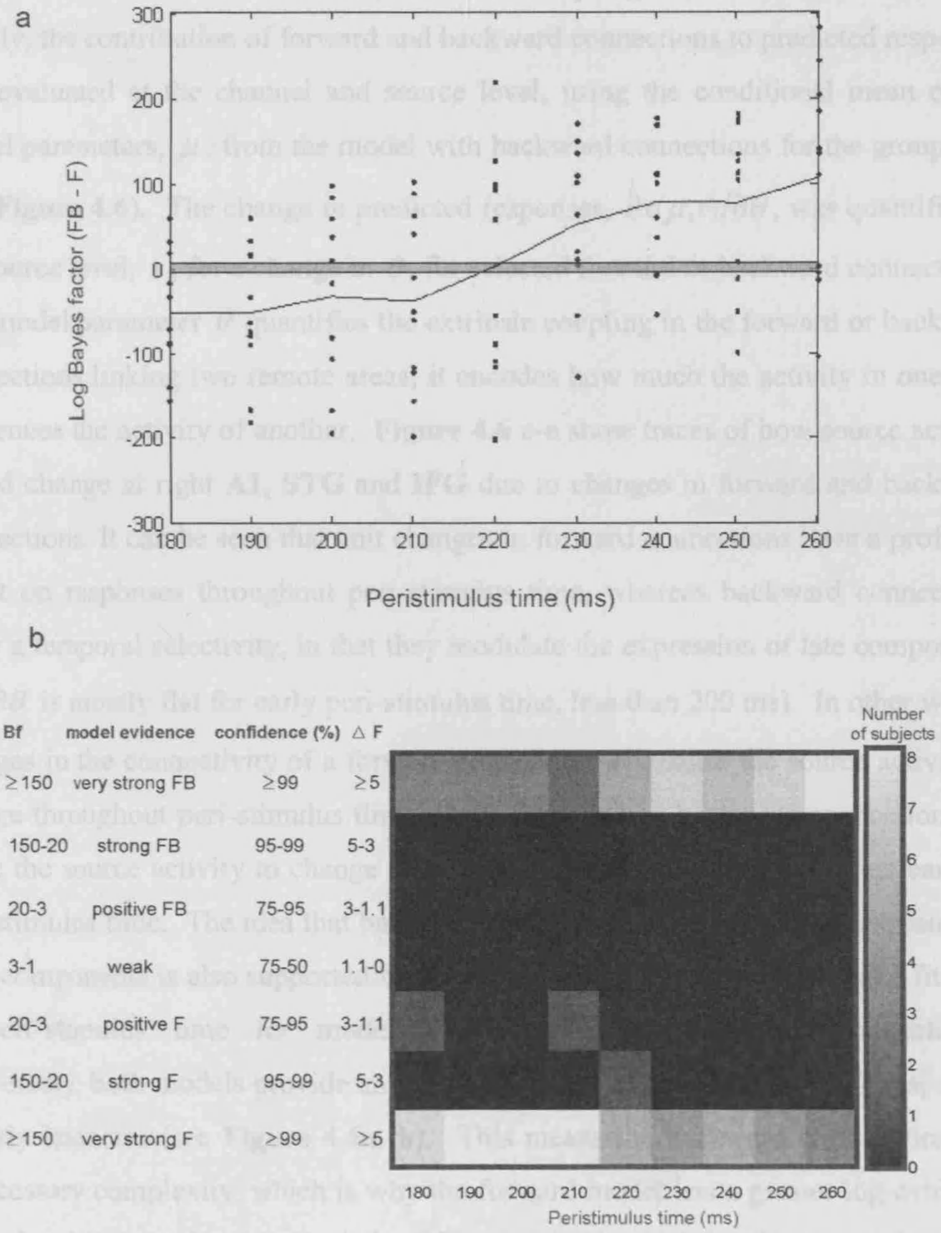


Figure 4.5 Bayesian model comparison across subjects. (a) Comparison of the model with backward connections (**FB**) against the model without (**F**), across all subjects over the peri-stimulus interval 180 to 260 ms. The dots correspond to differences in log-evidence for 11 subjects over time. The solid line shows the average log-evidence differences over subjects (this is proportional to the log-group Bayes factor (Bf) or to the differences in the free energy of the two models (ΔF)). The points outside the gray zone imply very strong inference (≥99% confidence that one model is more likely), *i.e.*, model **FB** supervenes over **F** for positive points and the converse for negative points. **(b)** Histogram showing the number of subjects in each of seven levels of inference on models with and without backward connections across the peri-stimulus interval 180 to 260 ms.

4.3.3 Conditional contribution of extrinsic coupling

Finally, the contribution of forward and backward connections to predicted responses was evaluated at the channel and source level, using the conditional mean of the model parameters, μ, from the model with backward connections for the group data (see **Figure 4.6**). The change in predicted responses, $\partial x(\mu,t)/\partial\theta$, was quantified at the source level, x, for a change in θ, for selected forward or backward connections. The model parameter θ quantifies the extrinsic coupling in the forward or backward connections linking two remote areas; it encodes how much the activity in one area influences the activity of another. **Figure 4.6 c-e** show traces of how source activity would change at right **A1**, **STG** and **IFG** due to changes in forward and backward connections. It can be seen that unit changes in forward connections have a profound effect on responses throughout peri-stimulus time, whereas backward connections show a temporal selectivity, in that they modulate the expression of late components ($\partial x/\partial\theta$ is mostly flat for early peri-stimulus time, less than 200 ms). In other words, changes in the connectivity of a forward connection will cause the source activity to change throughout peri-stimulus time, while changes in a backward connection will cause the source activity to change at long latencies but will have no effect early in peri-stimulus time. The idea that backward connections are necessary to explain late ERP components is also supported by the remarkable improvement of model fit later in peri-stimulus time for model **FB** (afforded by backward connections). Conversely, both models provide an equally good fit when modelling ERP responses at early latencies (see **Figure 4.6a, b**). This means the backward connections add unnecessary complexity, which is why the forward model has a greater log-evidence in, and only in, early phases of the ERP (see **Figure 4.5a**). These graphs show predicted and observed responses with and without backward connections projected onto the first eigenmode of a principal component analysis (PCA). It is evident that the model fit improves later in time, after about 200 ms, by adding backward connections to the network. This is coincident with the time interval for which Bayesian model comparison revealed that the model with backward connections explains the data much better at both group and individual subject levels.

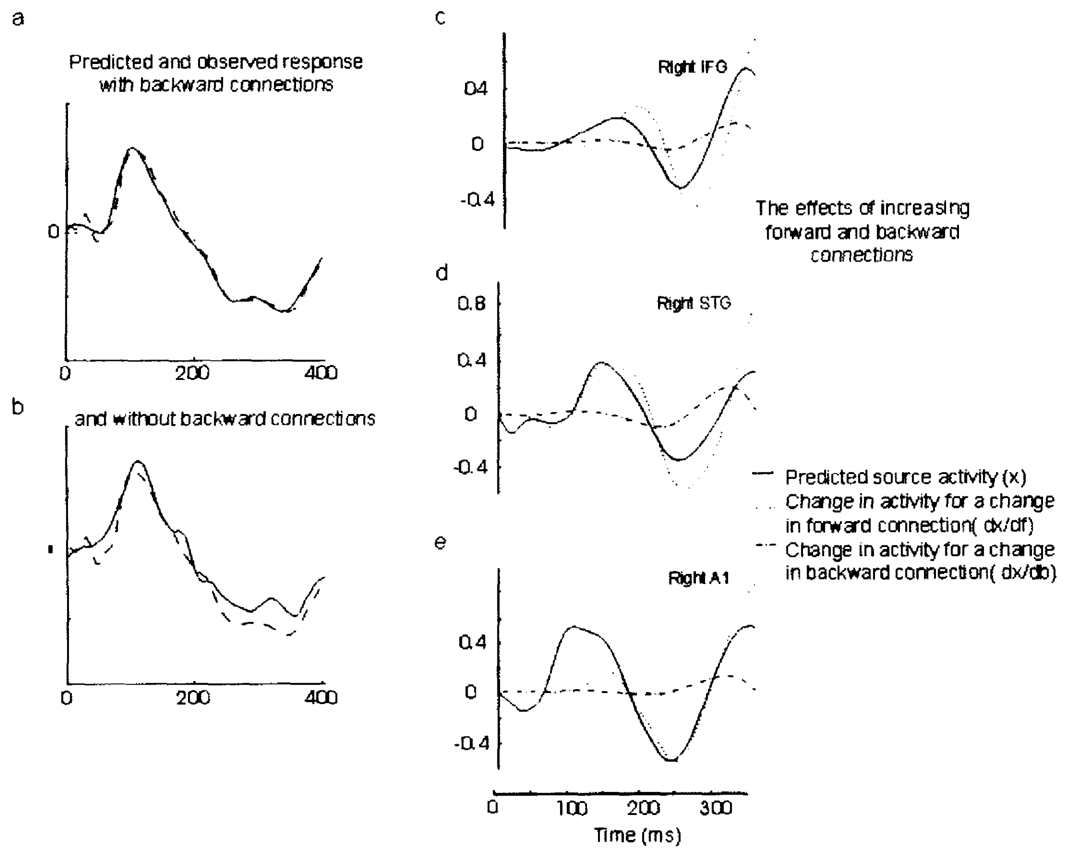


Figure 4.6 Contribution of extrinsic coupling to source activity. The graphs show predicted (solid) and observed (broken) responses in measurement space for the first spatial mode, which was obtained after projection of the scalp data onto eight spatial modes; for (a) **FB** model and (b) **F** model. The first mode accounts for the greatest amount of observed variance. The improvement of model fit due to backward connections for later components is evident. Predicted responses at each source (solid line) and changes in activity with respect to a unit change in forward (dotted line) and backward connection (dash-dotted line) for **(c)** right **IFG**, **(d)** right **STG** and **(e)** right **A1**. The gray bar covers the same period of peri-stimulus time as in **Figure 4.3**.

4.4 Discussion

In this chapter, DCM was used to provide direct evidence for the theoretical prediction that evoked brain responses are mediated by reentrant dynamics or top-down effects in cortical networks, and therefore rest on backward connections. This provides an explicit statistical test of the hypothesis that evoked responses depend on top-down effects (Syndulko et al., 1982; Gaillard et al., 1988). Furthermore, the relative contribution of bottom-up and top-down effects were quantified under a biologically informed model. These results show that backward connections are necessary to explain late neuronal responses. This evidence was furnished by comparison of models of oddball responses with and without backward connections. Bayesian model comparison revealed that the model with backward connections explains both group and individual data better than the model with forward connections only. This was particularly evident for long latencies (more than 200 ms, see **Figure 4.3c** and **Figure 4.5**). Furthermore, it was possible to quantify the contribution of backward connections to evoked responses (in both channel and source space) as a function of peri-stimulus time. As expected, these results show that forward connections have a profound effect on responses throughout peri-stimulus time, whereas backward connections show a temporal specificity, in that they mediate the expression of late components (more than 200 ms).

Backward connections are an important part of functional brain architectures, both empirically and theoretically. The distinction between forward and backward connections rests on the notion of cortical hierarchies and the laminar specificity of their cells of origin and termination (Boussaoud et al., 1990). Anatomically, backward connections are more abundant than forward connections (in the proportion of about 1:10/20) and show a greater divergence and convergence. Forward connections have sparse axonal bifurcations and are topographically organised whereas backward connections show abundant axonal bifurcation and diffuse topography transcending various hierarchical levels. Functionally, backward connections have a greater repertoire of synaptic effects: while forward connections mediate postsynaptic effects through fast AMPA and GABA$_A$ receptors (constant decay of about 1-6 ms), backward connections also mediate synaptic effects by slow NMDA receptor, which are voltage sensitive and therefore show non-linear dynamics or modulatory effects (with time-constants about 50 ms) (Rockland and Pandya, 1979; Salin and Bullier, 1995). Furthermore, the deployment of backward synaptic connections on the dendritic tree can endow them with nonlinear and veto-like properties (Mel, 1993). Backward connections play a central role in most theoretical and computational formulations of brain function (Douglas and Martin, 2004; 2007); ranging from the role of reentry in the theory of neuronal group selection (Edelman, 1993) to recurrent neural networks as universal non-linear approximators (Wray and Green, 1994). Several previous studies have highlighted the functional role of backward connections, especially in the visual domain. It has been suggested that visual perception or awareness emerges from neuronal activity in ascending and descending pathways that link multiple cortical areas (Pollen, 1999). Accordingly, recurrent processing or cortical feedback is necessary for object recognition (Lamme and Roelfsema, 2000) and has been found to be important in differentiation of figure from ground, particularly for stimuli with low salience (Hupe et al., 1998). Modern day formulations of Helmholtz's ideas about perception suggest that backward connections play a critical role in providing top-down predictions of bottom-up sensory input (Rao and Ballard, 1999). Indeed, the hypothesis that the brain tries to infer the causes of its sensory input, refers explicitly to hierarchical models that may be embodied by cortical hierarchies (Friston, 2005). In these formulations, the brain suppresses its free energy or prediction error to reconcile predictions at one level in the hierarchy with those in neighbouring levels. This entails passing prediction errors up the hierarchy (via forward connections) and passing predictions down the hierarchy (via backward connections), which is in conformity with a predictive coding framework based on hierarchical Bayes (Rao and Ballard, 1999; Friston, 2003; 2005). Experimental evidence, consistent with predictive coding models has been furnished by fMRI studies (Murray et al., 2002; Bar et al., 2006; Summerfield et al., 2006). It has been shown that activity in early visual areas is reduced through cortical feedback from high to low-level areas, which simplifies the description of a visual scene (Murray et al., 2002) and facilitates object recognition (Bar et al., 2006). A recent study has also used DCM to test predictive coding models in the context of perceptual decisions, and found an increase in top-down connectivity from the frontal cortex to face visual areas, when ambiguous sensory information is provided (Summerfield et al., 2006). In predictive coding, evoked responses correspond to prediction error that is explained away (within trial) by self-organising neuronal dynamics during perception and is suppressed (between trials) by changes in synaptic efficacy during learning. The study described in this chapter focused on the underlying mechanisms of responses evoked in an oddball paradigm. The results reported above are yet another piece of empirical evidence in favour of predictive coding. However, the recurrent dynamics that ensue are a plausible explanation for the form of evoked responses in general, and the theoretical cornerstone of most modern theories of perceptual inference and learning.

4.5 Summary

The aim of the work presented in this chapter was to investigate the role of backward connections in the generation of late ERP components. The hypothesis was that early components reflect a perturbation of neuronal dynamics by sensory inputs, which are associated with bottom-up processing. Conversely, later, endogenous components can be ascribed to recurrent dynamics among hierarchically disposed cortical processing levels which are associated with top-down effects. DCMs of responses to deviants, elicited in an oddball paradigm were modelled with and without backward connections. These two models were compared in order to assess their likelihood as a function of peri-stimulus time. This analysis revealed that evoked brain responses are generated by recurrent dynamics in cortical networks, and that backward connections are necessary to explain late components. Furthermore, it was possible to quantify the contribution of backward connections to evoked responses and to source activity, again as a function of peri-stimulus time. These results link a generic feature of brain responses to changes in the sensorium and a key architectural component of functional anatomy; namely backward connections are necessary for recurrent interactions among levels of cortical hierarchies.

Chapter 5

A Predictive Coding account of the Mismatch Negativity

In the two previous chapters (**Chapters 3** and **4**) a 5-area reciprocally connected network was used to model the MMN. It could be asked whether this network is correctly specified; in other words, does it correspond to the actual underlying network? For example, the network used might include an irrelevant source, it may be missing a relevant source, or have wrong connections. Although there is no such thing as a right or wrong model (only better or worse approximations to reality); these concerns can be addressed with model comparison and selection. This chapter deals with this issue by testing different biologically plausible networks. Furthermore, the network models tested attempt to map onto hypotheses debated in the MMN literature; with a specific focus on the relative roles of plastic change in extrinsic and intrinsic connections. These competing hypotheses or models were assessed with Bayesian model comparison, which provides a principled way for selecting among alternative models.

5.1 Introduction

Early work by Näätänen and colleagues suggested that the MMN results from a comparison between the auditory input and a memory trace of previous sounds. In agreement with this theory, others (Winkler et al., 1996; Näätänen and Winkler, 1999; Sussman and Winkler, 2001) have postulated that the MMN could reflect on-line modifications of the auditory system; in other words, the MMN would correspond to updates of the perceptual model during incorporation of a newly encountered stimulus into the model – *the model-adjustment hypothesis*. Hence, the MMN would be a specific response to stimulus change and not to the stimulus alone. This hypothesis has been supported by evidence that the prefrontal cortex is involved in a top-down modulation of the deviance detection system in the temporal cortices (Escera et al., 2003). In the light of the Näätänen model, it has been claimed that the MMN is caused by two underlying functional processes, a sensory memory mechanism related to temporal generators, and an automatic attention-switching process related to the frontal generators (Giard et al., 1990). Accordingly, it has been shown that the temporal and frontal MMN sources have distinct behaviours over time (Rinne et al., 2000) and that these sources interact with each other (Jemel et al., 2002). Thus, the MMN could be generated by a temporofrontal network (Opitz et al., 2002; Doeller et al., 2003). These two M/EEG and fMRI studies have linked the early component (in the range of about 100-140 ms) to a sensorial, or non-comparator account of the MMN elaborated in the temporal cortex. The later component (in the range of about 140-200 ms) has been associated with a cognitive account of the MMN involving the frontal cortex (Maess et al., 2007). However, a recent study (Jääskeläinen et al., 2004) has challenged the view that the MMN is generated by a temporal-frontal cortical network. Instead, according to Jääskeläinen et al. (2004) the observed response could result from a much simpler mechanism of local *adaptation* at the level of the auditory cortex that causes attenuation and delay of the N1 response. The N1 response is the negative component peaking at about 100 ms from stimulus onset and is associated with early auditory processing occurring at the level of A1. As a consequence of adaptation, the observed response (i.e., the MMN) would be erroneously interpreted as a separate component from the N1 wave – *the adaptation hypothesis*. According to this view, the fact that the neuronal elements within the auditory cortex become less responsive upon subsequent stimulation is sufficient to explain the generation of a supposed MMN. Generation of this response (a delayed and suppressed N1) would be confined to the auditory cortex, and the MMN reported in the literature would emerge as an artefact due to the subtraction procedure (see also **Chapter 1** for a review of the underlying mechanisms of MMN generation).

The aim of this chapter is to disambiguate between these hypotheses for the MMN generation: *adaptation* and *model-adjustment*. Competing mechanistic hypotheses were framed in terms of network models (or DCMs) characterised by repetition-dependent changes in coupling within and between cortical areas. Bayesian model comparison of DCMs was used to make inferences about the best model. The models examined were chosen to map onto the hypotheses that the MMN is generated by (i) local changes in coupling, i.e. *adaptation*; in other words, the MMN is best explained by neuronal disinhibition, confined to lower-order cortical areas (*cf.* Jääskeläinen et al., 2004); (ii) hypotheses entailed by interactions of a temporo-frontal network or *model-adjustment* (Winkler et al. 1996, Doeller et al. 2003), and (iii) a combination of both mechanisms, as suggested by formal models of perceptual learning (Friston, 2005; Baldeweg 2006), i.e., local adaptation within an area and changes in interactions between areas that reflect a reduced mismatch response mediated by top-down predictions. The experimental results suggest that rather than being mutually exclusive, both intrinsic (*adaptation*) and extrinsic (*model adjustment*) changes in coupling are required to explain the MMN. These results are discussed in terms of predictive coding and hierarchical inference in the brain.

5.2 Methods and Statistical analysis

5.2.1 Experimental design

5.2.1.1 Subjects

A group of twelve healthy volunteers aged 24–34 (4 female) gave signed informed consent before the study, which proceeded under local ethical committee guidelines.

5.2.1.2 Stimuli and Task

The experimental design used in this study is similar to that described in the previous two chapters (**Chapters 3** and **4**). However, in this experiment standards and deviants have a smaller relative difference (10% in this paradigm as opposed to 100% in the paradigm used previously). Here, the electroencephalographic activity was measured during a classical auditory 'oddball' paradigm, in which subjects were presented with "standard" (500 Hz) and "deviant" tones (550 Hz), occurring 80% (480 trials) and 20% (120 trials) of the time, respectively, in a pseudo-random sequence. This was done in order to prevent conflation of the MMN with the N1, because the bigger the difference between standards and deviants, the smaller the MMN latency (about 100 ms) which overlaps with the N1 peak. The stimuli were presented binaurally via headphones for 10 minutes every 0.5 seconds (*c.f.*, the ISI of 2 seconds, as used in previous chapters). The use of a shorter ISI was another strategy for dissociating the MMN and the N1 response.

In order to preclude emergence of a N2 and a P300 components, a second modification was introduced: subjects performed a distracting visual task and were instructed to ignore the sounds. The task consisted of button-pressing whenever a fixation cross changed its luminance. This occurred pseudo-randomly every 2 to 5 seconds (and did not coincide with auditory changes).

5.2.2 Data acquisition and pre-processing

Data acquisition and pre-processing procedures were very similar to those adopted in the previous chapters (**Chapter 3** and **4**). In this study, however, data were re-

referenced to the nose, instead of to the average of the right and left ear lobes. The method used for artefact removal was robust averaging. Robust averaging is an iterative algorithm that produces the best estimate of the average by weighting data points as a function of their distance from estimate of the mean for each iteration (*cf.* Wager et al., 2005). For computational expediency the dimensionality of the data was reduced to eight channel mixtures or spatial modes (as in **Chapter 4**). The use of eight principal eigenvariates explained more than 74% of the data in all subjects and preserved the interesting components of evoked responses (see **Figure 5.1**). As in **Chapter 3**, responses to both deviants and standards were modelled.

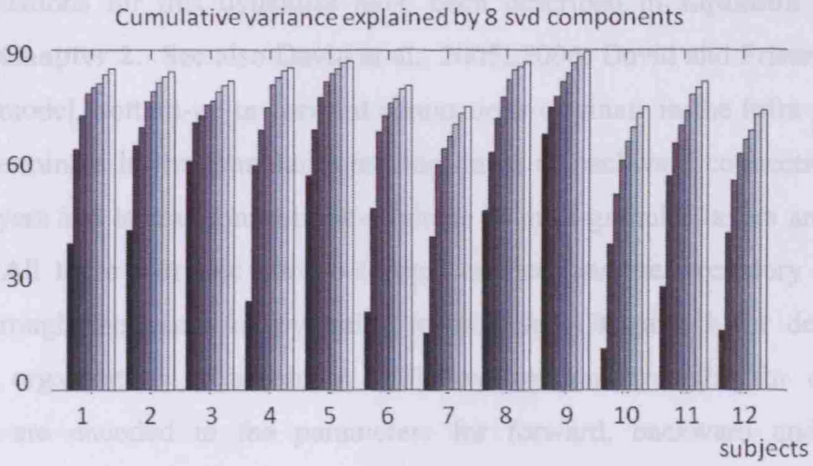


Figure 5.1 Cumulative variance explained by 8 svd components.

5.2.3 DCM extension: modulation of intrinsic connectivity

In this chapter, DCM is used to investigate connectivity models in a specific context: namely, explaining the basis of the MMN. Here, the main hypotheses, *adaptation* and *model-adjustment*, are framed in terms of connectivity models or DCMs, characterised by repetition-dependent changes in coupling within and between cortical areas. Specifically, adaptation effects are modelled in DCM by changes in intrinsic or self-connections that are confined to a cortical area (within-area effects). This section presents a brief summary of requisite methodological developments, which are described in detail in Kiebel et al. (2007).

As mentioned in **Chapter 2**, DCMs for MEG/EEG use neural mass models (David and Friston, 2003), to explain source activity in terms of the ensemble dynamics of the interacting inhibitory and excitatory subpopulations of neurons, based on the model of Jansen and Rit (1995). This model emulates the activity of a source using three neuronal subpopulations, each assigned to one of three cortical layers; an excitatory subpopulation in the granular layer, an inhibitory subpopulation in the supra-granular layer and a population of deep pyramidal cells in the infra-granular layer. The excitatory pyramidal cells receive excitatory and inhibitory input from local interneurons (via intrinsic connections, confined to the cortical sheet), and send excitatory outputs to remote cortical areas via extrinsic connections. The full set of the state equations for this dynamics have been described in **Equation 2.3** and **Figure 2.1**, **Chapter 2**. See also David et al., 2005; 2006; David and Friston, 2003. Within this model, bottom-up or forward connections originate in the infra-granular layers and terminate in the granular layer; top-down or backward connections link agranular layers and lateral connections originate in infra-granular layers and target all layers. All these extrinsic cortico-cortical connections are excitatory and are mediated through the axons of pyramidal cells. See **Chapter 1** for details on hierarchical organisation in the brain. The connection strengths for extrinsic connections are encoded in the parameters for forward, backward and lateral connections. Interactions among the subpopulations depend on internal coupling constants, which control the strength of intrinsic connections and reflect the total number of synapses expressed by each subpopulation. Intrinsic connectivity is encoded in the maximum amplitude of the synaptic kernel, H_e, which corresponds to the peak of the post-synaptic potential and the intrinsic excitability of cells (see **Equation 2.4**, **Chapter 2**). Hence, the modulation of intrinsic or self connections corresponds to gains on H_e and emulates local adaptation; in other words, it models a mechanism by which changes in source activity are caused by the dynamics of the source itself.

5.2.4 Model specification

The networks chosen were motivated by the results of previous studies of the generators of the MMN (Opitz et al., 2002; Doeller et al., 2003; Rinne et al., 2000;

Jääskeläinen et al., 2004) and formulated in terms of the hypotheses mentioned above: *adaptation*, *model-adjustment* and *predictive coding*.

Eight plausible models were specified in terms of specific architectures that mapped onto eight different explanatory mechanisms. The model search started with a parsimonious model that gradually increased in its complexity by addition of hierarchical levels (i.e., sources and extrinsic connections in the network). Model **S2** comprised two nodes in the left and right **A1**. Nodes and connections were added to the network to elaborate a symmetric three-level hierarchical model. This model was motivated by recent electrophysiological and neuroimaging studies that identified underlying sources for the MMN (Doeller et al., 2003; Opitz et al., 2002). These studies found bilateral sources located in the STG and the IFG, which are usually stronger and identified more consistently in the right hemisphere. All models can therefore be considered as special cases of this three-level hierarchical model, **S6i** (see **Figure 5.3**).

These models attempt to explain the generation of trial-type-specific individual responses (i.e., responses to standards and responses to deviants) under the constraint that differences among trial types have to be explained by, and only by, differences in coupling in specified connections. Therefore, all extrinsic connections (i.e., plasticity) were allowed to change. In addition, every model was considered with, and without, changes in intrinsic connections in the left and right **A1**. These two regions were chosen as cortical input stations for auditory information. As in the previous chapter, each active source, *i.e.*, each node in the network was modelled with a single ECD. The same priors on source location means and variances were used in this study (see **Figure 5.5**; see also **Chapters 3** and **4**). These parameters were used as priors to estimate the posterior locations and moments of the ECDs, for each subject.

The simplest model, **S2**, is a two source network that maps to the hypothesis that ERPs to standards and deviants are generated by bilateral activity in **A1**. This model does not allow changes in the connectivity parameters and cannot model a MMN difference. Model **S2i** is similar to **S2** but allows for intrinsic coupling changes within **A1**. Here, it is hypothesised that differences between responses to standards and deviants are caused by changes in **A1** due to differences in self-connections, i.e., within area. Model **S4** is a two-level hierarchical model comprising four sources. It is built upon **S2,** with left and right **STG**. These sources were reciprocally connected (i.e., connected through forward and backward connections) to their ipsilateral **A1**. This model allowed for plastic changes in extrinsic connections. Model **S4i** is analogous to **S4** with additional self-connections within **A1**. A third-level hierarchical model, comprising five sources, model **S5**, built upon **S4** and included right **IFG**, as well as plastic changes between cortical areas. Right **STG** was reciprocally connected with its ipsilateral **IFG**. Model **S5i** is like **S5** but had self-connections within **A1**. Models **S6** and **S6i**, six area models, are extensions of **S5** and **S5i** respectively that included the left **IFG** connected reciprocally to ipsilateral **STG,** as well as plastic changes between cortical areas.

In summary, the models differed in terms of their nodes and in the connections which could show putative learning-related changes, i.e., differences between listening to standard or deviant tones. Models **S2**, **S4**, **S5** and **S6** allowed changes in extrinsic, forward and backward connections, which map to hypotheses that ERP differences with standards and deviants are due to coupling changes in extrinsic connections. Models **S2i**, **S4i**, **S5i** and **S6i** allowed for changes in the same extrinsic connections plus changes in intrinsic connections within left and right **A1**. These models map to the hypothesis that differences in ERPs are due to conjoint coupling changes in extrinsic and intrinsic connections (see **Figure 5.3** for a graphical representation of the models).

5.2.5 Statistical analysis: Bayesian model comparison

The statistical analysis in this chapter uses the procedures described in **Chapter 2**, for Bayesian model comparison at both single-subject and group levels. See **section 2.2** for further details. The next subsection describes the inference made on the parameters of the best model pooled over subjects, i.e., at the group level.

5.3 Results

5.3.1 Event-related potentials

Data were recorded from 128 EEG sensors while subjects were presented with a classical auditory oddball paradigm. **Figure 5.2a** shows the grand mean responses (i.e., averaged across subjects) to the standard (prob. occurrence = 0.8, f = 500 Hz) and to the deviant tones (prob. occurrence = 0.2, f = 550 Hz). Responses to deviant tones exhibited a widespread negativity over the temporal and frontal electrodes, peaking at about 200 ms from change onset, which is consistent with previous studies. Note that this late latency for the MMN is due to the small difference between the two tones, standards and deviants. The difference between the ERPs evoked by the standard and deviant tones revealed a distinct MMN (see **Figure 5.2b**). **Figure 5.2c** shows the 2D scalp topography for the difference wave at its peak. This negativity had a broad spatial pattern encompassing electrodes associated with auditory and frontal areas. These spatiotemporal responses over sensors are characteristic of a typical MMN; however, the aim of this study was to identify the cortical network that generates it.

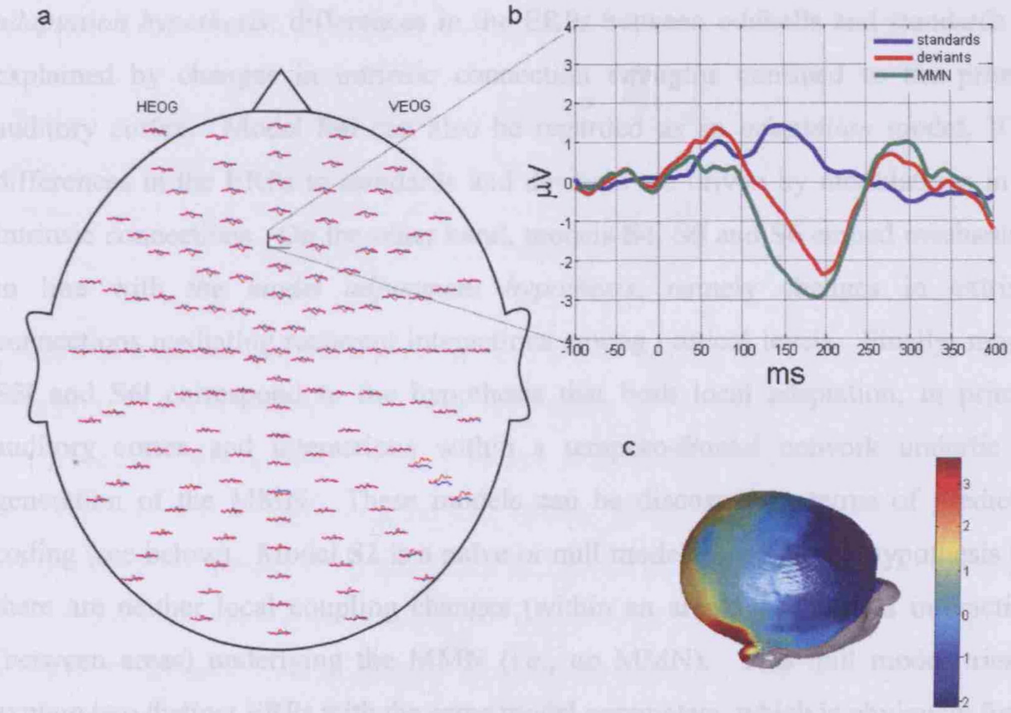


Figure 5.2 Grand mean ERPs, i.e., averaged over all subjects. (a) ERP responses to the standard and deviant tones overlaid on a whole scalp map of 128 EEG electrodes. **(b)** ERP responses to the

standard and deviant tones. The MMN difference wave was obtained by subtracting the grand-average ERP to standards from the ERP to deviants, at channel C21 (fronto-central). **(c)** Grand mean MMN at peak interpolated to give a 3D scalp topography.

5.3.2 Bayesian Model Selection

Eight network models (DCMs) were specified on the basis of the two proposed mechanisms for MMN generation (Winkler et al., 1996 and Jääskeläinen et al., 2004), as well as previous work on the localization of MMN generators (Opitz et al., 2002; Doeller at al., 2003). These eight DCMs were inverted for each of the twelve subjects (see **Figure 5.3** and **Materials and Methods** for a graphical description of the models). The model search started with the most parsimonious model (**S2**); one hierarchical level comprising two sources (bilateral primary auditory cortex - **A1**), and gradually increased its complexity by adding sources and connections, until it reached a three-level hierarchical model with six reciprocally interconnected sources (bilateral **A1**, **STG** and **IFG**). For each hierarchy, all extrinsic connections were allowed to change. In addition, each model was tested with and without coupling changes within the primary auditory cortex. All models are simpler versions of the most complex (**S6i**). Model **S2i** attempts to explain the MMN in terms of *the adaptation hypothesis*: differences in the ERPs between oddballs and standards are explained by changes in intrinsic connection strengths confined to the primary auditory cortex. Model **S4i** can also be regarded as an *adaptation* model, if the differences in the ERPs to standards and deviants are driven by modulations in the intrinsic connections. On the other hand, models **S4**, **S5** and **S6** embed mechanisms in line with *the model adjustment hypothesis*, namely changes in extrinsic connections mediating recurrent interactions among cortical levels. Finally, models **S5i** and **S6i** correspond to the hypothesis that both local adaptation, in primary auditory cortex and interactions within a temporo-frontal network underlie the generation of the MMN. These models can be discussed in terms of predictive coding (see below). Model **S2** is a naïve or null model encoding the hypothesis that there are neither local coupling changes (within an area), nor cortical interactions (between areas) underlying the MMN (i.e., no MMN). This null model tries to explain two distinct ERPs with the same model parameters, which is obviously futile.

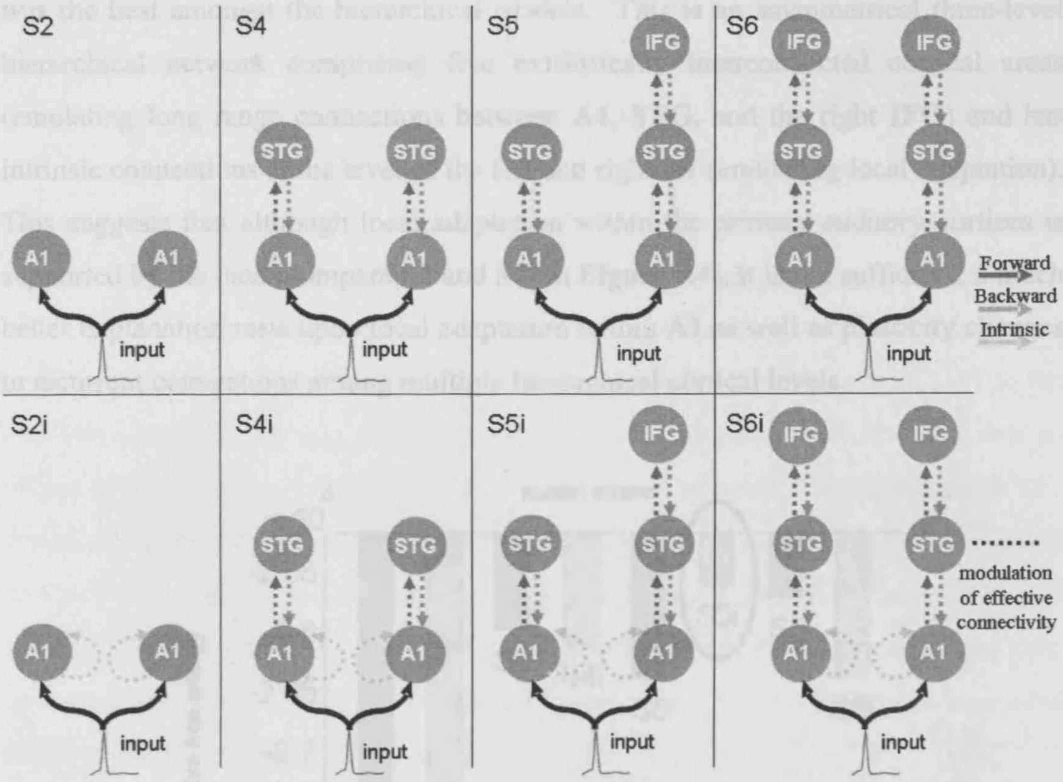


Figure 5.3 Model specifications. The sources comprising the networks are connected with forward (dark grey), backward (grey) or lateral (light grey) connections as shown. **A1**: primary auditory cortex, **STG**: superior temporal gyrus, **IFG**: inferior temporal gyrus. Eight different models with increasing hierarchical complexity were tested. The first row of models, [**S2, S4, S5, S6**], allowed for learning-related changes in only extrinsic (forward and backward) and the second row [**S2i, S4i, S5i, S6i**] allowed for conjoint extrinsic and intrinsic connectivity changes. Each column comprises similar network models, which differ only in allowing for changes of intrinsic connectivity within **A1**. From one column to the next, the number of active sources increased and were reciprocally connected (with forward and backward connections) to extant nodes.

Figure 5.4 shows the profile of negative free energy across the eight models considered; this is a lower-bound approximation to the model log-evidence, which measures how good a model is as compared to another. The model with the highest evidence explains the data with the best balance of accuracy and complexity. These results show that model **S5i** (highlighted in the figure) is the model that best explains the group data. The log-evidence at the group level, i.e., pooled over subjects, was determined under the assumption of data independence over subjects (see **Equation 2.9** in **Chapter 2**). Very strong evidence (ΔF > 5) was found for this model, relative to the remaining models tested. The two models with a single hierarchical level (**S2**

and **S2i**) performed very poorly, compared to the hierarchical models. Model **S5i** was the best amongst the hierarchical models. This is an asymmetrical three-level hierarchical network comprising five extrinsically interconnected cortical areas (emulating long range connections between **A1, STG,** and the right **IFG**) and has intrinsic connections at the level of the left and right **A1** (emulating local adaptation). This suggests that although local adaptation within the primary auditory cortices is supported by the data (compare **S2** and **S2i** in **Figure 5.4**), it is not sufficient; a much better explanation rests upon local adaptation within **A1** as well as plasticity changes in recurrent connections among multiple hierarchical cortical levels.

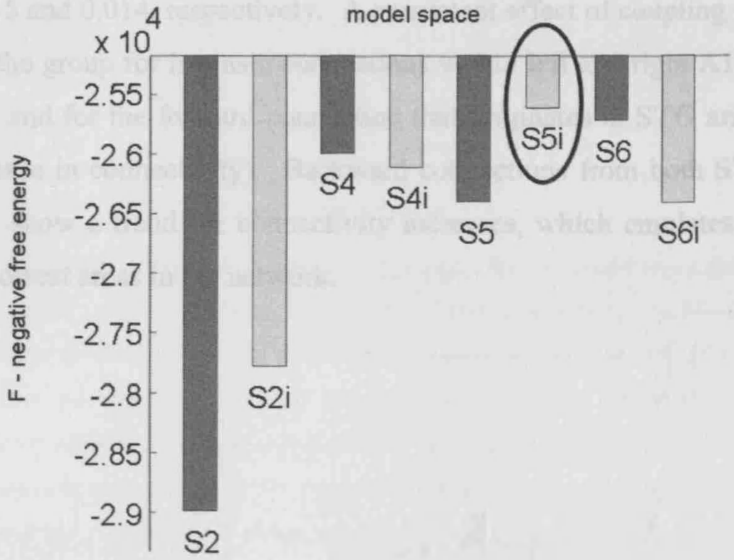


Figure 5.4 Bayesian model selection. This graph shows the free energy approximation to the log-evidence at the group level, *i.e.*, pooled over subjects, for the eight models. The best model is a three-level hierarchical network comprising five interconnected cortical areas with local adaptation within primary auditory cortices (model- **S5i).**

Figure 5.5 presents the results of model inversion quantitatively, for the best model, using the grand mean response over subjects. **Figure 5.5a** shows the prior locations of the dipolar sources overlaid in a MRI image of a standard brain. These six sources (bilateral **A1**, **STG** and **IFG**) were used to construct eight network models. **Figure 5.5b** illustrates the estimates of activity on inverting the best DCM, **S5i,** a three-level hierarchical network comprising five sources interconnected via reciprocal extrinsic connections linking **A1** to **STG,** and **STG** to **IFG**. All these connections were allowed to change between standards and oddballs, with additional plasticity in intrinsic connections at the level of left and right **A1**. The predicted responses at the source level and for each trial type (*i.e.*, standard or deviant) are shown at each node of the network. This figure demonstrates that differences in the ERPs to standards and deviants are expressed in differences in source activity that are caused by differences in coupling between and within these sources. The averaged (over subjects) coupling gains and associated *p*-values are shown against each connection. These values represent a scaling effect. For example, a coupling change of 0.52 from **lSTG** (left STG) to **lIFG** (left IFG) means that the effective connectivity decreased to 52% for rare events relative to frequent events, and a coupling gain of 1.47 in the intrinsic connection within right **A1,** means a 47% increase, with a significant *p*-values of 0.015 and 0.014, respectively. A consistent effect of coupling changes was found across the group for intrinsic connections within left and right **A1** (increase in connectivity), and for the forward connection that originates in **STG** and terminates in **IFG** (decrease in connectivity). Backward connections from both **STGs** to their ipsilateral **A1** show a trend for connectivity increases, which emulates a top-down effect on the lowest areas in the network.

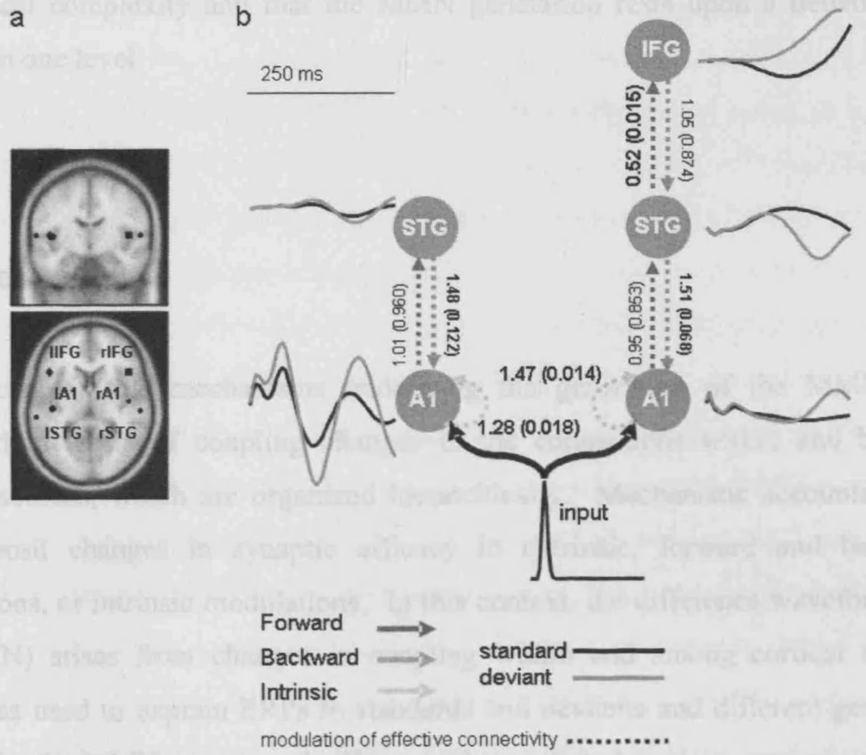


Figure 5.5 DCM estimates of states and coupling changes for the grand mean data under the winning model (S5i). **(a)** Prior locations for the nodes. Sources of activity, modelled as equivalent dipoles (estimated posterior moments and locations), are superimposed in an MRI of a standard brain in MNI space. Their prior mean locations are: **lA1** [-42, -22, 7], **rA1** [46, -14, 8], **lSTG** [-61, -32, 8], **rSTG** [59, -25, 8], **lIFG** [-46, 20, 8], **rIFG** [46, 20, 8] in mm. **(b)** Reconstructed responses to deviants and standards at each source. The mismatch response is expressed in nearly every source. There are widespread learning-related changes in most connections, expressed as coupling gains for deviants relative to standards. Mean coupling changes, i.e., averaged across subjects, and *p*-values lie beside the connections in the graph. Changes in coupling during oddball relative to standards increase consistently across subjects for intrinsic connections within **A1** and decrease between **STG** and **IFG** in the right hemisphere.

5.3.3 Between-subject consistency

A consistency test was performed for the best model across subjects. This was done with an ANOVA for repeated measures, using the log-evidence as the dependent measure. There was a main effect of hierarchical complexity, i.e., number of sources in the network ($F(2.23, 24.56) = 8.109$, p-value < 0.001) but there was no significant effect of intrinsic connectivity ($F(1.00,11.00) = 0.837$, $p = 0.380$). When **S2** and **S2i** were excluded there was no significant effect of hierarchical complexity or intrinsic connectivity. This suggests that the one-level models are driving the main effect of hierarchical complexity and that the MMN generation rests upon a network with more than one level.

5.4 Discussion

In this chapter, the mechanisms underlying the generation of the MMN were evaluated in terms of coupling changes in the connections within and between cortical sources, which are organized hierarchically. Mechanistic accounts of the MMN posit changes in synaptic efficacy in extrinsic, forward and backward connections, or intrinsic modulations. In this context, the difference waveform (i.e., the MMN) arises from changes in coupling within and among cortical sources. DCM was used to explain ERPs to standards and deviants and different generative models for the MMN were tested. The model space was based on previous accounts of MMN generation, specifically *adaptation* (Jääskeläinen et al., 2004), and *model-adjustment* (Winkler et al., 1996).

Model comparison addressed hierarchical implementations of multiple-level network models ranging from one to three levels. These models allowed for changes in extrinsic connections alone; among **A1**, **STG** and **IFG** (i.e., forward and backward connections) or combined with changes in intrinsic connections (i.e., confined to neural subpopulations within cortical units) at the level of **A1**; this models local adaptation as well as interactions between distant cortical areas, through changes in coupling. Significant coupling changes were found across the group, in intrinsic connections within bilateral primary auditory cortices, and extrinsically from right secondary auditory cortex to right frontal cortex. Bayesian model comparison revealed that the best model is a five source network, with intrinsic and extrinsic connections that change between trials. This is an important finding because it provides direct evidence that the MMN is generated by self-organized dynamics within a cortical hierarchy that is mediated by changes in both extrinsic and intrinsic coupling.

5.4.1 Technical issues

A crucial feature of DCM, and any hypothesis driven method in general, is that one cannot test all possibilities, i.e., one has to constrain the model space to a limited number of hypotheses. This study tested an extensive number of possibilities that map onto the major hypotheses concerning the generation of the MMN. Therefore, one can be confident that the winning model is a reasonable approximation to the real cortical network. However, this search on model space does not offer an exhaustive exploration of models; it only selects the best model amongst the models considered. This means there might be another plausible and better network models that explain the MMN. For example, it might be interesting to assess the evidence of a more complex model allowing for local adaptation at the level of **STG** and/or **IFG**. This is a question of model comparison, and provided that there is a good motivation for extending the space of models, one can use Bayesian model comparison to select the best model in the new set. Secondly, an important feature of DCM comparison is that the log-evidence accounts for both model accuracy and complexity; this allows one to compare models with different numbers of parameters (Friston et al., 2006b). Thirdly, DCM uses a conventional formulation of source localization (*cf.* ECD models; see Kiebel et al., 2006), but represents a departure from conventional inverse solutions to the EEG problem by using a full spatiotemporal forward model that embodies known constraints on the way EEG sources are generated. Put simply, these constraints are that electrical activity in one part of the brain must be caused by activity in another (David et al., 2006). Conventional methods localise an active source associated with a specific peak at a given latency. In contrast, DCM models the activity as it evolves over all peri-stimulus times selected; [0, 250] ms in this chapter. Therefore, the models considered here attempt to explain all the dynamics during that interval, including the MMN and any other components peaking under the limits of this interval, such as N1 or P3a. It is often difficult to disentangle the MMN and the N1 components (see Jacobsen and Schröger 2001), especially when there is a big difference between standards and deviants. This was the case in the two previous chapters where there was 100% difference between standards and deviants (**Chapters 3** and 4). In this chapter the conflation of the MMN and the N1 was minimized by using a frequency oddball paradigm with barely discriminable tones. In these cases, as in this study, the MMN peaks at about 200–300 ms, which is considerably later than the N1, peaking at about 100 ms (Näätänen and Alho, 1995).

In summary, this study presents the first attempt to map competing theoretical views onto specific mechanisms or cortical network models of the MMN; and given these models, to select the best in a principled way. The results presented in this chapter show that the MMN cannot be attributed to local adaptation in the primary auditory cortices alone; in other words, adaptation restricted to lower-level areas, is not a complete account of the MMN. This is consistent with a vast literature showing that there are temporal and frontal cortical sources underlying the MMN (Optiz et al., 2002; Doeller et al. 2003; Rinne et al., 2000; Jemel et al., 2002; Restuccia et al., 2005; Molholm et al., 2005). Accordingly, it has been claimed that the MMN is caused by two underlying functional processes, a sensory memory mechanism related to temporal generators and an automatic attention-switching process related to the frontal generators (Giard et al., 1990) and that the prefrontal cortex is involved in a top-down modulation of the deviance detection system in the temporal cortices (Escera et al., 2003). Indeed, the results of this research support the idea that the MMN rests upon a more complex architecture involving plastic interactions amongst multiple hierarchical levels, as well as local adaptation within the primary auditory cortices. This is in agreement with the *model adjustment hypothesis* (Winkler et al., 1996) in conjunction with local *adaptation* (Jääskeläinen et al., 2004). In fact, adaptation effects (neurons showing decreased activity due to stimulus repetition) as an explanation for the MMN had been discussed previously in terms of refractoriness (Näätänen, 1992) or stimulus specific adaptation (SSA) in single neurons in cat primary auditory cortex (Ulanovsky et al., 2003). Combined computational modelling and M/EEG measurements (May et al., 1999) have also addressed the question of MMN generation and discussed mechanisms of local suppressive effects, or adaptation, and lateral inhibition, i.e., synaptic changes in horizontal connections intrinsic to an area. They have shown that the MMN can be explained by both neuronal adaptation and lateral inhibition. Although this work suggests that mechanisms of adaptation also underlie the MMN, it is distinct from *the adaptation hypothesis*, which claims that the generation of the observed response is due to local adaptation alone, that is confined to the auditory cortex, and that there is no separate MMN (for a critical assessment see Näätänen et al., 2005).

The results obtained in this chapter are consistent with the conjecture that the MMN reflects perceptual learning of standards, using predictive coding (Friston, 2003; 2005; Garrido et al., 2007a), i.e., adaptive changes in connectivity during perceptual discrimination of sounds (standards and deviants). These ideas are rooted in predictive coding models based on hierarchical Bayes (Rao and Ballard, 1999). Predictive coding postulates that our perception of the world, under ambiguous sensory information, results from an interaction between our predictions, built from previous input, and the actual input from the environment (see also **Chapter 1**). In this framework, evoked responses correspond to prediction error that is explained away (within trial) by self-organising neuronal dynamics during perception and is suppressed (between trials) by changes in synaptic efficacy during perceptual learning. Therefore the MMN can be interpreted as a failure to suppress prediction error, which can be explained quantitatively in terms of coupling changes among and within cortical regions. The predictive coding framework encompasses the two distinct hypotheses, in the sense that it predicts the adjustment of a generative model of current stimulus trains (*cf. the model-adjustment hypothesis*) by using plastic changes in synaptic connections (*cf. the adaptation hypothesis*). The repeated presentation of standards may render suppression of prediction error more efficient; leading to a reduction in evoked responses under repetition and the emergence of a mismatch response, when an unlearned stimulus is presented.

The results described in this chapter support the hypothesis that the MMN reflects a failure to suppress prediction error, which can be framed in terms of predictive coding. Yet, the work described in Summerfield et al. (2006), suggests that this principle could presumably be extended to other sensory modalities in general such as vision. The predictive coding framework hypothesises that increases in intrinsic coupling encode progressive increases in the estimated precision of top-down predictions, which are responsible for suppressing prediction error. These changes in lateral interactions could be mediated by adaptation-like mechanisms within the auditory cortex to repeated sounds. Changes in forward connections may reflect changes in sensitivity to prediction error that is conveyed to higher levels. These higher levels form predictions so that backward connections can provide contextual guidance to lower levels. In this view, the MMN represents a failure to predict bottom-up input and consequently a failure to suppress prediction error. The requisite change in architecture, during the implicit learning of standards, is expressed in terms of quantifiable coupling changes among and within cortical regions.

5.5 Summary

The MMN, a well characterised response to violations in the regularity of an auditory structured sequence, is one of the most widely studied evoked responses. There have been several compelling mechanistic accounts of how the MMN might arise. It has been suggested that the MMN results from a comparison between sensory input and a memory trace of previous input – *the model adjustment hypothesis* – but others have argued that local adaptation due to stimulus repetition, is sufficient to explain an apparent MMN – *the adaptation hypothesis*. Thus, the precise mechanisms underlying the generation of the MMN remain unclear. This chapter tested biologically plausible mechanistic models of the MMN, which attempted to map onto these competing hypotheses. Comparison of DCMs revealed that the MMN is generated by a multi-level hierarchical network through coupling changes within and between cortical areas. This shows that both hypotheses, adaptation and model adjustment, operate in concert, and furnishes evidence that information processing in the brain is consistent with predictive coding.

Chapter 6

Repetition suppression and cortico-cortical plasticity in the human brain

The previous chapter tested different biologically plausible networks that attempt to map onto the mechanistic hypothesis pertaining to the MMN, *adaptation, model-adjustment* and *predictive coding.* It was shown that the MMN is best explained in the light of predictive coding, a general and unifying framework that encompasses both *adaptation* (Jääskeläinen et al., 2004) and *model-adjustment* (Winkler et al., 1996; Näätänen and Winkler 1999). The study described in this chapter had two aims: (i) to replicate these results with a different data set, elicited by a roving oddball paradigm; and (ii) to investigate the mechanisms underlying neuronal dynamics elicited by auditory stimulus repetition, particularly with regard to learning-related changes in brain connectivity. In contrast to traditional oddball paradigms, the roving paradigm used here is characterised by a continuously changing standard stimulus. This allows one to investigate *how a deviant becomes a standard*, while it retains the same physical properties. The previous studies described in **Chapters 3-5** used a categorical approach, modelling only one ERP (**Chapters 4**) or two (**Chapters 3 and 5**). In contrast, the modelling approach adopted here used weighting functions incorporated into a single DCM, which tried to explain multiple ERPs (across repetitions) in a parametric fashion. This approach provides information about the evolution of the connectivity parameters in the cortical network, as a function of repetition or learning. Moreover, it links auditory perceptual learning with repetition-dependent plasticity in the human brain.

6.1 Introduction

Few studies have explicitly explored the role of stimulus repetition during auditory memory-trace formation. Näätänen and Rinne (2002) found that later negative components, in contrast to earlier responses, are elicited only by sound repetition. Others found that increasing the number of standard tone repetitions induces enhanced activity of both early (30-50 ms) and later components (60-75 ms) (Dyson et al., 2005), localised in the primary and secondary auditory areas respectively (Liegeois-Chauvel et al., 1994). Similarly, Baldeweg et al. (2004) and Haenschel et al. (2005) found that the MMN increases with increasing number of standards and that this MMN is mediated by repetition-dependent enhancement of a slow positive wave (50-250 ms) in the standard ERP.

In the predictive coding framework (see also Friston, 2005; Baldeweg, 2006), evoked responses, corresponding to prediction error, drive perceptual inference and changes in synaptic efficacy (between trials) such that prediction error is suppressed with learning. Here, a roving MMN paradigm is used to test the hypothesis that the suppression of the MMN response to repetitive stimuli is due to plastic changes in connectivity. It is shown that learning the acoustic environment, through stimulus repetition, reduces effective connectivity, within and between cortical areas. This causes learning-induced decreases in prediction error which manifest as a suppression of the MMN as an oddball becomes a standard.

6.2 Methods and Statistical analysis

6.2.1 Experimental design

Twelve healthy volunteers aged 24–34 (4 female), gave signed informed consent before the study, which proceeded under local ethical committee guidelines. Subjects sat in front of a desk in a dimly illuminated room. Electroencephalographic activity was measured during an auditory roving oddball paradigm (see **Figure 6.3a**). The stimuli comprised a structured sequence of pure sinusoidal tones, with a roving, or sporadically changing tone. This paradigm resulted from few modifications to that used in Haenschel et al. (2005), originally designed by Cowan et al. (1993). Within each stimulus train, tones were of one frequency and were followed by a train of a different frequency. The first tone of a train was a deviant, which became a standard after few repetitions. This means deviants and standards have exactly the same physical properties, differing only in the number of times they have been presented. The number of times the same tone was presented varied pseudo-randomly between one and eleven. The frequency of the tones varied from 500 to 800 Hz in random steps with integer multiples of 50 Hz. Stimuli were presented binaurally via headphones for 15 minutes. The duration of each tone was 70 ms, with 5 ms rise and fall times, and the inter-stimulus interval was 500 ms. About 250 deviant trials (first tone presentation) were presented to each subject. The remaining trials included in this analysis had about 250 to 150 occurrences. Each subject adjusted the loudness of the tones to a comfortable level, which was maintained throughout the experiment. The subjects were instructed to ignore the sounds and performed the same distracting visual task as described in the previous chapter (see **Chapter 5**).

6.2.2 Data acquisition and pre-processing

Data acquisition and pre-processing procedures were similar to those adopted in the previous chapter (see **Chapter 5**), in terms of filtering, re-referencing, down-sampling, and artefact rejection procedures. In this study the data were low pass filtered at 30 Hz, which did not affect much the ERP forms, and is somewhat irrelevant for the analysis of the components of interest since their frequency is well below this. Trials were sorted in terms of tone repetition. In other words, trials one to eleven correspond to the responses elicited after one to eleven presentations of the same tone, collapsed across the whole range of frequencies. Trial one is the oddball, or the deviant trial. Averaged responses were obtained using robust averaging (as implemented in SPM5). Two subjects were excluded from the analysis due to artefacts and another two due to an undetectable MMN (on visual inspection of the scalp data). Data were transformed into scalp-map images (see **Figure 6.4a**). These were obtained after linear interpolation and smoothing (at FWHM 6:6:4 mm:mm:ms) of the difference wave response between the first presentation and the sixth presentation. For computational expediency, DCMs (see below) were computed on a reduced form of data that corresponded to eight channel mixtures or spatial modes. These were the eight principal modes of a singular value decomposition (SVD) of the channel data between 0 and 250 ms, over trial types of interest. In the first part of this study, where deviants and standards were analysed, the use of eight principal eigenvariates explained on average 80% of the variance in the data across the group (and more than 69% of the data in every subject; see **Figure 6.1**). For the multi-trial DCM, where five consecutive trials were analysed, these eight SVD components preserved on average 76% of the variance of the data across the group (and more than 66% in every subject).

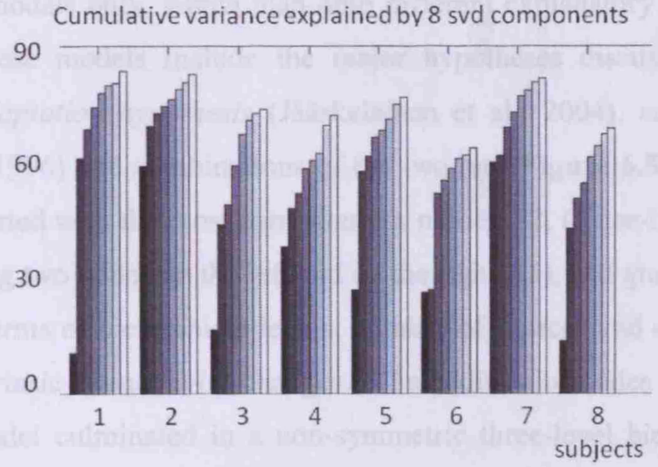


Figure 6.1 Cumulative variance explained by 8 svd components.

6.2.3 DCM specification: hypotheses tested

In the previous chapter, DCM was used to explore connectivity network models for explaining what causes the MMN (**Chapter 5**; Garrido et al., in submission). The research performed in this chapter investigates whether the same model underlies the MMN elicited in a roving paradigm and how tone repetition is expressed in terms of connectivity changes in the ensuing cortical network. As mentioned before, DCM does not explore all possible models but tests specific mechanistic hypotheses defined in terms of specific connectivity models. Bayesian model selection of DCMs can provide evidence in favour of one model relative to others. The chosen network architectures were motivated by the results of previous studies of MMN generators (Rinne et al., 2000; Opitz et al., 2002; Doeller et al., 2003; Grau et al., 2007; Garrido et al., 2007a). **Figure 6.2** shows the prior locations for nodes in the DCM graphs, which include the areas found to be active during an MMN: bilateral **A1** and **STG** and right **IFG**. See also **Table 3.1** in **Chapter 3** for the coordinates on these locations in MNI space (mm). The mechanistic models formulated here attempt to explain the generation of each individual response (i.e., responses to each tone presentation). These models are similar to those tested in the previous chapter of this thesis (see **Chapters 5**, **section 5.2.4** for details on model specification; see also Garrido et al., 2007a; in submission). However, the model search in this chapter focused on six models only, which map onto different explanatory mechanisms for the MMN. These models include the major hypotheses discussed in previous chapters: *the adaptation hypothesis* (Jääskeläinen et al., 2004), *model adjustment* (Winkler et al., 1996) and combinations of the two (see **Figure 6.5a**). In brief, the model search started with the most parsimonious model, **S2**, (a one-level hierarchical model comprising two nodes on the left and on the right **A1**), and gradually increased complexity; in terms of hierarchical levels, number of sources and manipulations of intrinsic and extrinsic connectivity changes. The addition of nodes and connections to the initial model culminated in a non-symmetric three-level hierarchical model that included bilateral **A1** and **STG**, and right **IFG**. All models can therefore be considered as a sub-model of the last one, model **S5i**).

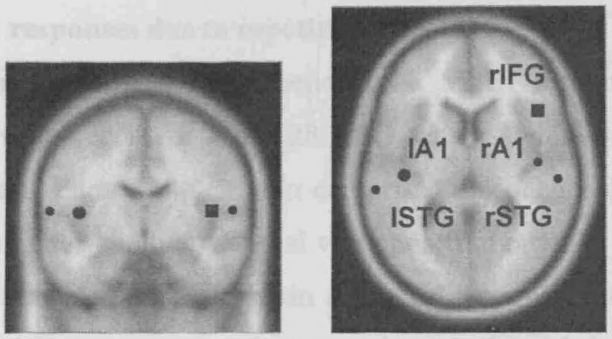


Figure 6.2 Prior locations for the nodes in the models. Sources of activity were modelled as equivalent dipoles. Their prior mean locations: **lA1** [-42, -22, 7], **rA1** [46, -14, 8], **lSTG** [-61, -32, 8], **rSTG** [59, -25, 8], **lIFG** [-46, 20, 8], **rIFG** [46, 20, 8] in mm are superimposed in an MRI of a standard brain in Montreal Neurological Institute (MNI) space.

6.2.4 Statistical analysis: Bayesian Model Comparison

The statistical analysis in this chapter uses the procedures described in **Chapter 2**, for Bayesian model comparison at both single-subject and group levels. See **section 2.2** for further details. The next subsection describes the inference made on the parameters of the best model pooled over subjects, i.e., at the group level.

6.3 Results

The results of this chapter show that learning the acoustic environment through stimulus repetition reduced connectivity, within and between hierarchically organised cortical areas. This analysis comprised three parts: (i) confirmation that there is a significant difference response (MMN) between the first and sixth tone presentation; (ii) hypotheses or model testing to establish the best DCM; and (iii) analyses of plasticity, in terms of connectivity strengths that change as a function of tone repetition.

6.3.1 Mismatch responses due to repetition effects

An initial analysis confirmed the presence of a MMN response in this roving paradigm. Data were recorded from 128 EEG sensors while subjects listened to trains of pure tones. Each stimulus train comprised a sequence of tones, presented every 500ms. This inter-stimulus interval was chosen in order to ensure that simple pre-synaptic facilitation could not explain any short-term plasticity observed. Each tone was presented between one and eleven times before changing in frequency. The first presentation of a tone with a different frequency from the preceding tone was defined as a deviant (see **Methods and Statistical analysis** and **Figure 6.3a** for details on experimental design). **Figure 6.3b** shows the grand mean responses (over subjects) to first tone presentation; the deviant or oddball trial (gray), and responses to the sixth presentation (black), when it is assumed that a standard response has been attained. This assumption is based on the ERP waveforms shown in **Figure 6.3c** for the first five presentations. These data came from a fronto-central electrode (C21), where the MMN was most evident. This MMN response was found over frontal and temporal electrodes, peaking at about 180 ms from change onset, which is consistent with previous studies (Cowan et al., 1993; Baldeweg et al., 2004).

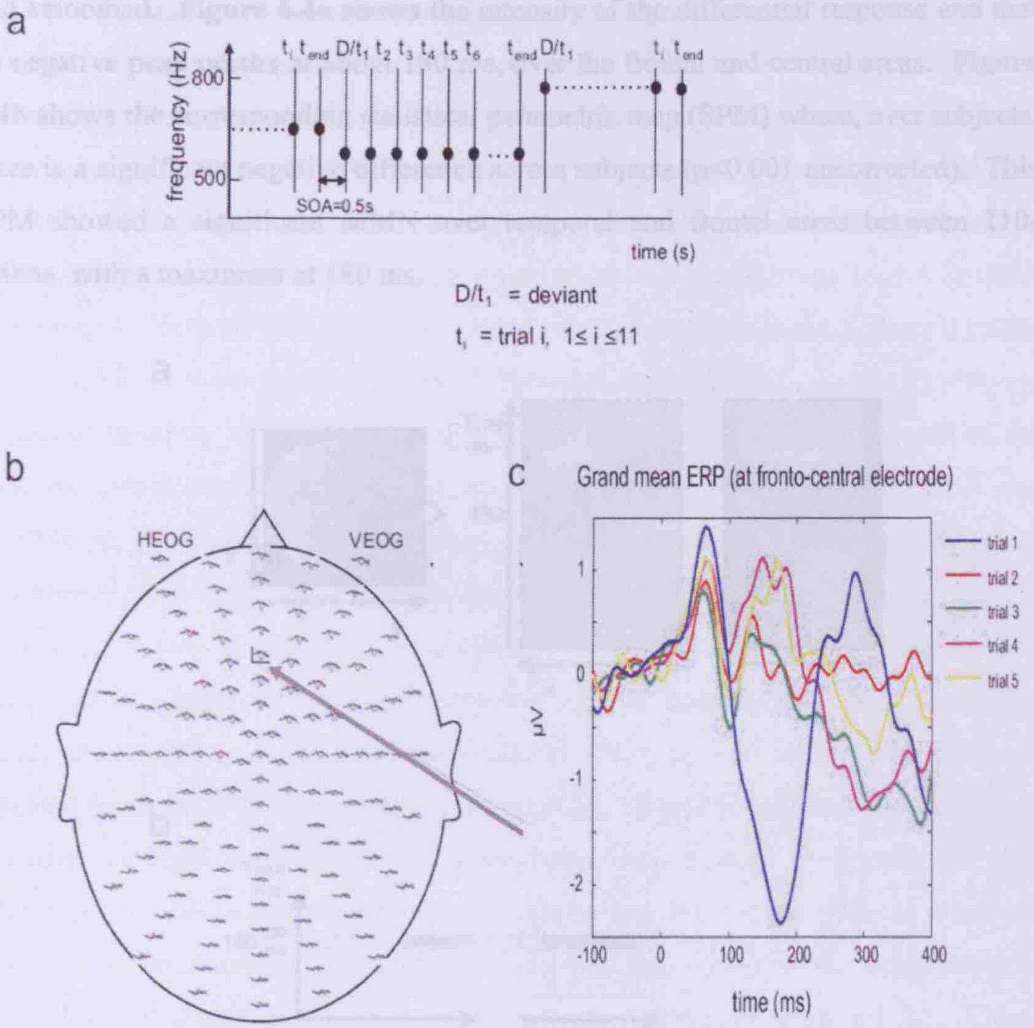


Figure 6.3 Design and responses elicited in a roving paradigm. (a) Stimulus design is characterised by a sporadically changing standard stimulus. The first presentation of a novel tone is a *deviant* $D = t_1$ that becomes a *standard*, through repetition $(t_2,\ldots,t_{end})$. However, in this paradigm, deviants and standards have exactly the same physical properties. **(b)** Grand mean (averaged over subjects) ERP responses to the sixth tone presentation, the established "standard" (t_6 in black) and deviant tone (t_1, in gray) overlaid on a scalp-map of 128 EEG electrodes. **(c)** enlarged ERP responses to the standard and deviant tones at channel C21 (fronto-central) where the MMN response peaks at about 180 ms from change onset.

Figure 6.4 shows a 3D spatiotemporal characterisation of the grand mean difference wave response, using conventional statistical parametric mapping to compare the first and the sixth presentations, the *deviant* and the *standard*, respectively. This analysis searched for differences over 2D sensor-space and all peri-stimulus time [-100, 400]. The scalp topography at any time-bin was interpolated from 128 channels and smoothed. **Figure 6.4a** shows the intensity of the differential response and that its negative peak occurs at about 180 ms, over the frontal and central areas. **Figure 6.4b** shows the corresponding statistical parametric map (SPM) where, over subjects, there is a significant negative difference across subjects ($p<0.001$ uncorrected). This SPM showed a significant MMN over temporal and frontal areas between 110-200ms, with a maximum at 180 ms.

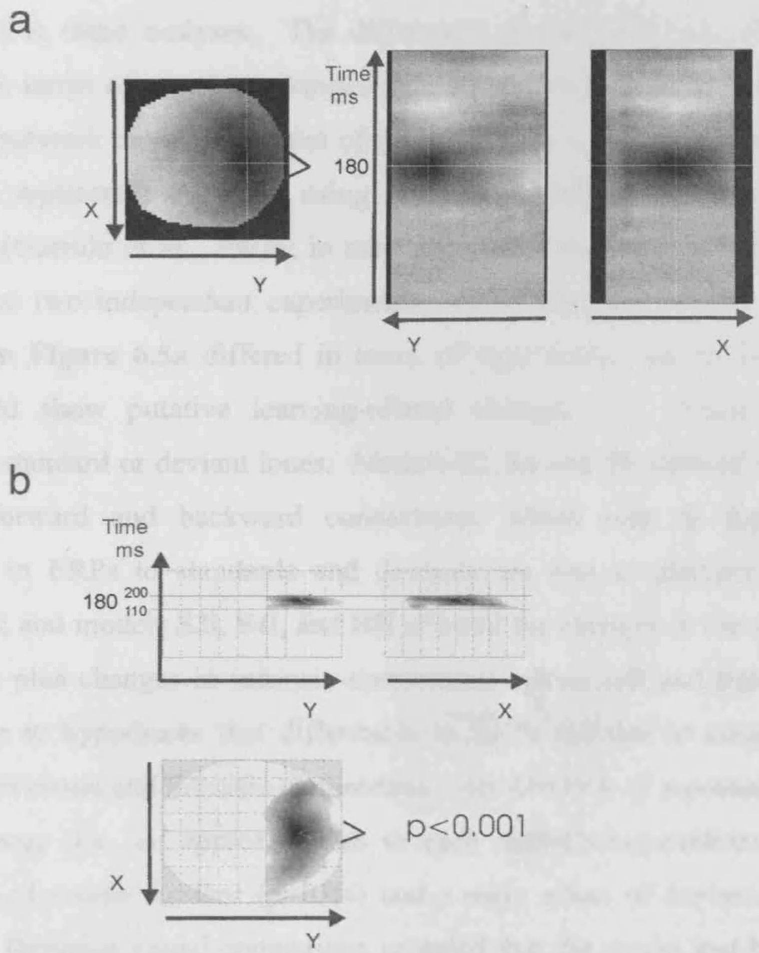


Figure 6.4 3D-spatiotemporal SPM analysis of the grand mean difference between the first presentation and the sixth presentation at the between-subject level. The measurement space corresponds to a 2D-scalp topography (interpolated from the 128 channels) and peri-stimulus time (-100 to 400). **(a)** Differential response with a negative peak at about 180 ms. **(b)** SPM showing areas where there is a significant negative difference across subjects ($p<0.001$ uncorrected). Significant effects were found over temporal and frontal areas in the range of 110 to 200 ms peaking at 180 ms (see marker). Significant positive effects (not shown) were found in the time window of 250-350 ms, which correlate with P300.

6.3.2 Hypotheses testing –underlying connectivity models of the MMN

Next, six different hierarchical models were tested. These models represent specific mechanistic hypotheses about MMN generation: *adaptation* (mapped onto **S2i**-model), *model-adjustment* (**S4**- and **S5**-models) and *predictive coding* (**S4i**- and **S5i**-models). Model **S4i** can also be regarded as an *adaptation* model, if the differences in the ERPs to standards and deviants are driven by modulations in the intrinsic connections. Both responses, ERPs to standards and deviants, were explained by the same model in these analyses. The differences in the ERPs, i.e., the MMN, are explained in terms of coupling changes within and among the cortical areas of the underlying network model. The aim of these analyses was to assess whether it was possible to replicate the results using classical oddball paradigms described in **Chapter 5** (Garrido et al., 2007a; in submission). Indeed, the best model was the same for the two independent experiments, model **S5i** (see below). The models illustrated in **Figure 6.5a** differed in terms of their nodes and in the connections which could show putative learning-related changes, *i.e.*, differences between listening to standard or deviant tones. Models **S2**, **S4** and **S5** allowed for changes in extrinsic, forward and backward connections, which map to hypotheses that differences in ERPs to standards and deviants are due to plasticity in extrinsic connections; and models **S2i**, **S4i**, and **S5i** allowed for changes in the same extrinsic connections plus changes in intrinsic connections within left and right **A1**. These models map to hypotheses that differences in ERPs are due to conjoint coupling changes in extrinsic and intrinsic connections. An ANOVA of repeated measures on the free-energy (i.e., an approximation to each model's log-evidence) revealed a main effect of *source number* ($p<0.04$) and a main effect of *intrinsic connectivity* ($p<0.001$). Bayesian model comparison revealed that the model that best explained the data is model **S5i**, a three-level network composed of bilateral **A1** and **STG** and right **IFG** (see **Figure 6.5b**). This replicates the result obtained in **Chapter 5** (see also Garrido et al., in submission). For the winning model **S5i**, a post hoc t-test confirmed a significant coupling decrease ($p<0.003$) for the backward connection linking **rIFG** to **rSTG** and a trend increase ($p<0.1$) for the intrinsic connection within **rA1** and the forward connection linking **lA1** to **lSTG**.

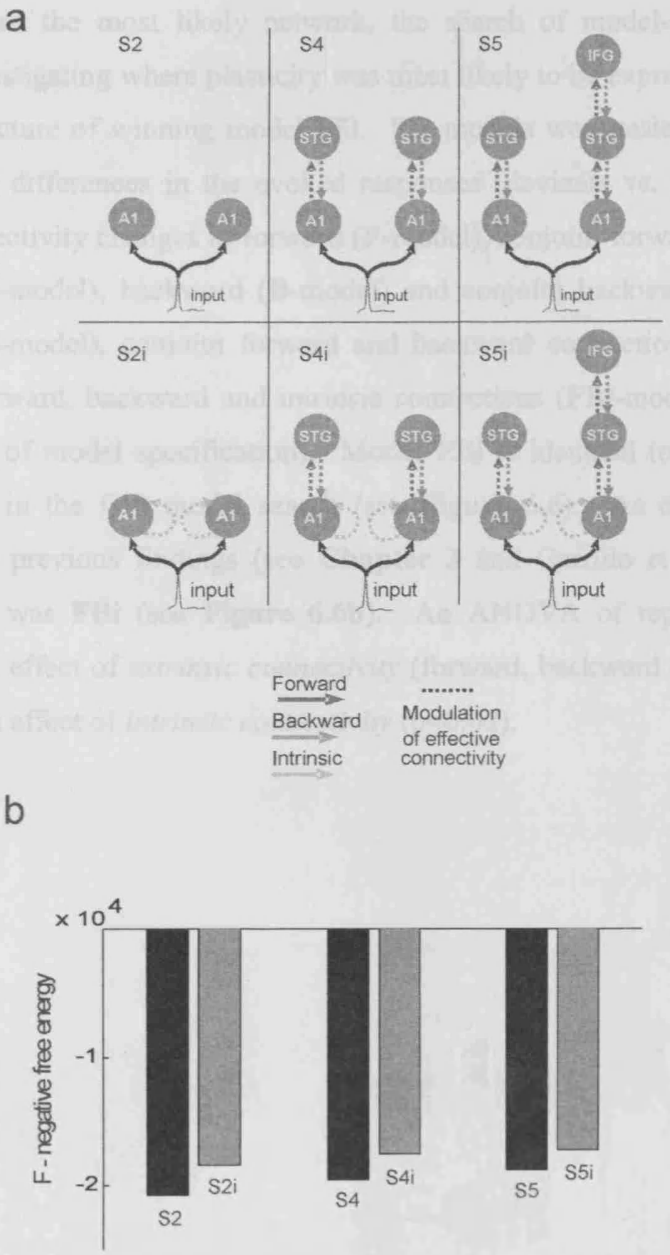


Figure 6.5 Model specification and Bayesian model comparison for the six networks tested. (a) The sources comprising the networks: **A1**: primary auditory cortex, **STG**: superior temporal gyrus and **IFG**: inferior temporal gyrus are connected with forward (dark grey), backward (grey) and intrinsic (light grey) connections. The first row of models, **[S2, S4, S5]**, allowed for learning-related changes in only extrinsic (forward and backward), while the second row **[S2i, S4i, S5i]** allowed for conjoint changes in extrinsic and intrinsic connections. Each column is filled with two similar network models, which differ only in allowing for modulations of intrinsic connectivity within **A1**. The columns differ in the number of active sources, which were reciprocally connected (with forward and backward connections). **(b)** The graph shows the free-energy approximation to the log-evidence at the group level, *i.e.*, pooled over subjects, for the six models. The best model is a 3-level hierarchical network, comprising five interconnected cortical areas allowing for local adaptation within primary auditory cortices and plastic changes in extrinsic connections (model- **S5i**).

Having identified the most likely network, the search of model-space was then finessed by investigating where plasticity was most likely to be expressed; within the network architecture of winning model **S5i**. Six models were tested, encoding the hypotheses that differences in the evoked responses (deviants vs. standards) were caused by connectivity changes in forward (**F**-model), conjoint forward and intrinsic connections (**Fi**-model), backward (**B**-model) and conjoint backward and intrinsic connections (**Bi**-model), conjoint forward and backward connections, (**FB**-model), and conjoint forward, backward and intrinsic connections (**FBi**-model) (see **Figure 6.6a** for details of model specification). Model **FBi** is identical to model **S5i**, the winning model in the first model search (see **Figure 6.6**). As expected, and in agreement with previous findings (see **Chapter 3** and Garrido et al., 2007a) the winning model was **FBi** (see **Figure 6.6b**). An ANOVA of repeated measures revealed a trend effect of *extrinsic connectivity* (forward, backward or both) (p<0.1) and a significant effect of *intrinsic connectivity* (p<0.02).

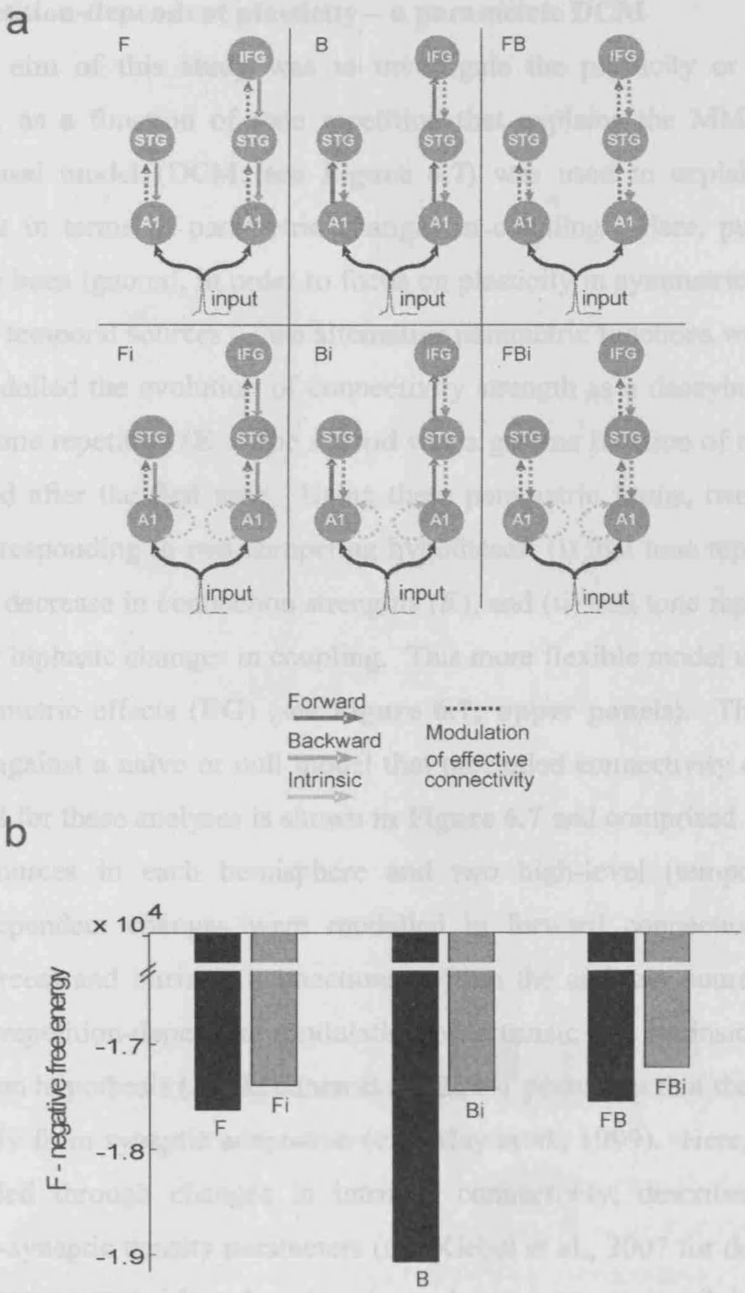


Figure 6.6 Model specification and Bayesian model comparison for the six variants of model S5i. **(a)** Six models comprising three hierarchical cortical levels. Bilateral **A1** are reciprocally connected with bilateral **STG**, and right **STG** is reciprocally connected with right **IFG**. The first row of models, [**F**, **B**, **FB**], allowed for learning-related changes in only extrinsic connections: forward, backward and conjoint forward and backward connections, respectively. The second row [**Fi**, **Bi**, **FBi**] allowed for conjoint extrinsic and intrinsic (within **A1**) connections. Each column shows similar network models, which differ only in allowing for changes of intrinsic connectivity within **A1**. **(b)** The graph shows the free energy approximation to the log-evidence at the group level, *i.e.*, pooled over subjects, for the six models. The best model is **FBi** which allows for modulations of all extrinsic and intrinsic connections. This is in fact exactly the same as the parent model S5i in which all connections could change. Models **Fi** and **FBi** are better than **Bi**.

6.3.3 Repetition-dependent plasticity – a parametric DCM

The second aim of this study was to investigate the plasticity or dynamics of connectivity, as a function of tone repetition that explains the MMN above. A dynamic causal model (DCM; see **Figure 6.7**) was used to explain differences among ERPs in terms of parametric changes in coupling. Here, putative frontal sources have been ignored, in order to focus on plasticity in symmetrically deployed auditory and temporal sources. Two alternative parametric functions were evaluated. The first modelled the evolution of connectivity strength as a decaying exponential function of tone repetition (**E**). The second was a gamma function of repetition (**G**), which peaked after the first tone. Using these parametric forms, two DCMs were inverted, corresponding to two competing hypotheses: (i) that tone repetition causes a monotonic decrease in connection strengths (**E**), and (ii) that tone repetition causes 'one-shot' or biphasic changes in coupling. This more flexible model used a mixture of both parametric effects (**EG**) (see **Figure 6.7; upper panels**). The two DCMs were tested against a naïve or null model that precluded connectivity changes. The network used for these analyses is shown in **Figure 6.7** and comprised two low-level (auditory) sources in each hemisphere and two high-level (temporal) sources. Repetition-dependent changes were modelled in forward connections (from the auditory sources) and intrinsic connections (within the auditory sources), allowing for separate repetition-dependent modulation of extrinsic and intrinsic connections. The adaptation hypothesis (Jääskeläinen et al., 2004) postulates that the MMN arises predominantly from synaptic adaptation (*c.f.*, May et al., 1999). Here, these effects were modelled through changes in intrinsic connectivity, described by source-specific post-synaptic density parameters (see Kiebel et al., 2007 for details). These effects could be mediated by adaptation (*e.g.*, due to increase in calcium-dependent potassium conductances, leading to after-hyperpolarizing currents; Powers et al., 1999) or subsequent calcium-dependent intracellular mechanisms that underlie phenomena like paired-pulse depression (*e.g.*, Davies et al., 1990). Putative short-term changes in the synaptic efficacy of intrinsic afferents modify lateral interactions within primary auditory cortex. In the predictive coding framework, these encode the uncertainty of predictions. Similar changes in extrinsic (forward connections) correspond to perceptual learning or model adjustment (see Winkler et al., 1996;

Näätänen and Winkler, 1999; Friston, 2005). This DCM has been validated extensively in previous studies (Garrido et al., 2007a; b).

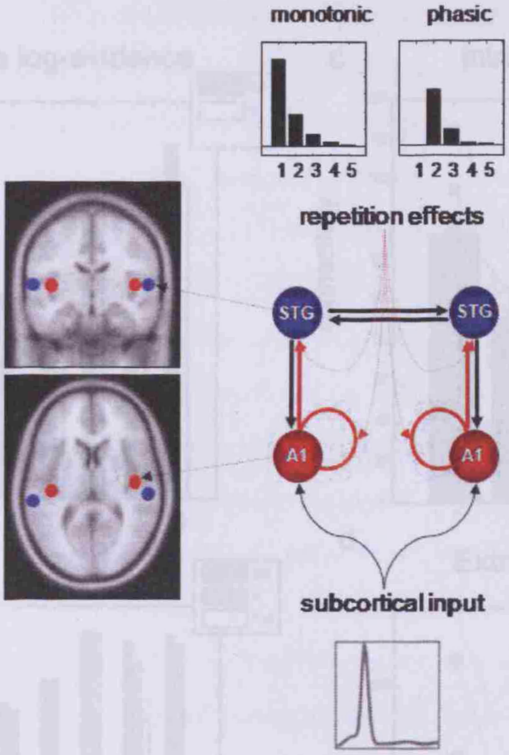


Figure 6.7 Model specification. (right) The sources comprising the networks: **A1**: primary auditory cortex, and **STG**: superior temporal gyrus are connected with forward (red), backward (black) and intrinsic (red) connections. **(left)** The locations of the sources in the models. Sources of activity were modelled as a mixture of eight local cortical basis functions over all dipoles within 16 mm of the source locations. Their prior mean locations: **lA1** [-42, -22, 7], **rA1** [46, -14, 8], **lSTG** [-61, -32, 8], **rSTG** [59, -25, 8], in mm are superimposed in an MRI of a standard brain in Montreal Neurological Institute (MNI) space. The DCM receives (parameterised) subcortical input at the **A1** sources, which elicit transient perturbations in the remaining sources. Repetition effects are modelled by changes in intrinsic and forward connections (red) that are a mixture of monotonic (**upper left**) and phasic (**upper right**) repetition-specific effects.

Bayesian model comparison revealed that the **EG** model supervened over the simple monotonic model **E,** in all but one subject; and in all subjects both parametric models were substantially better than the null model that precluded plasticity. This means that there is consistent evidence for changes in connectivity, above and beyond a simple exponential decay, in one or more connections (see **Figure 6.8A**). Figure 4B shows the equivalent results for accuracy expressed as the proportion of variance explained over channels and trials by each of the three models assessed (The **EG** model explained 80% of variance on average and at least 68% in each subject).

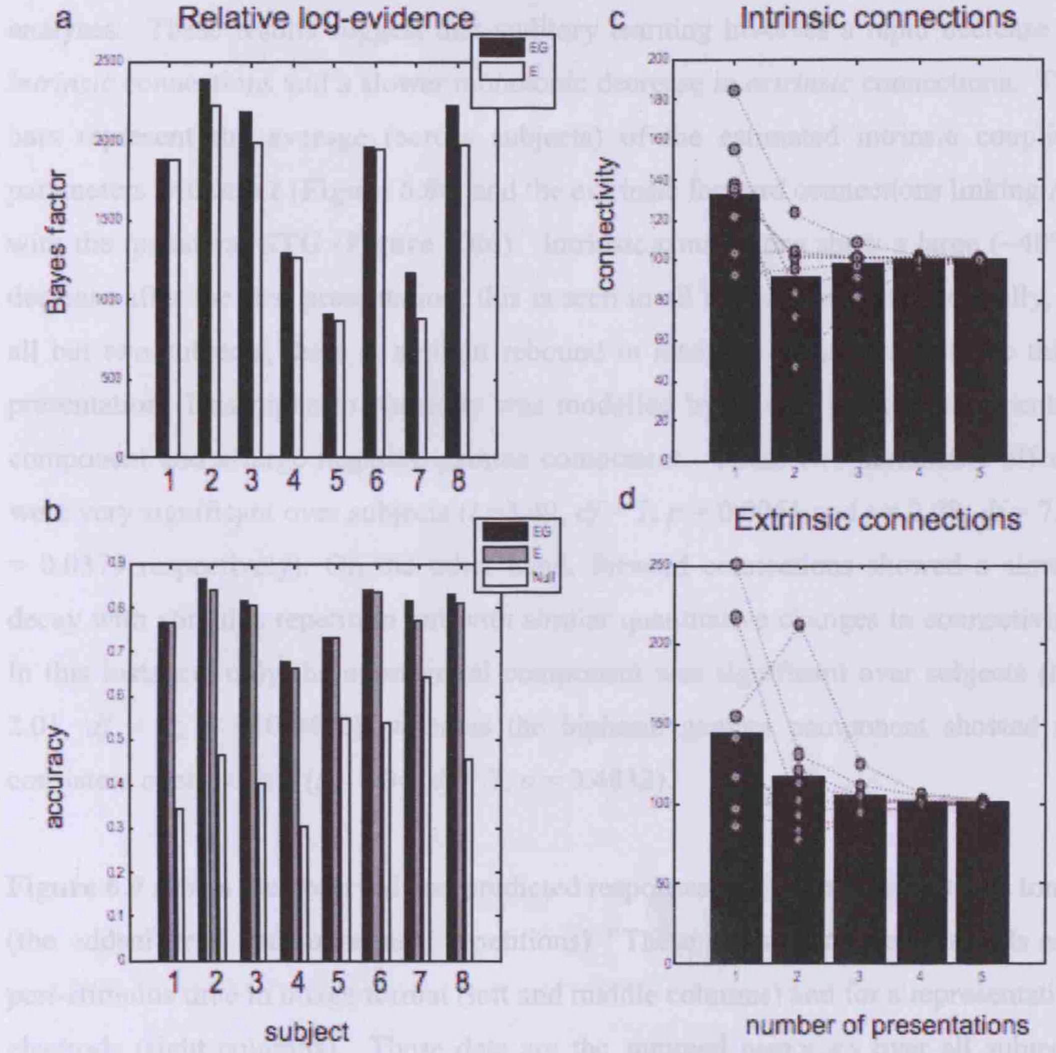


Figure 6.8 Model comparisons and conditional expectations for repetition-dependent connectivity changes as a function of learning. (a) Bayesian model comparison shows that model **EG** supervenes over model **E** in 7 out of 8 subjects; data from the first subject were best explained by the exponential (monotonic) model but this effect was trivial in relation to the log-evidence with respect to the null model. **(b)** Corresponding results for accuracy, expressed in terms of the proportion of variance explained by the model (*i.e.*, the coefficient of determination). **(c)** Connectivity reductions with repetition. This shows the temporal evolution of connectivity as a function of time, or repetition for the intrinsic connections within **A1**, expressed as the average conditional expectation over subjects (bars) and for each subject separately (circles). By design these repetition effects are normalised so that the connection strength is a percentage of strength after learning. There is a very consistent and marked decrease in coupling after the first presentation that appears to rebound slightly on subsequent presentations. **(d)** The same results for the extrinsic connections. Here the changes are expressed more slowly as a function of repetition, exhibiting a monotonic decrease over time.

Figure 6.8 also shows the changes in coupling strengths of the intrinsic (**Figure 6.8c**) and extrinsic (**Figure 6.8d**) connections. Here, the evolution of connectivity is shown as a function of repetition in terms of conditional expectations from the DCM analyses. These results suggest that auditory learning involves a rapid decrease in *intrinsic* connections and a slower monotonic decrease in *extrinsic* connections. The bars represent the average (across subjects) of the estimated intrinsic coupling parameters within **A1** (**Figure 6.8c**) and the extrinsic forward connections linking **A1** with the ipsilateral **STG** (**Figure 6.8d**). Intrinsic connections show a large (~40%) decrease after the first presentation; this is seen in all but two subjects. Critically, in all but two subjects, there is a slight rebound in intrinsic connectivity on the third presentation. This biphasic plasticity was modelled by a large positive exponential component and a large negative gamma component. These two parametric effects were very significant over subjects (t =3.49, df= 7; p = 0.0051 and t = 2.08, df= 7; p = 0.0379 respectively). On the other hand, forward connections showed a slower decay with stimulus repetition but with similar quantitative changes in connectivity. In this instance, only the exponential component was significant over subjects (t = 2.01, df = 7; p = 0.0422), whereas the biphasic gamma component showed no consistent contribution (t = 0.04, df= 7; p = 0.4832).

Figure 6.9 shows the observed and predicted responses elicited by the first five tones (the oddball trial and subsequent repetitions). These are shown over channels and peri-stimulus time in image format (left and middle columns) and for a representative electrode (right columns). These data are the summed responses over all subjects (after applying a Hanning window). The response to the first presentation or oddball shows a peak after 100ms (that subtends the N1 component) and an enhanced response with its maximum at about 180ms. Visual inspection of the scalp data (not shown) suggested that the later peak conforms to the spatial deployment of the MMN. The second and subsequent presentations elicit a response with a similar temporal profile, but the MMN component is greatly attenuated. This suggests that, after one presentation of a new tone, the brain has re-learned the auditory context; in other words, the "standard" is largely learned (*c.f.*, Baldeweg et al., 2004; Haenschel et al., 2005; Dyson et al., 2005).

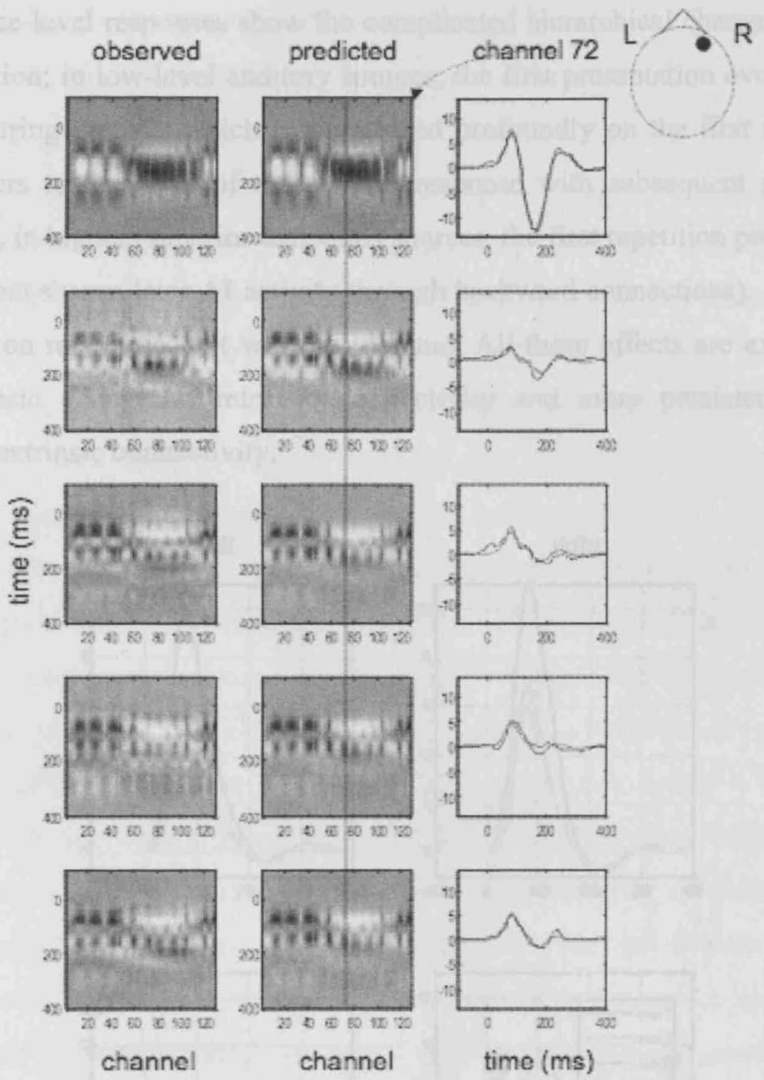


Figure 6.9 Predicted and observed responses in channel space, averaged over all subjects. (Hanning window from -100ms to 400ms): (left) image format responses over peri-stimulus time and channels for each of the five repetitions of a tone. Profound mismatch negativity is seen in the upper panels (first presentation) that disappear quickly to produce the standard response by the fifth presentation (**right column**). The predicted (red) and observed (black) responses for channel 72 are shown on the right for illustration. The agreement is self-evident. Responses to repeating tones show a decrease in the N1 component (peaking at about 100ms) and later in the MMN, which vanishes after two repetitions.

Figure 6.10 shows the reconstructed responses (summed over subjects) at the source level for bilateral **A1** and bilateral **STG**. The generators for the N1 component lie in **A1** but not at the level of **STG**. Activity peaking between 100 and 200ms is seen in **STG** that might underlie the MMN. This spatiotemporal dissociation of N1 and MMN generators is very reminiscent of the findings of Jääskeläinen et al. (2004).

These source-level responses show the complicated hierarchical changes in the ERP with repetition; in low-level auditory sources, the first presentation evokes a greater response during the N1, which is suppressed profoundly on the first repetition. It then recovers to the level of the oddball response with subsequent presentations. Conversely, in higher (superior temporal) sources, the first repetition produces a later response (that shapes later **AI** activity through backward connections). This again is suppressed on repetition, but with no rebound. All these effects are explained by a rapid biphasic change in intrinsic connectivity and more persistent monotonic changes in extrinsic connectivity.

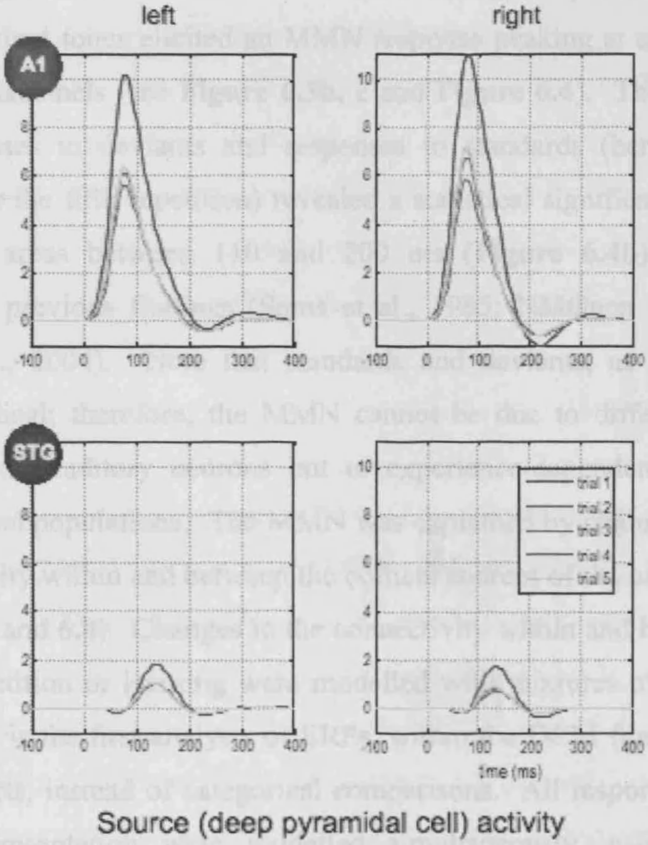


Figure 6.10 Reconstructed responses at the source level for bilateral primary auditory cortex (upper panels) and bilateral superior temporal gyrus (lower panels). These are the averages over all subjects. Right and left **A1** show peak activity at about 80ms that is suppressed to about half its amplitude after the first presentation; indeed it is suppressed so much that it recovers slightly on subsequent presentations. In bilateral **STG** peak activity is observed at about 140ms, which has the greatest amplitude on the first presentations. There is also a slight negative wave at about 50ms that may correspond to the response positivity in scalp space.

6.4 Discussion

6.4.1 Summary of findings

In summary, this study presents the first attempt to quantify plasticity underlying sensory-memory formation, caused by stimulus repetition with a network of interacting cortical areas using EEG. In this chapter, the effect of stimulus repetition on scalp electroencephalographic responses was investigated. Moreover, the underlying dynamics of the cortical network that generates these responses was explored. Subjects were presented with a roving paradigm, a modified auditory oddball paradigm with standard tones that changed sporadically to another frequency. Deviant tones elicited an MMN response peaking at about 180 ms over temporofrontal channels (see **Figure 6.3b, c** and **Figure 6.4**). The difference wave between responses to deviants and responses to standards (here assumed to be established after the fifth repetition) revealed a statistical significant negativity over temporofrontal areas between 110 and 200 ms (**Figure 6.4b**). This result is consistent with previous findings (Sams et al., 1985; Näätänen and Rinne, 2002; Baldeweg et al., 2004). Note that standards and deviants, as defined here, are physically identical; therefore, the MMN cannot be due to different responses in frequency-specific auditory neurons but to experience-dependent changes in the same neuronal subpopulations. The MMN was explained by changes in the strength of the connectivity within and between the cortical sources of the underlying network (see **Figure 6.7** and **6.8**). Changes in the connectivity within and between areas as a function of repetition or learning were modelled with mixtures of parametric basis functions. This is the first analysis of ERPs, within the DCM framework, that uses parametric effects, instead of categorical comparisons. All responses from the first to the fifth presentation were modelled simultaneously, using parameterised connectivity changes. Bayesian model comparison revealed that as the plasticity of the underlying cortical network unfolds, connection strengths show a progressive decrease with some connections exhibiting fast or biphasic changes (model **EG**, see **Figure 6.8**). Specifically, intrinsic connections within bilateral **A1** show a fast depression, followed by a slight rebound, whereas forward connections show a slower decay. These results suggest that perceptual learning, caused by stimulus repetition, can be explained by plasticity in intrinsic (adaptation) and extrinsic (model learning) brain connections. An interesting finding is that the MMN vanishes after one or two repetitions (see **Figure 6.9**), suggesting that the brain learns the context established by auditory trains within a second. These findings accord with Liegeois-Chauvel et al. (1994), who found that the generators of early components are distributed along **A1** and support the propagation hypothesis (Baldeweg, 2006): that a sensory memory trace can be detected earlier and earlier, at the level of **A1**, with an increasing number of repetitions. This is also consistent with the idea that stimulus-specific adaptation in **A1** contributes to the emergence of the MMN (Ulanovsky et al., 2003); although this modelling suggests that the adaptation effect is expressed vicariously through later responses in secondary or higher temporal sources. The decrease in inter-regional connection strengths over repetitions is consistent with predictive coding theories (Rao and Ballard, 1999; Friston, 2005). From this perspective, perceptual learning of the auditory context may be understood as a process of (between-trial) prediction error suppression, implemented neurophysiologically through changes in connection strengths within a hierarchical cortical network (Friston, 2005; Baldeweg, 2006).

6.4.2 Implications

The work described in this chapter tests the hypothesis that the MMN could be explained by repetition suppression of the sort that has been studied extensively in the visual system (Desimone, 1998). The relative contribution of plasticity in intrinsic (*i.e.*, adaptation) and extrinsic (*i.e.*, hierarchical learning) connectivity was assessed. It was found that, as anticipated, both exhibited repetition-dependent changes; however, the time-courses of these changes were surprisingly distinct. This speaks to distinct pre or post-synaptic mechanisms. This is important, both from the perspective of computational theories of sensory or perpetual learning and how these computations are implemented physiologically. The connection with paired-pulse paradigms used to study synaptic facilitation and depression (Davies et al., 1990) is self-evident and suggests that the relative time-scales of intrinsic and extrinsic plasticity could be characterised by varying the inter-stimulus interval in roving paradigms. Furthermore, combining this with pharmacological interventions is motivated easily by existing psycho-pharmacological studies of the MMN; see

Baldeweg et al. (2004) for a discussion of these studies, in relation to schizophrenia research.

6.4.3 Technical issues

A feature of DCM, and hypothesis driven methods in general, is that one cannot test all possibilities; i.e., one has to constrain the model-space in order to reduce it to a limited number of testable hypotheses. Two alternative models for the connectivity changes due to stimulus repetition were tested here. For simplicity, these hypotheses were tested under a symmetric network comprising bilateral primary and secondary auditory cortex (c.f., model **4Si**, see **Figure 6.5**). The true underlying network is probably more complex than this, and in fact, Bayesian model comparison revealed that the model that best explains these data rests upon a more complex architecture (see **Figures 6.5b and 6.6b**). The model chosen for this analysis is not the best amongst all possible, and yet, it still explains on average 76% of the variance of the data across the group (and a least 66% for every subject). Although a simple model, it was sufficient to make inferences on the connectivity changes and to recover consistent effects at the between-subject level. A similar result would be expected under model **S5i.** The search in model space does not offer an exhaustive exploration and selection of models; it only selects the best model amongst the models considered. Hence, there might be other models that explain connectivity changes during stimulus repetition or learning. This is a question of model comparison, and provided that there is good motivation for adding another model to the space of models, one can use Bayesian model comparison to evaluate any new model. This applies to alternative networks and alternative basis sets for the parametric effects of repetitions. An important consideration, when comparing DCMs, is that the free-energy takes into account both model accuracy and complexity. This allows for comparison of models with different numbers of parameters (Friston et al., 2006b).

Finally, as mentioned previously, the models considered here attempt to explain the dynamics during 0 to 250 ms, which includes the MMN and any other component peaking under these time limits, such as repetition positivity (RP), N1 or P3a. In addition, the parametric multi-trial DCM analysis attempts to explain the dynamics caused by five successive tone presentations. Hence, the Bayes factor, a measure of the goodness of one model relative to another, pertains to the dynamics evoked over multiple stimulus conditions.

6.4.4 Mechanisms of MMN generation

Mechanistic accounts of MMN generation posit changes in plasticity in extrinsic, forward and backward connections, or intrinsic (local) connections between and within hierarchical sources (Friston, 2005). In this context, the difference waveform (and the MMN) arises from changes in coupling within and among cortical sources. In previous chapters (**Chapters 3** and **5**; see also Garrido et al., in submission), DCMs were used to explain ERPs to standards and deviants, which tested alternative mechanisms or generative models for the MMN. For internal consistency, the same models were tested here, given new data. The choice of models was based on previous theoretical formulations of MMN generation, specifically *adaptation* (Jääskeläinen et al., 2004), *model-adjustment* (Winkler et al., 1996), and conjugations of the two, required by predictive coding. The predictive coding framework encompasses the two distinct hypotheses, in the sense that it predicts the adjustment of a generative model of current stimulus trains (*cf. the model-adjustment hypothesis*) combined with local changes in post-synaptic sensitivity (*cf. the adaptation hypothesis*). In agreement with the studies previously described (**Chapters 3** and **5**), the best model comprised five reciprocally connected sources (bilateral **A1** and **STG**, and right **IFG**). This is an important finding because it offers a comprehensive framework to explain the MMN; and furnishes direct evidence that the MMN is caused by self-organized changes in a cortical network with multiple hierarchical levels.

6.4.5 The MMN, a marker for auditory perceptual learning

Functionally, the MMN reflects error detection caused by an unexpected or unlearned event that follows the perceptual learning of standards. This has been formulated under hierarchical models of learning (Rao and Ballard, 1999; Friston

2003, 2005; Garrido et al., 2007a). In this framework, evoked responses correspond to prediction error that is explained away (within-trial) by neuronal dynamics during perception and is suppressed (between trials) by changes in connectivity during learning. Therefore, the MMN can be interpreted as a failure to suppress prediction error, which can be explained quantitatively in terms of coupling changes among and within cortical regions. The repeated presentation of tones leads to learning or establishing a representation of a standard that is accessed more efficiently. This renders suppression of prediction error more efficient, leading to a reduction in evoked responses and the emergence of a mismatch response, when novel and therefore unlearned stimuli are presented. The suppression of evoked responses, due to a repeated event, is a ubiquitous phenomenon in neuroscience. It is seen at the level of single-unit responses (where it is referred to as repetition suppression; Desimone, 1996) and is a long-standing observation in human neuroimaging (where it is often referred to as adaptation *e.g.,* cerebellar adaptation during motor repetitions; Friston et al., 1992 or repetition effects in visual studies; Henson et al., 2003). From an empirical Bayesian perspective (*c.f.*, predictive coding), modulations in intrinsic connectivity may encode changes in the precision of top-down predictions, responsible for suppressing prediction error. Changes in forward connections may reflect changes in prediction error that is conveyed to higher levels. These higher levels form predictions so that backward connections can provide contextual guidance to lower levels. In this view, the MMN represents a failure to predict sensory input and consequently a failure to suppress prediction error. The repetition-suppression of the MMN can be explained quantitatively in terms of coupling changes among and within cortical regions.

6.5 Conclusion

The key contribution of this work is to show that the plasticity underlying perceptual learning can occur very quickly and is effectively complete after a few presentations of a stimulus. Furthermore, the putative experience-dependent plasticity that underlies this learning (as observed electrophysiologically) involves distinct changes in intrinsic and extrinsic connections and, implicitly, distributed interactions among multiple sources.

6.6 Summary

The aim of this study was to investigate the mechanisms underlying neuronal responses elicited by repeated auditory stimuli, particularly with regard to experience-dependent changes in connectivity. Subjects were exposed to a roving oddball paradigm characterized by sporadic stimulus changes, while ERPs where recorded. A significant MMN response was found between the first (oddball) and the sixth presentation (standard), distributed over frontal electrodes across subjects and peaking at about 180 ms after change onset. Bayesian inversion of DCMs revealed systematic changes in both intrinsic and extrinsic connections, within a hierarchical cortical network, as a function of repetition. Intrinsic (within-source) connections showed biphasic changes that were much faster than changes in extrinsic (between-source) connections, which decreased monotonically with repetition. This study shows that auditory perceptual learning is associated with repetition-dependent plasticity in human brain connectivity.

Chapter 7

General Discussion and Conclusion

The previous four chapters described the empirical work pursued in this thesis. This chapter presents a summary of the work performed and discusses its general implications for perception. Predictive coding is proposed as a theoretical framework that explains perceptual learning and inference, in the particular context of the MMN paradigm. A critical assessment on the limitations of methodology used (DCM) is presented. Finally, directions for future research are considered.

7.1 Short synopsis

The work reported in this thesis attempted to test mechanistic hypotheses, framed in terms of connectivity models that explain the generation of evoked brain responses. These models were inspired by predictive coding and Bayesian inference, which state that perception involves processing current input from the environment, and predictions of this input. The working hypothesis of this thesis was that bottom-up flow of sensory input and top-down predictions have a biological substrate, which is mediated by synaptic efficacy or coupling strength. Therefore, perceptual learning should be associated with changes in forward and backward connections that link cortical areas and form a cortical hierarchy. This hypothesis was tested using responses to oddballs or unpredictable events, the MMN. DCM was employed to model the data, after establishing its reproducibility, using group analyses (**Chapter 3**). Bayesian model comparison was used to make inferences on model space and the connectivity parameters. The connectivity model that best explained the MMN responses, elicited in the diverse paradigms used, was consistent across subjects and studies (**Chapters 3-6**). This connectivity model is a multi-level hierarchical network that embodies coupling changes in forward, backward and intrinsic connections. Backward connections, that are associated with top-down effects, have been found to play a fundamental role in the expression of ERP and source activity, especially at long latencies (**Chapter 4**). Forward, backward and intrinsic connections were found to decrease in their strength as a function of stimulus repetition, reflecting successful encoding of a sensory regularity (**Chapter 6**).

7.2 Implications of this work: discussion

The aim of the work described in the first result chapter, **Chapter 3**, was to establish the validity of DCM by assessing its reproducibility across subjects. An oddball paradigm was used to elicit mismatch responses. Different connectivity models were specified to evaluate three different hypotheses: differences in the ERPs to deviant and standard events are mediated by changes in forward connections (**F**-model), backward connections (**B**-model) or both (**FB**-model). Bayesian inversion provided estimates of changes in coupling among sources and the marginal likelihood of each model. In all but one subject, the forward model (**F**-model) was better than the backward model (**B**-model). Furthermore, the **FB**-model was significantly better than both, in seven out of eleven subjects. At the group level the **FB**-model supervened. These findings are important because they establish the validity and usefulness of DCM in characterising EEG/MEG data and its ability to model ERPs in a mechanistic fashion (see also Garrido et al., 2007a). Moreover, this work offers a methodological approach to infer on models and their parameters, at the group level. However, DCM is not limited to answering questions related to the mechanisms underlying the MMN. DCM is a robust and powerful tool that can be used, in principle, to test mechanistic hypotheses underlying any physiologic processes in any sensory modality, such as face processing (David et al., 2006) and somatosensory evoked potentials (Kiebel et al., 2006). An interesting application would be to use it in the context of cross modal integration or in higher-level cognitive processes.

Chapter 4 presents the work performed on the investigation of the role of backward connections on the generation of late ERP components. DCMs of responses to deviants, elicited in an oddball paradigm were modelled with and without backward connections. Bayesian model comparison revealed that the model with backward connections explains the data better than the model without backward connections, especially at long latencies. Hence, this result supports the hypothesis that evoked brain responses are generated by recurrent dynamics among levels of cortical hierarchies. In addition, it was demonstrated that backward connections play a fundamental role in ERP generation, especially for later components, which are expressed at both the cortical and scalp level (see also Garrido et al., 2007b). This is consistent with top-down processing in perception and cognition. Importantly, it supports the working hypothesis of this thesis; i.e., that perception can be framed in terms of predictive coding. Moreover, the results prove that connection strength is a physical substrate or parameter that can be associated with learning and inference in the brain. A point of controversy though, is the inductive logic that any late ERP component involves backward connections. The rationale here is based on an exemplar response, an ERP to an auditory oddball. Ideally, one would corroborate this conclusion with ERPs elicited within different paradigms and in other sensory modalities. Yet, there is evidence from vision research that not every process in the brain engages backward connections. A possible explanation is that only forward connections are engaged in the context of explicit information, i.e., when bottom-up cues are sufficiently clear and unambiguous (Yuille and Kersten 2006). A similar conjecture has been proposed for the role of LGN (lateral geniculate nucleus) neurons in the context of thalamo-cortical feedback (Mumford, 1991).

Chapter 5 tested and compared different accounts of the mechanisms underlying the MMN using dynamic causal models (see also Garrido et al., in submission). The range of models tested covered (i) the *adaptation* hypothesis, which proposes that the MMN is best explained by a deviant-induced interruption of neuronal adaptation that is confined to lower-order auditory areas (*cf.* Jääskeläinen et al., 2004); (ii) the *model-adjustment* hypothesis (Winkler et al., 1996, Doeller et al., 2003), which assumes that the MMN results from deviant-induced changes in temporo-frontal connections, i.e., short-term synaptic plasticity; and (iii) combinations of these two hypotheses, which accommodate intra-areal adaptation combined with plasticity of inter-areal connections. The latter group of models are consistent with the *predictive coding* formulation (Friston, 2005; Baldeweg, 2006). These biologically plausible hypotheses for the MMN generation were mapped onto mechanistic models or DCMs that attempted to model responses to standards and deviants elicited in a classical auditory oddball paradigm. Bayesian model comparison revealed that the MMN is generated by a multi-level hierarchical network, through coupling changes within and between cortical areas. This shows that both mechanisms, adaptation and model adjustment, operate in concert, and furnishes evidence that information processing in the brain is consistent with predictive coding. The results of model comparison support the idea that the MMN cannot be attributed to a mechanism of local adaptation in the primary auditory cortex alone. In other words, neuronal adaptation *per se* is not sufficient to explain the MMN. If our results indicate that the adaptation hypothesis is not sufficient to explain MMN generation, nor do they favour model adjustment alone. In other words, the MMN can not be explained by connectivity changes in intrinsic connections, only, nor can it be explained by changes in extrinsic connections, only. The results reported in this thesis support the idea that the MMN rests upon a more complex mechanism. The mechanisms of MMN generation involve plasticity of inter-areal connections amongst multiple hierarchical levels, as well as local adaptation within the primary auditory cortices. This result is important because it combines both the *model adjustment hypothesis* (Winkler et al., 1996) and the local *adaptation* hypothesis (Jääskeläinen et al., 2004) into the unified and more general framework of predictive coding. Moreover it can accommodate the findings of a multitude of studies showing that there are temporal and frontal cortical sources underlying the MMN generation (Rinne et al., 2000; Jemel et al., 2002; Opitz et al., 2002; Doeller et al., 2003; Liebenthal et al., 2003; Molholm et al., 2005; Restuccia et al., 2005). Crucially, this result promotes a better understanding of the mechanisms that subtend the MMN, which have been a topic of great interest and debate for many years.

Chapter 6 investigated the mechanisms underlying neuronal dynamics elicited by auditory stimulus repetition, particularly with regard to learning-related changes in brain connectivity. This experiment used a roving paradigm, characterized by continuously changing standard stimuli, which enabled memory formation to be tracked (Cowan et al., 1993; Baldeweg et al., 2004; Haenschel et al., 2005). The results show that repeated stimuli, that share exactly the same physical properties, give rise to different brain responses. This suggests that only perception or the internal representations change, which is likely to be caused by learning. As in the previous experiment (**Chapter 5**), a set of mechanistic models were tested. These DCMs mapped onto different accounts of the mechanisms underlying the MMN generation: *adaptation*, *model-adjustment* and *predictive coding*. The results of this experiment lead to similar conclusions; namely, the mechanisms underlying the MMN involve plasticity of inter-areal connections amongst multiple hierarchical levels, as well as local adaptation within the primary auditory cortices. Bayesian inversion of a parametric multi-trial DCM revealed modulations of connections within a temporo-frontal cortical network. Over tone repetition these connections showed a bi-exponential decrease of plasticity. This study shows that perceptual learning is associated with connectivity changes in the brain. The suppression of evoked responses, to a repeated event, is a ubiquitous phenomenon in neuroscience. It is seen at the level of single-unit responses (where it is referred to as repetition suppression; Desimone, 1996) and is a long-standing observation in human neuroimaging (where it is often referred to as adaptation *e.g.,* cerebellar adaptation during motor repetitions; Friston et al., 1992 or repetition effects in visual studies; Henson et al., 2003).

7.3 Predictive Coding: the proposed model

Perception arises from the product of sensations and predictions of these sensations. It has been suggested that perception involves adapting an internal model of the world (our predictions) to match what the world seems to be, given the sensory input (our sensations) (Mumford, 1992). Similarly, predictive coding formulations propose that what we register is the difference between predictions and actual sensory input, i.e., prediction error (Rao and Ballard, 1999). These ideas have been combined with empirical Bayes for describing perceptual learning and inference (Friston, 2003; 2005). In this view, ERPs are elicited by prediction error; in other words, ERPs are an expression of unpredictable events. In this view, prediction error is conveyed to higher cortical areas via forward connections, where predictions are updated in the light of new data. These predictions (posteriors or empirical priors in the subsequent loop) are then sent back to the lower cortical areas. This is a recurrent process that ceases when reconciliation between predictions and sensory input is reached. The research performed in this thesis provides experimental evidence that the MMN is an ER that can be interpreted in the light of predictive coding (Friston, 2005). In summary, the predictive coding framework postulates that evoked responses correspond to prediction error that is explained away during perception and is suppressed by changes in synaptic efficacy during perceptual learning. In this context, the MMN would be the result of prediction error, which is due to an unexpected deviant, or oddball, embeded in learnt a sequence of standard events. The MMN would arise when there is a mismatch between current stimulus input (unpredictable deviants) and a memory trace of previous input (predictable standards). Crucially, the predictive coding framework encompasses the two distinct hypotheses, *model-adjustment hypothesis* (*cf.* Winkler et al., 1996; Näätänen and Winkler, 1999) and *adaptation* (*cf.* May et al., 1999; Ulanovsky et al., 2003; Jääskeläinen et al., 2004), proposed in the literature to explain the mechanisms underlying the MMN. Moreover it can accommodate the findings of a multitude of studies showing that there are temporal and frontal cortical sources underlying the MMN generation (Rinne et al., 2000; Jemel et al., 2002; Opitz et al., 2002; Doeller et al., 2003; Liebenthal et al., 2003; Molholm et al., 2005; Restuccia et al., 2005). According to predictive coding, increases in intrinsic connectivity may encode progressive increases in the estimated precision of top-down predictions, responsible for suppressing prediction error. Changes in forward connections may reflect changes in sensitivity to prediction error that is conveyed to higher levels. These higher levels form predictions so that backward connections can provide contextual guidance to lower levels. In this view, the MMN represents a failure to predict bottom-up input and consequently a failure to suppress prediction error. The work described in this thesis supports this idea by showing that both neuronal adaptation within areas and short-term synaptic plasticity of inter-areal connections coexist.

An example of experimental evidence that can be interpreted in terms of predictive coding is that dipole intensity is stronger for large deviants (100%) compared with medium deviants (30%) at the temporal sources (Opitz et al., 2002). On the other hand, a reversed pattern was observed in the right frontal cortex; i.e., a bigger dipole strength with smaller deviances. The authors discuss these findings in terms of alternative explanations and suggest that the prefrontal cortex (**IFG**) contributes to a top-down process that modulates the deviance detection system in the temporal cortex (**STG**) (see also Doeller et al., 2003). Under Bayesian models of perception (Yuille & Kersten, 2006) this dissociation can be interpreted easily as greater prediction error in low-level sources for large deviants. Conversely, in higher levels, ambiguous bottom-up cues may induce prediction errors that cannot be explained away by supraordinate levels. Very similar dissociations between high and low-level responses to predictable and unpredictable stimuli have been reported in the visual cortex (e.g., Murray et al, 2004; Harrison et al, 2007).

7.4 DCM: methodological and theoretical considerations

DCM is a spatiotemporal model that explains ER on the basis of an underlying cortical network. Differences in ER wave forms are explained by means of changes in effective connectivity. This is motivated by theoretical considerations that perceptual learning involves changes in synaptic strength. The work described in this thesis involved validation and applications of DCM to ERP data.

DCM as a hypothesis driven method

DCM is a hypothesis driven or model-based method; not an explorative technique. DCM does not make a full search on all possible hypotheses or models underlying a certain brain response. Instead, DCM tests specific and well-formulated hypotheses. Therefore, there is no point in performing a DCM if one does not have a testable hypothesis about the underlying model, or connectivity graph. Critically, with DCM one can test specific models constrained by biologically plausible hypotheses. The danger is, though, that the model is far from the true architecture. Questions might arise as to whether a model is specified correctly in terms of sources included or excluded and how these sources are connected. However, there is no such thing as a right or wrong model, only better or worse approximations to reality, which can be evaluated objectively with Bayesian model comparison. Comparison of DCMs has proved to be useful for disambiguating between strong competing hypotheses. The log-evidence, estimated for each model affords an objective measure for selection of the best model amongst competing alternatives. Inference based on the log-evidence of a model is meaningless unless compared to log-evidence of a competing model. Hence, although DCM does not make an exhaustive search over all possible models, it provides a framework that can accommodate comparisons for alternative mechanistic hypothesis or network models; as many as one likes.

DCM and source localisation

Source localisation refers to inversion of a forward electromagnetic model, which is implicit in the implementation of DCM as used in the previous chapters. DCM embeds the same formalism of ECD solutions with the additional constraint that the activity in each source is caused by the activity in its adjacent sources. In **Chapters 3-6**, DCM has been furnished with soft spatial priors (relatively informative priors; 16-32 mm^2 Gaussian dispersion) on source locations based on the relevant literature. The orientations were estimated by DCM under uninformative or flat priors. When possible, the best practice would be to choose priors on source locations based on the actual data set in question. However, evaluations of DCM with somatosensory evoked potentials revealed that precision on the orientation is substantially greater than the precision on location (Kiebel et al., 2006). Therefore, informed priors on location can be derived from conventional source reconstruction techniques such as distributed sources (David et al., 2006), classical ECDs, from fMRI analysis, or from the literature, as performed here.

DCM and priors

A critical point in DCM is the use of a considerable amount of prior information. This is also the case in other domains where one needs to make certain assumptions to solve an ill-posed problem. This can be done explicitly as in DCM, or implicitly in the choice of electrodes for further analysis. DCM requires certain assumptions or priors on the model parameters. The choice of priors can be a delicate issue and while some might call it prejudice, others would call it scientific judgement. Indeed, there is a vast space for debate on whether one should use priors; what a sensible prior is; and where do the priors come from. For instance, how does one choose the prior means on the connectivity parameters, and the areas or the connections to include in the DCM graph? Ideally, this information should come from established knowledge in the literature, or from the data itself, which lives on the lowest level of this hierarchical framework. It should be noted that Bayesian model comparison can be used to compare priors because the priors are part of the model. This means, in principle, it is possible to optimise the priors *per se*.

The use of eigenmodes (SVD) in DCM

In this thesis three or eight eigenmodes were used to invert the DCM. *Does DCM deal only with few channel data?* No, DCM will fit any number of channels or modes. For computational reasons, the channel data are projected onto their principal eigenmodes to reduce the size of the matrices the inversion scheme has to handle. Generally, one would use the same number of modes as there are sources. This is because data generated by an *n*-source model can only span an *n*-dimensional subspace of sensor space. Provided the signal is large, relative to noise, the first *n* principal eigenvariates should capture the majority of signal.

DCM does not model subcortical activity

The activity at the level of the thalamus or in any other subcortical structures is not modelled in DCM. Inputs are directly fed into the cortical areas. The assumption here is that the activity arising from the thalamus is the same, regardless of whether standard or deviant tones are presented (David et al., 2006). However, differences in the responsiveness of the thalamus could be modelled by simply adding an extra source to the DCM. Another assumption in DCM is that all extrinsic, or long-range connections, are excitatory (Felleman and Van Essen, 1991). In principle this is a fair assumption for modelling cortico-cortical networks, yet this is not the case if one were to model networks involving basal ganglia, which are known to use inhibitory connections (Mumford, 1991). Incorporating this into the model would, however, require additional technical developments.

Modelling evoked responses vs. labelling components

A well established procedure in ERP research is to associate a specific component to a specific physiologic or cognitive process. Deviations from the usual wave form seen at a given electrode are useful for tracing a cognitive deficit or disorder. Yet, DCM does not attempt to model individual components but all components in one go. Hence, all peaks or components are seen as manifestations of an underlying dynamic process, which involves multiple hierarchical processing levels.

7.5 Directions for future research

This section puts forward some ideas for potential future research to further explore the mechanisms of MMN generation. Ideas of predictive coding are extrapolated for other domains of cognitive neuroscience and suggestions for methodological advances are proposed.

Bayesian random effects analysis

The experimental results described in the previous chapters (**Chapters 3-6**) relied mainly on Bayesian model comparison of different DCMs. Selection of the best model is based on the log Bayes factor for the group, which is obtained by adding individual log Bayes factor from each subject. Adopting this approach assumes correctly that the data from each subject are independent. However, this does not take into account the random effects from subject to subject, in terms of which model was actually engaged. Critically, one outlier subject may bias the result, at the group level, towards one model, whereas most of the other subjects show preference for a different model. One way of addressing this issue would involve a full hierarchical Bayesian model that includes random effects from each subject, this relaxes the assumption that all subjects use the same underlying perceptual mechanism. One could argue that this is not the case; in that different subjects may use different mechanisms or information processing strategies and therefore engage different cortical networks. For these reasons, the research performed in this thesis did not rely on the sum of the Bf only, but also on supporting information from classical statistics on the models and on the parameters at the second (between-subject) level.

Time specificity of backward modulations in the MMN

The work presented in **Chapters 5 and 6** is important because it provides a direct statistical measure for the likelihood of the competing hypotheses concerning the generation of the mismatch negativity, namely *adaptation*, *model adjustment* and *predictive coding*. The results suggest that the underlying mechanisms for the MMN rest on a multi-level cortical network involving connectivity changes in forward, backward and intrinsic connections. It would be interesting however, to examine whether the results obtained in **Chapter 4** still hold for the different data analysed in **Chapters 5 and 6**. In other words, one could ask the question: *When in time do modulations in backward connections become important for explaining the MMN?* This could be addressed easily with a similar type of analysis as described in **Chapter 4**, i.e., by testing a FBi-model against a Fi-model, as a function of peri-stimulus time. In this analysis, however, data of the two trial types would be used. These data, elicited within a finer oddball paradigm, would have the advantage of being free of overlapping components such as N1 and P300, as encountered in the data used in **Chapters 3 and 4**.

The role of attention in the MMN

There has been some speculation about whether and how attention influences the MMN. Some studies report that the MMN is independent or seldom affected by attention, whereas other studies suggest that the MMN is attenuated when the subject's attention is outside the focus of the auditory stimulus (Arnott and Alain, 2002; Müller et al., 2002). On the other hand, the degree to which attention is paid to the visual stimulus does not seem to influence the MMN (Otten et al., 2000). It would be interesting to investigate the role of attention in the MMN and how it relates to changes in backward connections. One would expect that top-down attention would be expressed in changes in backward connections. A possible study would have a 2x2 design with attention vs. no-attention and visual vs. auditory stimulus. In this way one could ask two specific questions: *Does attention have an* *effect on the MMN, as well as backward connections? If yes, is that dependency specific to stimulus' modality?*

MMN in schizophrenia

There have been several studies showing significant reductions in MMN amplitude in schizophrenia (Umbricht & Krljes, 2005). Moreover, individual MMN amplitudes correlate with disease severity and cognitive dysfunction (Baldeweg et al., 2004). It has also been suggested that schizophrenia is a disconnection syndrome (Friston and Frith, 1995; Javitt, 2004; Harrison & Weinberger, 2005; Stephan et al., 2006). From a predictive coding perspective, the fact that schizophrenic patients have hallucinations suggest that they do not make very good predictions of their sensory input. Is this due to a failure in the flow of information from the environment upstream to higher cortical areas, to where predictions are processed? Is it a failure to update predictions given new (contradictory) information? Or can we not dissociate these two processes of learning and inference? In brief, if this is a disconnection syndrome, one could ask: *Where is the disruption in the loop, in the forward connections, in the backwards or in both?* This could be addressed with the sort of connectivity models proposed in this thesis, using MMN data from schizophrenics and controls. Presumably, if the disruption occurred at the level of the forward or backward connections, these connections would show aberrant learning-dependent changes. A potentially useful approach would be to identify relevant conditional parameters estimated with DCM (e.g. the connectivity parameters, or the conduction delays) that correlate with cognitive dysfunction or symptoms. This might lead to important improvements in the diagnosis and classification of schizophrenic patients into different subgroups.

MMN and pharmacological manipulations

Neuromodulators, such as acetylcholine and dopamine, modify the spatio-temporal pattern of the neuronal circuitry by acting on excitatory and inhibitory synaptic transmission (Hasselmo, 1995). Pharmacologically induced changes in the MMN have been investigated in numerous studies, using a variety of drugs affecting different neurotransmitter systems. The most robust, and perhaps also the most important neuropharmacological effect, given its importance for relating the MMN to schizophrenia, is exerted through NMDA receptors: several studies have found strong reductions of MMN amplitude under the NMDA antagonist ketamine (Javitt et al., 1996; Kreitschmann-Andermahr et al., 2001; Umbricht et al., 2000; 2002), with only a single study failing to find an effect of ketamine (Oranje et al., 2002). In contrast to NMDA receptors, the roles of dopamine, serotonin, nicotinic, muscarinic and GABA receptors for MMN generation are more controversial. Concerning nicotinic receptors, for example, whereas most studies reported an increase in the MMN amplitude by nicotinic receptor stimulation (Baldeweg et al., 2006; Dunbar et al., 2007; Engeland et al., 2002), other studies found different effects (Harkrider and Hedrick, 2005; Inami et al. 2005), and one study did not find any effect at all (Knott et al., 2006). The only two available studies of the role of muscarinic receptors in the MMN, performed by the same authors, gave contradictory results (Pekkonen et al., 2001; 2005). Finally, inconsistent results have also been obtained in studies manipulating $GABA_A$ receptor function, with some studies reporting a significant reduction of MMN amplitude by benzodiazepines (Nakagome et al., 1998; Rosburg et al., 2004), whereas other studies failed to observe a significant modulation of the MMN (Kasai et al., 2002; Murakami et al., 2002; Smolnik et al., 1998).

Overall, one might conclude that the roles of dopaminergic, serotoninergic, nicotinic, muscarinic and GABA receptors in MMN generation are currently not well established and require further research. In contrast, there is broad agreement amongst studies that blockage of NMDA receptors leads to significant reductions in MMN amplitude. It would be interesting to evaluate the action of classical neuromodulators on changes in cortical activity and their expression in DCM parameters. In other words, if neuromodulators modify synaptic plasticity and cortical excitability, this should have an effect on the connectivity parameters of a DCM.

DCM and induced responses

It is important to mention that evoked responses convey only part of the information contained in EEG and MEG measurements. Induced or spectral responses might also contribute for the MMN in which case it would be interesting to study the dynamics of the underlying network in terms of linear and non-linear coupling. See Chen et al. (2008) for a recent development of a DCM for induced responses and its application in the context of a face-perception experiment.

7.6 Concluding remarks

The work presented in this thesis explored mechanistic hypotheses that map onto connectivity models or DCMs that underlie a specific evoked response, the MMN. This work has been performed under the premises of predictive coding and empirical Bayes as a general framework for understanding perception. In other words, perception arises from an interplay between inputs from the environment and predictions of these inputs. Inputs, predictions and prediction errors (difference between input and predictions) travel across the different hierarchical levels of an interconnected cortical network. The research described in this thesis provides experimental evidence that perception can be framed within predictive coding. Moreover it demonstrates the usefulness of DCM in addressing core problems in neuroscience such as connectivity, inference, and learning in the brain.

This thesis is based on the following publications numbered from I-VII (see below). These publications comprise original research described in the previous chapters of this thesis (**Chapters 3-6**) and a review of the literature on the underlying mechanisms of the MMN (related to **Section 1.2**).

7.6.1 Summary of original contributions

The original contributions of this thesis are summarised as follows:

- Predictive validity of DCM for ERPs was established. A methodological procedure for inferences on model and connectivity parameters at the group level is proposed. (**Chapter 3**, Publication I)
- Backward connections are necessary to explain late ERP components. (**Chapter 4**, Publication III)
- The generation of the MMN can be explained with a connectivity model comprising multiple hierarchical cortical levels. This model entails both *adaptation* and *model-adjustment* which is in line with predictive coding. (**Chapter 5**, see also Publication II for

methodological advances and Publications IV and V for description of the studies)

- Learning by repetition suppresses intrinsic and extrinsic connectivity in the brain. (**Chapter 6**, Publication VI)

7.6.2 Publications arising from work in this thesis

Published:

I. **M.I. Garrido**, J.M. Kilner, S.J. Kiebel, K.E. Stephan, K.J. Friston (2007) Dynamic Causal Modelling of evoked potentials: A reproducibility study. NeuroImage 36: 571-580.

II. S. J. Kiebel, **M. I. Garrido**, K. J. Friston (2007) Dynamical causal modelling of evoked responses: The role of intrinsic connections. NeuroImage 36: 332-345.

III. **M.I. Garrido**, J.M. Kilner, S.J. Kiebel, K.J. Friston (2007) Evoked brain responses are generated by feedback loops. Proc Natl Acad Sci U S A 104: 20961-20966.

In submission:

IV. **M.I. Garrido**, J.M. Kilner, S.J. Kiebel, K.J. Friston. A predictive coding account of the mismatch negativity.

V. **M.I. Garrido**, K.J. Friston, S.J. Kiebel, K.E. Stephan, T. Baldeweg, J.M. Kilner. The functional anatomy of the MMN: a DCM study of the roving paradigm.

VI. **M.I. Garrido**, K.J. Friston, S.J. Kiebel, K.E. Stephan, T. Baldeweg, J.M. Kilner. Repetition suppression and cortico-cortical plasticity in the human brain

VII. **M.I. Garrido**, J.M. Kilner, K.E. Stephan, K.J. Friston. The mismatch negativity: a review of underlying mechanisms.

References

Akatsuka K, Wasaka T, Nakata H, Inui K, Hoshiyama M, Kakigi, R. Mismatch responses to temporal discrimination of somatosensory stimulation. Clin. Neurophysiol. 116, 1930-1937 (2005).

Alain C, Woods DL, Knight RT. A distributed cortical network for auditory sensory memory in humans. Brain Res. 812, 23-37 (1998).

Alho K, Sainio K, Sajaniemi N, Reinikainen K, Näätänen R. Event-related brain potential of human newborns to pitch change of an acoustic stimulus. Electroencephalogr. Clin. Neurophysiol. 77, 151-155 (1990).

Alho K. Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. Ear Hear. 16, 38-51 (1995).

Arnott SR, Allan C. Stepping out of the spotlight: MMN attenuation as a function of distance from the attended location. Neuroreport 13, 2209-2212 (2002).

Astikainen P, Ruusuvirta T, Wikgren J, Korhonen T. The human brain processes visual changes that are not cued by attended auditory stimulation. Neurosci. Lett. 368, 231-234 (2004).

Atienza M, Cantero JL. Complex sound processing during human REM sleep by recovering information from long-term memory as revealed by the mismatch negativity (MMN). Brain Res. 901, 151-160 (2001).

Atienza M, Cantero JL, Dominguez-Marin E. The Time Course of Neural Changes Underlying Auditory Perceptual Learning. Learn. Mem. 9, 138-150 (2002).

Atienza M, Cantero JL, Quiroga RQ. Precise timing accounts for postraining sleep-dependent enhancements of the auditory mismatch negativity. Neuroimage. 26, 628-634 (2005).

Baldeweg T, Richardson A, Watkins S, Foale C., Gruzelier J. Impaired frequency discrimination in dyslexia detected with mismatch evoked potentials. Ann. Neurol. 45, 495-503 (1999).

Baldeweg T, Klugman A, Gruzelier J, Hirsch SR. Mismatch negativity potentials and cognitive impairment in schizophrenia. Schizophr. Res. 69, 203-217 (2004).

Baldeweg T, Wong D, Stephan KE. Nicotinic modulation of human auditory sensory memory: evidence from mismatch negativity potentials. Int. J. Psychophysiol. 59, 49-58 (2006).

Baldeweg T. Repetition effects to sounds: evidence for predictive coding in the auditory system. Trends Cogn. Sci. 10, 93-94 (2006).

Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, Dale AM, Hamalainen MS, Marinkovic K, Schacter DL, Rosen BR, Halgren E. Top-down facilitation of visual recognition. Proc. Natl. Acad. Sci. U S A 103, 449-454(2006).

Bays PM, Flanagan JR, Wolpert DM Attenuation of Self-Generated Tactile Sensations is Predictive, not Pstdictive. PLoS Biol. 4, e28 (2006).

Behrens TE, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214-1221(2007).

Böttcher-Gandor C, Ullperger P. Mismatch negativity in event-related potentials to auditory stimuli as a function of varying interstimulus interval. Psychophysiology 29, 546-550 (1992).

Boussaoud D, Ungerleider LG, Desimone R. Pathways for motion analysis: cortical connections of the medial superior temporal and fundus of the superior temporal visual areas in the macauqe. J. Comp. Neurol. 296, 462-95(1990).

Bramon E, Croft RJ, McDonald C, Virdi GK, Gruzelier JG, Baldeweg T, Sham PC, Frangou S, Murray RM. Mismatch negativity in schizophrenia: a family study. Schizophr. Res. 67,1-10 (2004).

Chater N, Tenenbaum JB, Yuille A. Probabilistic models of cognition: Conceptual foundations. Trends. Cogn. Sci. 10, 287-293 (2006).

Chen CC, Kiebel SJ, Friston KJ. Dynamic causal modelling of induced responses. Neuroimage 41, 1293-1312 (2008).

Cowan N, Winkler I, Teder W, Näätänen R. Memory pre-requisites of mismatch negativity in the auditory even-related potential (ERP). J. Exp. Psychol. Learn. Mem. Cogn. 19, 909-921 (1993).

Cusack R, Deeks J, Aikman G, Carlyon RP. Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J. Exp. Psychol. Hum. Percept. Perform. 30, 643-656 (2004).

Czigler I, Balázs L, Pató LG. Visual change detection: event-related potentials are dependent on stimulus location in humans. Neurosci. Lett. 364, 149-153 (2004).

David O, and Friston KJ. A neural mass model for MEG/EEG: coupling and neuronal dynamics. Neuroimage 20, 1743-1755 (2003).

David O, Harrison L, Friston KJ. Modelling event-related responses in the brain. Neuroimage 25, 756-770 (2005).

David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, Friston KJ. Dynamic causal modelling of evoked responses in EEG and MEG. Neuroimage 30, 1255-1272 (2006).

Deneve S, and Pouget A. Bayesian multisensory integration and cross-modal spatial links. J. Physiol. Paris 98, 249-258(2004).

Denham SL, and Winkler I. The role of predictive models in the formation of auditory streams. J. Physiol. Paris 100, 154-170 (2006).

Deouell LY, Bentin S, Giard M-H. Mismatch negativity in dichotic listening: Evidence for interhemispheric differences and multiple generators. Psychophysiology 35, 355-365 (1998).

Desimone R. Neural mechanisms for visual memory and their role in attention. Proc. Natl. Acad. Sci. U S A 93, 13494-13499 (1996).

Doeller CF, Opitz B, Mecklinger A, Krick C, Reith W, Schröger E. Prefrontal cortex involvement in preattentive auditory deviance detection: neuroimaging and electrophysiological evidence. Neuroimage 20,1270-1282 (2003).

Douglas RJ, and Martin KA. Neuronal circuits of the neocortex. Annu. Rev. Neurosci. 27, 419-451 (2004).

Douglas RJ, and Martin KA. Recurrent neuronal circuits in the neocortex. Curr. Biol. 3, R496-500 (2007).

Dunbar G, Boeijinga PH, Demazieres A, Cisterni C, Kuchibhatla R, Wesnes K, Luthringer R. Effects of TC-1734 (AZD3480), a selective neuronal nicotinc receptor agonist, on cognitive performance and the EEG of young heathy male volunteers. Psychopharmacology (Berl.) 191, 919-929 (2007).

Dyson BJ, Alain C, He Y. I've heard it all before: Perceptual invariance represented by early cortical auditory-evoked responses. Brain Res. Cogn. Brain Res. 23, 457-460 (2005).

Edelman GM. Neural Darwinism: selection and reentrant signalling in higher brain function. Neuron 10, 115-25 (1993).

Engeland C, Mahoney C, Mohr E, Ilivitsky V, Knott VJ. Acute nicotine effects on auditory sensory memory in tacrine-treated and nontreated patients with Alzheimer's disease: an event-related potential study. Pharmacol. Biochem. Behav. 72, 457-464 (2002).

Escera C, Alho K, Winkler I, Näätänen R. Neural Mechanisms of Involuntary Attention to Acoustic Novelty and Change. J. Cogn. Neurosci. 10, 590-604 (1998).

Escera C, Yago E, Corral M-J, Corbera S, Nunez MI. Attention capture by auditory significant stimuli: semantic analysis follows attention switching. Eur. J. Neursci. 18, 2408-2412 (2003).

Felleman DJ, and Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cereb.Cortex 1, 1-47 (1991).

Friston KJ, Frith CD, Passingham RE, Liddle PF, Frackowiak RS. Motor practice and neurophysiological adaptation in the cerebellum: a positron tomography study. Proc. Biol. Sci. 248, 223-228 (1992).

Friston KJ, and Frith CD. Schizophrenia: a disconnection syndrome? Clin. Neurosci. 3, 89-97. Review (1995).

Friston KJ. Bayesian estimation of dynamical systems: an application to fMRI. Neuroimage 16, 513–530 (2002).

Friston K. Learning and inference in the brain. Neural Netw. 16,1325-1352 (2003).

Friston KJ, Harrison L, Penny W. Dynamic causal modelling. Neuroimage 19, 1273-1302 (2003).

Friston, K. A theory of cortical responses. Philos. Trans. R. Soc. Lond., B. Biol. Sci. 360, 815-836 (2005).

Friston K, Kilner J, Harrison L. A free energy principle for the brain. J. Physiol Paris 100, 70-87 (2006a).

Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W. Variational free energy and the Laplace approximation. Neuroimage 34, 220-234 (2006b).

Friston K, Kiebel S, Garrido M, David O. Dynamic causal models for EEG. Statistical Parametric Mapping 2007. Chapter 42, pages 561-576.

Gaillard AW. Problems and paradigms in ERP research. Biol. Psychol. 26, 91-109 (1988).

Garrido MI, Kilner JM, Kiebel SJ, Stephan KE, Friston KJ. Dynamic Causal Modelling of evoked potentials: A reproducibility study. Neuroimage 36, 571-580 (2007a).

Garrido MI, Kilner JM, Kiebel SJ, Friston KJ. Evoked brain responses are generated by feedback loops. Proc. Natl. Acad. Sci. U S A 104, 20961-20966 (2007b).

Giard MH, Perrin F, Pernier J, Bouchet P. Brain generators implicated in the processing of auditory stimulus deviance: a topographic event-related potential study. Psychophysiology 27, 627-640 (1990).

Gomot M, Giard M-H, Roux S, Barthelemy C, Bruneau N. Maturation of frontal and temporal components of mismatch negativity (MMN) in children. Neuroreport 14, 3109-3112 (2000).

Grau C, Fuentemilla L, Marco-Pallarés J. Functional neural dynamics underlying auditory event-related N1 and N1 suppression response. Neuroimage 36, 522-531 (2007).

Griffiths TL, Tenenbaum JB. Structure and strength in causal induction. Cognit. Psychol. 51, 334-384(2005).

Haenschel C, Vernon DJ, Prabuddh D, Gruzelier JH, Baldeweg T. Event-Related Brain Potential Correlates of Human Auditory Sensory Memory-Trace Formation. J. Neurosci. 25, 10494-10501 (2005).

Hari R, Hämäläinen M, Ilmoniemi R, Kaukoranta E, Reinikainen K, Salminen J, Alho K, Näätänen R, Sams M. Responses of primary auditory cortex to pitch changes in a sequence of tone pips: neuromagnetic recordings in man. Neurosci. Lett. 50, 127-132 (1984).

Hari R, Rif J, Tiihonen J, Sams M. Neuromagnetic mismatch fields to single and paired tones. Electroencephalogr. Clin. Neurophysiol. 82, 152-154 (1992).

Harkrider AW, and Hedrick MS. Acute effect of nicotine on auditory gating in smokers and non-smokers. Hear. Res. 202, 114-128 (2005).

Harrison LM, Stephan KE, Rees G, Friston KJ. Extra-classical receptive field effects measured in striate cortex with fMRI. Neuroimage 34, 1199-208 (2007).

Harrison PJ, and Weinberger DR. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. Mol. Psychiatry 10, 40-68 (2005).

Hasselmo ME. Neuromodulators and cortical function: modelling the physiological basis of behaviour. Behav. Brain Res. 67, 1-27 (1995).

Henson RN, Goshen-Gottstein Y, Ganel T, Otten LJ, Quayle A, Rugg MD Electrophysiological and haemodynamic correlates of face perception, recognition and priming. Cereb. Cortex 13, 793-805 (2003).

Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. (Lond.) 160, 106-154 (1962).

Hupe JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature 394, 784-787 (1998).

Inami R, Kirino E, Inoue R, Arai H. Transdermal nicotine administration enhances automatic auditory processing reflected by mismatch negativity. Pharmacol. Biochem. Behav. 80, 453-461 (2005).

Jääskeläinen IP, Lehtokoski A, Alho K, Kujala T, Pekkonen E, Sinclair JD, Näätänen R, Sillanaukee P. Low doses of ethanol suppresses mismatch negativity of auditory event-related potentials. Alcohol Clin. Exp. Res. 19, 607-610 (1995).

Jääskeläinen IP, Ahveninen J, Bonmassar G, Dale AM, Ilmoniemi RJ, Levänen S, Lin FH, May P, Melcher J, Stufflebeam S, Tiitinen H, Belliveau JW. Human posterior auditory cortex gates novel sounds to consciousness. Proc. Natl. Acad. Sci. U S A 101:6809-6814 (2004).

Jacobsen T, and Schröger E. Is there pre-attentive memory-based comparison of pitch? Psychophysiology 38, 723-727 (2001).

Jansen BH, and Rit VG. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. Biol. Cybern. 73, 357-366 (1995).

Javitt DC, Steinscheneider M, Schroeder CE, Arezzo JC. Role of cortical N-methyl-D-aspartate receptors in auditory sensory memory and mismatch negativity generation: Implications for schizophrenia. Proc. Natl. Acad. Sci. U S A 93, 11962-11967 (1996).

Javitt DC. Glutamate as a therapeutic target in psychiatric disorders. Mol. Psychiatry 9, 984-997 (2004).

Jemel B, Achenbach C, Muller BW, Ropcke B, Oades RD. Mismatch negativity results from bilateral asymmetric dipole sources in the frontal and temporal lobes. Brain Topogr. 15, 13-27 (2002).

Kane NM, Curry SH, Butler SR, Cummins BH. Electrophysiological indicator for awakening from coma. Lancet 341, 688 (1993).

Kasai K, Yamada H, Kamio S, Nakagome K, Iwanami A, Fukuda M, Yumoto M, Itoh K, Koshida I, Abe O, Kato N. Do high or low doses of anxiolytics and hypnotics affect mismatch negativity in schizophrenic subjects? An EEG and MEG study. Clin. Neurophysiol. 113, 141-150 (2002).

Kekoni J, Hämäläinen H, Saarinen M, Gröhn J, Reinikainen K, Lehtokoski A, Näätänen R. Rate effect and mismatch responses in the somatosensory system: ERP-recordings in humans. Biol. Psychol. 46, 125-142 (1997).

Kersten D, Mamassian P, Yuille A. Object Perception as Bayesian Inference. Annu. Rev. Psychol. 55, 271-304 (2004).

Kiebel SJ, David O, Friston KJ. Dynamic causal modelling of evoked responses in EEG/MEG with lead field parameterization. Neuroimage 30, 1273-1284 (2006).

Kiebel SJ, Garrido MI, Friston KJ Dynamic Causal Modelling of evoked responses: the role of intrinsic connections. Neuroimage 36, 332-345 (2007).

Kilner JM, Friston KJ, Frith CD. The mirror-neuron system: a Bayesian perspective. Neuroreport 18, 619-623 (2007).

Kilner JM, Penny WD, Friston KJ. Application of robust averaging in artifact removal of EEG time-series. In submission.

Knott V, Blais C, Scherling C, Camarda J, Millar A, Fisher D, McIntosh J. Neural effects of nicotine during auditory selective attention in smokers: an event-related potential study. Neuropsychobiology 53, 115-126 (2006).

Körding KP, Wolpert DM. Bayesian integration in sensorimotor learning. Nature 427, 244-247 (2004).

Körding KP, Wolpert DM. Bayesian decision theory in sensorimotor control. Trends Cogn. Sci. 10, 319-326 (2006).

Kraus N, McGee TJ, Carrell TD, Zecker SG, Nicol TG, Koch DB. Auditory neurophysiologic responses and discrimination deficits in children with learning problems. Science 273, 971-973 (1996).

Kreitschmann-Andermahr I, Rosburg T, Demme U, Gaser E, Nowak H, Sauer H. Effect of ketamine on the neuromagnetic mismatch negativity field in healthy humans. Brain Res. Cogn. Brain Res. 12, 109-116 (2001).

Kujala T, Tervaniemi M, Schröger E. The mismatch in cognitive and clinical neuroscience: Theoretical and methodological considerations. Biol. Psychol. 74, 1-19 (2007).

Lamme VA, and Roelfsema PR. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci. 23, 571-579 (2000).

Levänen S, Ahonen A, Hari R, McEvoy L, Sams M. Deviant auditory stimuli activate human left and right auditory cortex differently. Cereb. Cortex 6, 288-296 (1996).

Liebenthal E, Ellingson ML, Spanaki MV, Prieto TE, Ropella KM, Binder JR. Simultaneous ERP and fMRI of the auditory cortex in a passive oddball paradigm. Neuroimage 19, 1395-1404 (2003).

Liegeois-Chauvel C, Musolino A, Badier JM, Marquis P, Chauvel P. Evoked potentials recorded from the auditory cortex in the man: evaluation and topography of the middle latency components. Electroencephalogr. Clin. Neurophysiol. 92, 204-214 (1994).

Light GA, and Braff D.L. Stability of mismatch negativity deficits and their relationship to functional impairments in chronic schizophrenia. American Journal of Psychiatry 162, 1741-1743 (2005).

Loveless N, Levanen S, Jousmaki V, Sams M, Hari R. Temporal integration in the auditory sensory memory: neuromagnetic evidence. Electroencephalogr. Clin. Neurophysiol. 100, 220-228 (1996).

Maess B, Jacobsen T, Schröger E, Friederici AD. Localizing pre-attentive auditory memory-based comparison: Magnetic mismatch negativity to pitch change. Neuroimage 37, 561-571 (2007).

Makeig S, Westerfield M, Jung T-P, Enghoff S, Townsend J, Courchesne E, Sejnowski TJ. Dynamic Brain Responses of Visual Evoked Responses. Science 295, 690-693 (2002).

Marco-Pallarés J, Grau C, Ruffini G. Combined ICA-LORETA analysis of mismatch negativity. Neuroimage 25, 471-477 (2005).

May P, Tiitinen H, Ilmoniemi RJ, Nyman G, Taylor JG, Näätänen R Frequency change detection in human auditory cortex. J. Comput. Neurosci. 6, 99-120 (1999).

Mazaheri A, Jensen O. Posterior α activity is not phase-reset by visual stimuli. Proc. Natl. Acad. Sci. U S A 103, 2948-2952 (2006).

McCormick DA, Wang Z, Huguenard J. Neurotransmitter Control of Neocortical Neuronal Activity and Excitability. Cereb. Cortex 3, 387-389 (1993).

McCormick DA, Williamson A. Convergence and divergence of neurotransmitter action in human cerebral cortex. Proc. Natl. Acad. Sci. U S A 86, 8098-8102 (1989).

Mel BW. Synaptic integration in an excitable dendritic tree. J. Neurophysiol. 70, 1086-1101 (1993).

Michie PT, Innes-Brown H, Todd J, Jablensky AV. Duration mismatch negativity in biological relatives of patients with schizophrenia spectrum disorders. Biol. Psychiatry 52, 749-758 (2002).

Molholm S, Martinez A, Ritter W, Javitt DC, Foxe JJ. The Neural Circuitry of Pre-attentive Auditory Change-detection: An fMRI Study of Pitch and Duration Mismatch Negativity generators. Cereb. Cortex 15, 545-551 (2005).

Müller BW, Achenbach C, Oades RO, Bender S, Schall U. Modulation of mismatch negativity by stimulus deviance and modality of attention. Neuroreport 13, 1317-1320 (2002).

Mumford D. On the computational architecture of the neocortex. I. The role of the thalamo-cortical loop. Biol. Cybern. 65, 135-145 (1991).

Mumford D. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern. 66, 241-251 (1992).

Murakami T, Nakagome K, Kamio S, Kasai K, Iwanami A, Hiramatsu K, Fukuda M, Hata A, Honda M, Watanabe A, Kato N. The effects of benzodiazepines on event-related potential indices of automatic and controlled processing in schizophrenia: a preliminary report. Prog. Neuropsychopharmacol. Biol. Psychiatry 26, 651-661 (2002).

Murray SO, Kersten D, Olshausen BA, Schrater P, Woods DL. Proc. Natl. Acad. Sci. U S A 99, 15164-15169 (2002).

Murray SO, Schrater P, Kersten D. Perceptual grouping and the interactions between visual cortical areas. Neural Netw. 17, 695-705 (2004).

Nakagome K, Ichikawa I, Kanno O, Akaho R, Suzuki M, Takazawa S, Watanabe H, Kazamatsuri H. Overnight effects of triazolam on cognitive function: an event-related potentials study. Neuropsychobiology 38, 232-240 (1998).

Näätänen R. The role of attention in auditory information processing as revealed by event related potentials and other brain measures of cognitive function. Beha. Brain Sci. 13, 201-288 (1990).

Näätänen R. Attention and Brain Function (Lawrence Erlbaum, Hillsdale, New Jersey, 1992).

Näätänen R, Schröger E, Karakas S, Tervaniemi M, Paavilainen P. Development of a memory trace for a complex sound in the human brain. Neuroreport 4, 503-506 (1993).

Näätänen R, and Alho K. Mismatch negativity – a unique measure of sensory processing in audition. Int. J. Neurosci. 80, 317-337 (1995).

Näätänen R, Lehtokoski A, Lennes M, Cheour M, Houtilainen M, Ilvonen A, Vainio M, Alku P, Ilmoniemi RJ, Luuk A, Allik J, Sinkkonen J, Alho K. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385, 432-434 (1997).

Näätänen R, and Winkler I. The concept of auditory stimulus representation in cognitive neuroscience. Psychol. Bull. 125, 826-859 (1999).

Näätänen R. Mismatch negativity (MMN): perspectives for application. Int. J. Psychophysiol. 37, 3-10 (2000).

Näätänen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I. “Primitive intelligence” in the auditory cortex. Trends. Neurosci. 24, 283-288 (2001).

Näätänen R, and Rinne T. Electric brain response to sound repetition in humans: an index of long-term-memory – trace formation? Neurosci. Lett. 318, 49-51 (2002).

Näätänen R. Mismatch negativity: clinical research and possible applications. Int. J. Psychophysiol. 48, 179-188 (2003).

Näätänen R, Pakarinen S, Rinne T, Takegata R. The mismatch negativity (MMN): towards the optimal paradigm. Clin. Neurophysiol. 115, 140-144 (2004).

Näätänen R, Jacobsen T, Winkler I. Memory-based or afferent process in mismatch negativity (MMN): a review of the evidence. Psychophysiology 42, 25-32 (2005).

Neumann J, and Lohmann G. Bayesian second-level analysis of functional magnetic resonance images. Neuroimage 20, 1346-1355 (2003).

Nikulin VV, Linkenkaer-Hansen K, Nolte G, Lemm S, Müller KR, Ilmoniemi RJ, Curio G. A novel mechanism for evoked responses in the human brain. Eur J Neurosci. 25, 3146-3154 (2007).

Opitz B, Rinne T, Mecklinger A, von Cramon DY, Schröger E. Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. Neuroimage 15, 167-174 (2002).

Oranje B, van Berckel BN, Kemner C, van Ree JM, Kahn RS, Verbaten MN. The effects of a sub-anaesthetic dose of ketamine on human selective attention. Neuropsychopharmacology 22, 293-302 (2002).

Otten LJ, Alain C, Picton TW. Effects of visual attentional load on auditory processing. Neuroreport 11, 875-880 (2000).

Paavilainen P, Simola J, Jaramillo M, Näätänen R, Winkler I. Preattentive extraction of abstract feature conjunctions from auditory stimulation as reflected by the mismatch negativity (MMN). Psychophysiology 38, 359-365 (2001).

Pazo-Alvarez P, Cadaveira F, Amenedo E. MMN in the visual modality: a review. Biol. Pshycol. 63, 199-236 (2003).

Pekkonen E, Hirvonen J, Jääskeläinen IP, Kaakkola S, Huttunen J. Auditory sensory memory and the cholinergic system: implications for Alzheimer's disease. Neuroimage 14, 376-382 (2001).

Pekkonen E, Jääskeläinen IP, Kaakkola S, Ahveninen J. Cholinergic modulation of preattentive auditory processing in aging. Neuroimage 27, 387-392 (2005).

Penny WD, Stephan KE, Mechelli A, Friston KJ. Comparing dynamic causal models. Neuroimage 22, 1157-1172 (2004).

Picton TW, Alain C, Otten L, Ritter W. Mismatch Negativity: Different Water in the Same River. Audiol Neurootol 5, 111-139 (2000).

Pollen AD. On the Neural Correlates of Visual Perception. Cereb Cortex 9, 4-19 (1999).

Pulvermüller F. Brain reflections of words and their meaning. Trends Cogn. Sci. 5, 517-523 (2001).

Rademacher J, Morosan P, Schormann T, Schleicher A, Werner C, Freund H-J, Zilles K. Probabilistic mapping and volume measurement of human primary auditory cortex. Neuroimage 13, 669-683 (2001).

Rao RP, and Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79-87 (1999).

Restuccia D, Della Marca G, Marra C, Rubino M, Valeriani M. Attentional load of the primary task influences the frontal but not the temporal generators of the mismatch negativity. Brain Res. Cogn. Brain Res. 25, 891-899 (2005).

Rinne T, Alho K, Holi M, Sinkkonen J, Virtanen J, Bertrand O, Näätänen R. Analysis of speech sounds is left-hemisphere predominant at 100-150 ms after sound onset. Neuroreport 10, 1113-1117 (1999).

Rinne T, Alho K, Ilmoniemi RJ, Virtanen J, Näätänen R. Separate time behaviors of the temporal and frontal mismatch negativity sources. Neuroimage 12, 14-19 (2000).

Rinne T, Degerman A, Alho K. Superior temporal and inferior frontal cortices are activated by infrequent sound duration decrements: an fMRI study. Neuroimage 26, 66-72 (2005).

Rockland KS, and Pandya DN. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain Res. 179, 3-20 (1979).

Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat. Neurosci. 2, 1131–1136 (1999).

Rosburg T, Marinou V, Haueisen J, Smesny S, Sauer H. Effects of lorazepam on the neuromagnetic mismatch negativity (MMNm) and auditory evoked field component N100m. Neuropsychopharmacology 29, 1723-1733 (2004).

Rosburg T, Trautner P, Dietl T, Korzyukov OA, Boutros NN, Schaller C, Elger CE, Kurthen M. Subdural recordings of the mismatch negativity (MMN) in patients with focal epilepsy. Brain 128, 819-828 (2005).

Saarinen J, Paavilanen P, Schröger E, Tervaniemi M, Näätänen R. Representation of abstract attributes of auditory stimuli in the human brain. Neuroreport 3, 1149-1151 (1992).

Salin PA, and BullierJ. Corticocortical connections in the visual system: structure and function. Physiol. Rev. 75, 107-154 (1995).

Sallinen M, Kaartinen J, Lyytinen H. Is the appearance of mismatch negaticity during stage 2 sleep related to the elicitation of K-complex? Electroencephalogr. Clin. Neurophysiol. 91, 140-148 (1994).

Sams M, Paavilainen P, Alho K, Näätänen R. Auditory frequency discrimination and event-related potentials. Electroenceph, clin. Neurophys. 62, 437-48 (1985).

Schiff S, Mapelli D, Vallesi A, Orsato R, Gatta A, Umilta C, Amodio P. Clin Neurophysiol. 117, 1728-1736 (2006).

Smolnik R, Pietrowsky R, Fehm HL, Born J. Enhanced selective attention after low-dose administration of the benzodiazepine antagonist flumazenil. J. Clin. Psychopharmacol. 18, 241-247 (1998).

Stephan KE, Baldeweg T, Friston KJ. Synaptic plasticity and dysconnection in schizophrenia. Biol. Psychiatry 59, 929-939 (2006).

Steyvers M, Griffiths TL, Dennis S. Probabilistic inference in human semantic memory. Trends Cogn. Sci. 10, 327-334 (2006).

Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J. Predictive codes for forthcoming perception in the frontal cortex. Science 314, 1311-1314 (2006).

Sussman E, Ritter W, Vaugham HG. Attention affects the organization of auditory input associated with the mismatch negativity system. Brain Res. 789, 130-138 (1998).

Sussman E, and Winkler I. Dynamic sensory updating in the auditory system. Brain Res. Cogn. Brain Res. 12, 431-439 (2001).

Sussman E, and Steinschneider M. Neurophysiological evidence for context-dependent encoding of sensory input in human auditory cortex. Brain Res. 1075, 165-174 (2006).

Sussman ES, Horváth J, Winkler, Orr M. The role of attention in formation of auditory streems. Percept. Psychophys. 69, 136-152 (2007).

Syndulko K, Cohen SN, Tourtellotte WW, Potvin AR Endogenous event-related potentials: prospective applications in neuropsychology and behavioral neurology. Bull. Los Angeles Neurol. Soc. 47, 124-140 (1982).

Tenenbaum JB, Griffiths TL, Kemp C. Theory-based Bayesian models of inductive learning and reasoning. Trends Cogn Sci 10, 309-318 (2006).

Tervaniemi M, Maury S, Näätänen R. Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. Neuroreport 5, 844-846 (1994).

Tervaniemi M, Ilvonen T, Sinkkonen J, Kujala A, Alho K, Huotilainen M, Näätänen, R. Harmonic partials facilitate pitch discrimination in humans: electrophysiological and behavioural evidence. Neurosci. Lett. 279, 29-32 (2000a).

Tervaniemi M, Medvedev SV, Alho K, Pakhomov SV, Roudas MS, van Zuijen TL Näätänen R. Lateralized Automatic Auditory Processing of Phonetic Versus Musical Information: A PET Study. Hum. Brain Mapp. 10, 74-79 (2000b).

Tiitinen H, Alho K, Huotilainen M, Ilmoniemi RJ, Simola J, Näätänen R. Tonotopic auditory cortex and the magnetoencephalographic (MEG) equivalent of the mismatch negativity. Psychophysiology 30, 537-540 (1993).

Tiitinen H, May P, Reinikainen, Näätänen R. Attentive novelty detection in humans is governed by pre-attentive sensory memory. Nature 372, 90-92 (1994).

Tiitinen H, Salminen NH, Palomäki KJ, Mäkinen VT, Alku P, May PJC. Neuromagnetic recordings reveal the temporal dynamics of auditory spatial processing in the human cortex. Neurosci. Lett. 396, 17-22 (2006).

Tremblay K, Kraus N, McGee T. The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. Neuroreport 9, 3557-3560 (1998).

Ulanovsky N, Las L, Nelken I. Processing of low-probability sounds by cortical neurons. Nat. Neurosci. 6, 391-398 (2003).

Umbricht D, Schmid L, Koller R, Vollenweider FX, Hell D, Javitt DC. Ketamine-induced deficits in auditory and visual context-dependent processing in healthy volunteers: implications for models of cognitive deficits in schizophrenia. Arch. Gen. Psychiatry 57, 1139-1147 (2000).

Umbricht D, Koller R, Vollenweider FX, Schmid L. Mismatch negativity predicts psychotic experiences induced by NMDA receptor antagonist in healthy volunteers. Biol. Psychiatry 51, 400-406 (2002).

Umbricht D, Koller R, Schmid L, Skrabo A, Grübel C, Huber T, Stassen H. How specific are deficits in mismatch negativity generation to schizophrenia? Biol Psychiatry 53, 1120-1131 (2003).

Umbricht D, and Krljes S. Mismatch negativity in schizophrenia: a meta-analysis. Schizophr. Res. 76, 1-23 (2005).

van Zuijen TL, Sussman E, Winkler I, Näätänen R, Tervaniemi M. Grouping of sequential sounds – an event-related potential study comparing musicians and non-musicians. J. Cogn. Neurosci. 16, 331-338 (2004).

van Zuijen TL, Simoens VL, Paavilainen P, Näätänen R, Tervaniemi M. Implicit, Intuitive, and Explicit Knowledge of Abstract Regularities in Sound Sequence: An Event-related Brain Potential Study. J. Cogn. Neurosci. 18, 1292-1303 (2006).

Varela F, Lachaux, J-P, Rodriguez E, Martinerie J. The brainweb: phase synchronization and large-scale integration. Nat. Rev. Neurosci. 2, 229-239 (2001).

Vuust P, Pallesen KJ, Bailey C, van Zuijen TL, Gjedde A, Roepstorff A, Østergaard L. To musicians, the message is in the meter pre-attentive neuronal responses to incrongruent rhythm are left-lateralized in musicians. Neuroimage 24, 560-564 (2005).

Wager TD, Keller MC, Lacey SC, and Jonides J. Increased sensitivity in neuroimaging analysis using robust regression. NeuroImage 26, 99-113 (2005).

Winkler I, Karmos G, and Näätänen R. Adaptive modelling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. Brain Res. 742, 239-252 (1996).

Winkler I, Tervaniemi M, and Näätänen R. Two separate codes for missing-fundamental pitch in the human auditory cortex. J. Acoust. Soc. Am. 102, 1072-1082 (1997).

Woldorff MG, Gallen, CC, Hampson SA, Hillyard SA, Pantev C, Sobel D, Bloom FE. Modulation of early sensory processing in human auditory cortex during auditory selective attention. Proc. Natl. Acad. Sci. U S A 90, 8722-8726 (1993).

Wolpert DM, Ghahramani Z, Jordan MI An internal model for sensorimotor integration. Science 269, 1880-1882 (1995).

Wolpert DM, Doya K, Kawato M. A unifying computational framework for motor control and social interaction. Phil. Trans. R. Soc. Lond. B 358, 593-602 (2003).

Wray J, and Green GGR. Calculation of the Volterra kernels of non-linear dynamic systems using an artificial neural network. Bio. Cybern 71, 187-195 (1994).

Yabe H, Tervaniemi M, Reinikainen K, Näätänen R. Temporal window of integration revealed by MMN to sound omission. Neuroreport 8, 1971-1974 (1997).

Yuille A, and Kersten, D. Vision as Bayesian inference: analysis by synthesis? Trends Cogn. Sci. 10, 301-308 (2006).