KU LEUVEN

FACULTEIT INGENIEURSWETENSCHAPPEN

Decoding EEG responses during perception and imagination of music

Marthe Tibo

Thesis submitted for the degree of
Master of Science in
Biomedical Engineering

Thesis supervisor:
Prof. dr. ir. Alexander Bertrand

Assessors:
Prof. dr. ir. Hugo Van hamme
Prof. dr. ir. Tom Francart

Mentor:
Ir. Simon Geirnaert

Academic year 2019 – 2020

Preface

From the very first course of neurophysiology I had in my curriculum, I loved to learn about the workings of the brain and its mystery still keeps on fascinating me today. Especially, visual and auditory processing have always been intriguing to me. It is hard to grasp how the brain, with such complex and high amounts of data, can make sense of it all. The fact that, in this study, I could combine this fascination for the human brain with other interests, such as music, mathematics and biomedical data processing, made this a very interesting topic for me to study.

I would therefore like to thank everyone who made this study possible. Thank you Simon, for being an absolutely great mentor, for sharpening my critical thinking, or for simply listening when I needed that the most. Thank you to my supervisor, professor Alexander Bertrand, for making this topic possible and for the helpful input. I would like to thank my assessors, professor Hugo Van hamme and professor Tom Francart, for their feedback and questions during the evaluation moments.

I would like to thank Di Liberto et al. and Stober et al. for making their data sets publicly available.

Thank you to my parents, Lucas, my friends and the rest of the family for their support throughout the years. Nonkel Jan, het indienen van deze thesis heb je jammer genoeg net niet meer mogen meemaken, hoewel je daar nog graag bij geweest was. Hopelijk maakt dit je vandaag blij. Deze thesis draag ik dan ook mee op aan jou.

Marthe Tibo

Contents

Abstract

In this study, we will decode EEG responses during perception and imagination of music. During perception, subjects listen to an auditory stimulus, after which a neural processing pathway transforms the auditory cues into a brain response. During imagination of music, no physical auditory source is present. Stimuli are simply 'heard' by the brain.

In order to study these conditions, we will implement a linear regression model with regularization, which will reconstruct stimulus envelopes out of their EEG responses. This model will show varying results over all studied perception and imagination conditions. As an overall conclusion, we find that perception experiments significantly outperform music imagination. After adding an extension to this model, explicitly incorporating the possibility of a latency in the EEG recording, the model will show vastly improved results. However, we see that these results seem to have a periodical behavior in function of this new parameter. Possible effects of beat tracking will therefore be investigated.

Moreover, an analysis of effects within the results will be done, investigating the influence of subject musicality, song categories and imagination techniques. We will look whether songs can be distinguished from each other via various classifiers and discuss the influence of the preprocessing filter range. It was observed that the musical stimuli used in this study achieve equal or better results when using a bandpass filter between 1-10 Hz compared to a 1-30 Hz bandpass filter.

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

BCI	Brain computer interface
CCA	Canonical correlation analysis
CV	Cross-validation
EEG	Electroencephalogram
fMRI	Functional magnetic resonance imaging
LS	Least squares
MSE	Mean squared error
MIIR	Music imagination information retrieval
MIR	Music information retrieval

Symbols

ρ	Pearson correlation
$s(t)$, $\mathbf{S}$	Musical stimulus
$\mathbf{g}$	Decoder
$r(t)$, $\mathbf{R}$	EEG response
λ	hyperparameter
ν	Time shift
τ	Time lag

Vectors and matrices are printed **bold**.

Chapter 1

Introduction

Listening to music is an activity that is performed daily by many people all over the world. Many of our regular tasks have a connection to music, whether that is noticeable or not. From the soothing background song in a restaurant, to singing along to the radio, to watching a YouTube video... When you think about it, the list of activities involving music is endless.

Every day, this exposure to sound sends millions of different cues to the brain, leading to a range of neural responses. Our brains register different rhythmic and melodic properties of musical pieces as we listen [33, 36]. This broad palette of properties (e.g., pitch, tempo...) is what makes a certain song unique. These features are very helpful in practice: they allow us to recognize music in future repetitions. One could then anticipate what part of the song comes next, which is a necessary skill in for example playing an instrument or singing [2, 34].

Furthermore, a significant amount of emotional response is present while listening to music. A song may have a calming effect or evoke a feeling of tension to the listener. One can enjoy a song or dislike it. Songs can remind us of a specific person or situation [36]. One thing to note is that these responses are not always completely voluntary or noticeable. Music in for example adds can be a handy way to influence emotions and make you spend more money, although it may not be very apparent.

Apart from listening to music, which is named 'music perception', the human brain is capable of an even more remarkable task. It cannot only remember existing musical pieces, but also imagine them without the presence of a physical source. An example of this phenomenon is a so-called 'ear worm', where people keep singing or humming the same song for hours because the piece is stuck inside their head. This so-called 'music imagination' can take on many different forms and responses [36, 34]. One person can for example solely 'hear' songs inside his/her head, while others can envision themselves singing along or playing the song on an instrument.

When brain responses are recorded for the above settings, they can form great tools to use in research. Most commonly, the registration of brain activity is done by means of an electroencephalogram (EEG) or functional magnetic resonance imaging (fMRI). Studies of these brain responses may lead to many useful insights. They could be used to shed more light on the complex neural processing of music, which is still mostly unknown to this day [36, 34, 37].

Recent research domains try to find a relation between the original songs and their evoked brain responses. Two main subfields of this problem are seen in practice. The term 'Music Information Retrieval' (MIR) is used for studying the responses of music perception, whereas 'Music Imagination Information Retrieval' (MIIR) is the term used in a music imagination setting [36, 34]. Different techniques could be used to quantify the desired relation. Generally, these techniques are borrowed from the neighboring domain of speech decoding, which tries to find this relation in the case of speech signals.

Furthermore, one could think of more futuristic uses of these brain responses. Algorithms to estimate an original song out of its imagined brain response might find their use in brain computer interfaces (BCI's). BCI's form, as the name suggests, an interface between electronics and neural tissue. This way, they could restore lost functions of the human body or provide new ones. People's abilities to communicate with the external world could (partially) be taken over or extended by a machine.

For speech, this could mean that a person's imagined words are decoded into natural speech. Although the direct use of music decoding algorithms in this way is far more limited than for speech decoding, BCI's could as well have some futuristic implementations with regards to music. Imagine for example that songs could be streamed on a device by simply thinking of them (i.e., 'Shazam for the brain').

In the next subsections, a deeper overview of the topic is given. First, the studied music processing pathway is looked upon more closely. This pathway can be subdivided into two parts: the processing in the ear and in the brain. Next, the focus is laid on the different ways to record brain responses. A comparison of the most frequently used modalities is given, with for each their respective advantages and disadvantages. Lastly, this is followed by a summary of previous results of research in literature, as well as a short link to the neighboring domain of speech decoding.

1.1 The processing pathway of music

Music, as other sounds, is composed of waves which travel at a certain frequency in time and space. Humans can register a remarkable range of frequencies, roughly estimated from 20Hz to 20 kHz. With such amount of frequencies available, a complex and effective processing needs to be carried out. In the human body, this is done in two sequential steps: the ear, as a sensory organ, and the processing in the brain.

1.1.1 Processing in the ear

The auditory processing pathway of stimuli, such as speech and music, starts in the ear. Three parts are distinguished: the outer, middle, and inner ear. The outer ear has an external part, called the pinna. The pinna fulfills complex functions in the journey of sound to the brain. It has a distinct shape, which plays an important role in the directionality detection of sound. Moreover, the pinna guides the sound wave into the auditory canal. At the end of this canal, the sound wave makes the tympanic membrane oscillate, propagating the signal to the middle ear [2, 8].

In the middle ear, the sounds traveling via the tympanic membrane are amplified by the ossicles: the *malleus*, *incus* and *stapes*. This amplification is needed because the medium through which the sound waves travel changes: from air (outer and middle ear) to fluid (inner ear). The third ossicle in the chain, the *stapes*, deforms the oval window. This deformation propagates the amplified waves to the inner ear [2, 8].

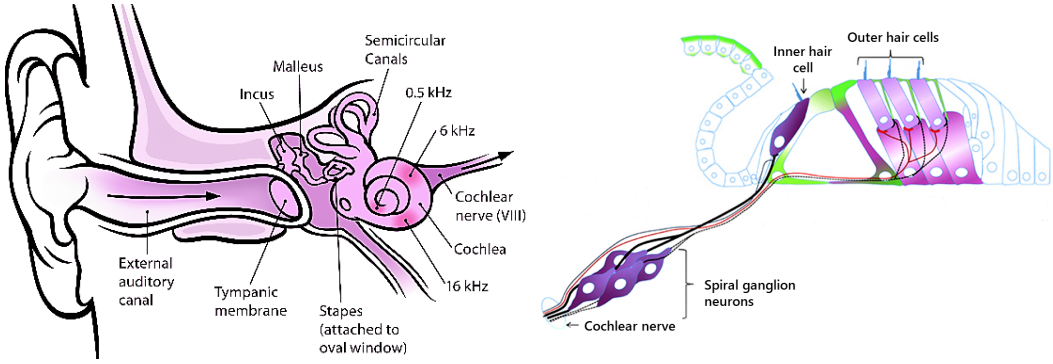

FIGURE 1.1: Left: The structure of the ear. Adopted from [8]. Right: the basilar membrane. Adopted from [19].

The inner ear is the site where the main processing of sounds takes place before their journey to the brain. It is made up out of two large components. The first one, called the labyrinth, is part of the vestibular system which controls a person's balance. The second component is the cochlea. The cochlea is a complex system of fluid filled chambers, which are twisted into a form resembling a snail shell (Figure 1.1, left). In between the chambers, a sensitive basilar membrane is found. This membrane supports tissues with complex functions, such as the organ of Corti, which contains hair cells that induce activation of the auditory nerve (Figure 1.1, right) [2, 19].

When a sound wave propagates from the oval window into the inner ear, the basilar membrane deforms and oscillates at a particular site, dependent on the frequencies present in the wave. High frequency sounds only deform the base of the basilar membrane. Very low frequencies are registered up until its apex. The lower the frequency, the higher the deformation on the basilar membrane appears [2]. This phenomenon was famously discovered by G. von Békésy, who won the Nobel Prize for medicine and physiology in 1961 for his research in audiology (Figure 1.2) [29].

The local deformation is transferred to the organ of Corti, embedded on the basilar membrane [19]. Its hair cells, also named auditory receptor cells, start to bend. A chain of biochemical events is induced by this bending, leading to the release of neurotransmitters. These neurotransmitters synaptically activate the first layer of neurons in this pathway, the spiral ganglion cells, and thereby induce firing of the auditory nerve [2, 40].

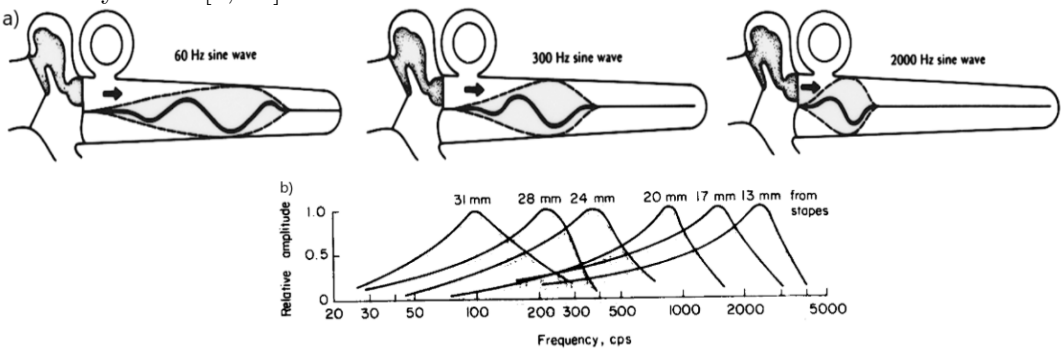

Figure 1.2: a) Frequency dependent deformation of an unfolded basilar membrane. b) The basilar membrane as a filter bank. Adopted from [40], after G. von Békésy.

Because the deformation of the basilar membrane and the subsequent events happen locally, the frequency dependency is also present further in the chain [2]. Research by Rose, Hind, Anderson and Brugge [1] shows that a single neuron fiber in the auditory nerve is also tuned to a certain characteristic frequency. The conclusion is that frequencies have a certain tonotopy or fixed spatial separation on the basilar membrane. This tonotopy is propagated to the auditory nerve, in distinct frequency bins dependent on the activation of the hair cells. In other words, the basilar membrane thus acts as a biological filter bank in the ear (Figure 1.2) [2].

1.1.2 Processing in the brain

After the first processing in the ear, the signal travels via the auditory nerve to the brain. The auditory nerve is made up of the axons of spiral ganglion cells, which possess a characteristic frequency and firing rate (measure of intensity). The nerve fibers lead the signal to the next processing step in the pathway: the nuclei of the brain stem (Figure 1.3, right) [2].

In these nuclei, more neural cell layers are added and complex connections are formed. This leads to a gradually more diverse and nonlinear processing of the signal. As an example, one can find cells that can detect specific frequency changes overtime. A nucleus in the thalamus, called the *superior olive*, is the first to receive input from both ears, leading to our ability to process sounds binaurally and derive sound location information. Also, a substantial amount of feedback is present between the brain cortex/nuclei and the nuclei/ear (hair cells) [2, 15]. It is exactly this growing complexity and nonlinearity, ascending from the sensory organs, that makes a study of brain such an advanced problem in research.

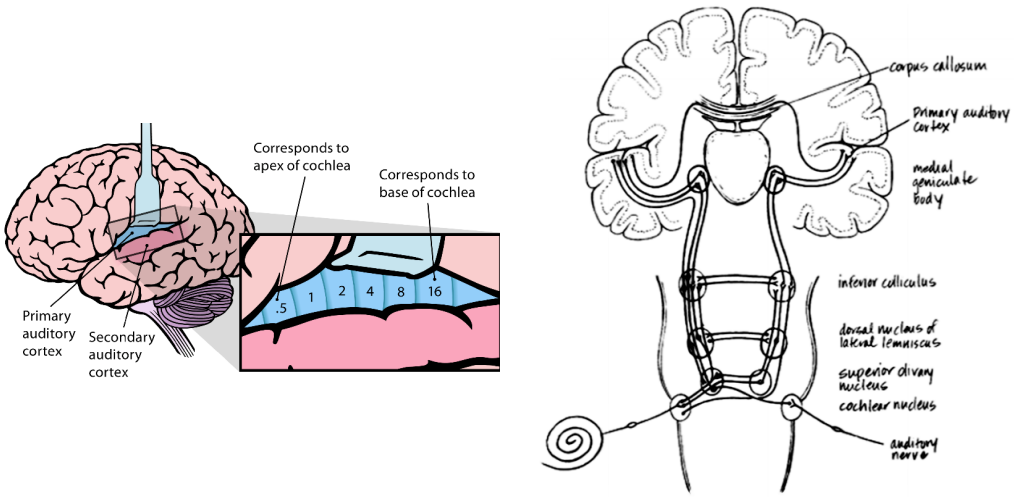

FIGURE 1.3: Left: Tonotopy in the auditory cortex. Adopted from [8]. Right: Auditory pathway to the brain. Adopted from [15].

The *medial geniculate nucleus* (MGN) propagates the signal to the auditory regions in the brain, located laterally on the head [15]. The primary and secondary auditory cortex are areas in the brain that are closely related, but still not fully understood. They play a primary role in handling frequency content, intensity, and binaural information to form a neural response to other brain areas. In the primary auditory cortex, we again see a certain tonotopic organization (Figure 1.3, left). Low frequencies are found on the anterior part, while high frequencies (up to 20 kHz) are located on the posterior part of the region [2, 8].

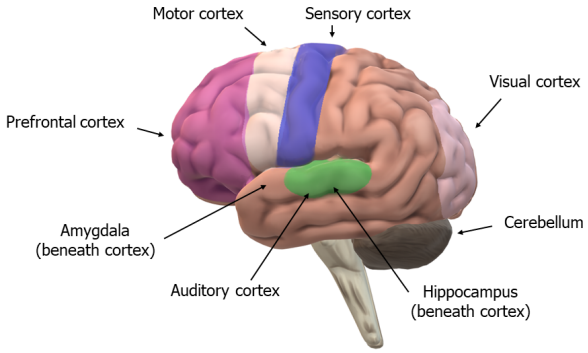

FIGURE 1.4: Active brain regions during music perception [42] [26].

In addition to the primary and secondary auditory cortex, a wide range of areas work together to induce further neural responses (Figure 1.4). For example, the hippocampus, which is a control center of memory in the brain, can link a musical piece with past experiences. The motor cortex on the other hand is responsible for movement (for example while playing an instrument). The auditory processing areas also share regions with the visual system, which might explain why a person can have visualizations while imagining music. The amygdala can induce and emotional response while listening to a song [42].

This complex collaboration between brain areas leads to interesting observations, such as beat tracking, which are not fully uncovered yet from a computational point of view. Music consists of a complex train of events, which mostly have a certain periodicity. The brain extracts this periodical beat from songs, signaling a strong tendency to tap or move to the music [39]. However, noticing this periodicity in a song is a complex task, as it requires the brain to look through the temporal structure of music and anticipate when the next beat will follow [39].

Previous research proposes that a certain musical entrainment is taking place in the brain, in which oscillatory signals at the frequency of this beat are present along the neural pathway, synchronizing brain waves to the beat of the song [39]. This synchronization might also explain why music can bring up feelings of relaxation or tension, as the beat frequency can be located in the same range used by these processes. Meditation sounds are an example of the calming effect of music.

Additionally, research is conducted on the neural processing of different music categories. Pop songs mostly have a regular beat and contain lyrics, which intertwine the processing of music with speech understanding. On the other hand, a voice could distract the listener from other musical properties. Therefore, purely instrumental music could be seen as more suitable on some occasions, for example as background music while studying. Studies using multiple music categories could shed more light on possible differences with respect to their neural processing.

Furthermore, another interesting characteristic of the brain is its plasticity, in which neural responses grow more complex or expand to larger brain areas due to for example experience and training. In a more and more studies, the brain plasticity with respect to musical training is investigated. In musicians, playing a song results in more selective attention and concentration to instruments in comparison to non-musicians, blocking irrelevant sounds and environmental cues [45]. One could then pose the interesting question if this musical training has any effect on the study of this neural processing with respect to non-musicians.

In a nutshell, the mentioned areas in Figure 1.4 are broad. It is still relatively unknown which specific areas are activated and how they work together during these complex processes [42]. Apart from that, it is also unclear how this processing might change when imagining music. These questions are the main driving force for MIR and MIIR research.

1.2 Recording brain responses

In MIR and MIIR research, recordings of the available brain responses can lead to very useful insights. These recordings can be done via different modalities. The two main approaches used in previous research, fMRI and EEG, each have their advantages and drawbacks. What follows is a short summary of their workings, along with a comparison with respect to their use in research.

1.2.1 Functional magnetic resonance imaging (fMRI)

Medical imaging techniques, more specifically magnetic resonance imaging (MRI), are a popular non-invasive way to visualize the brain. Its workings are based on electromagnetism, i.e., the disturbance of the spin magnetic moments of protons in the object to visualize [38]. A special application of this technique is functional magnetic resonance imaging, which captures an image of the blood flow and oxygen consumption in tissue via the protein hemoglobin. The two forms of hemoglobin in the blood circulation, oxygenated and deoxygenated hemoglobin, have different magnetic properties. When a brain area is active, a larger flow of oxygenated blood to those regions is seen. This difference can be exploited in an MRI setting, comparing images of activity to control images at rest [38].

The reason why fMRI is often used to capture brain responses, also for music processing studies, can be seen in the following advantages. First, fMRI is a non-invasive technique, which can be used in larger scale research. Moreover, its spatial resolution is good, capturing subtle differences even in soft tissue (such as the brain) which usually suffers from low contrast in other imaging techniques [30, 38].

There are also a few disadvantages to this technique. First, the presence of the magnetic fields and gradients requires magnets and coils. This results in a machine with large dimensions, which cannot be placed everywhere, and a high cost. Secondly, the charged coils make a loud noise during imaging. This might not only be cumbersome for the subject, but can also disturb the measurement of brain responses or interfere with a musical stimulus. Moreover, acquiring an MR image has a long duration, resulting in a relatively low temporal resolution [27].

1.2.2 Electroencephalogram (EEG)

With the upcoming BCI research, the EEG is making a steep incline as modality to register brain responses in studies. It is also the recording modality used in this work. The EEG technique uses a very different approach to quantify brain activity. When neurons fire, electrical signals are passed which can be measured as potential differences [32]. When electrodes are placed on the head of the subject, creating measuring channels, these differences can be captured relative to a designated reference electrode. In practical settings, the electrodes are usually embedded in a cap and spatially distributed according to the internationally recognized 10-20 system (Figure 1.5). To maintain good electrical contact (low impedance), a gel is applied between the electrodes and the scalp [32].

Due to its low relative cost and non-invasive nature, the EEG is a good candidate for many practical applications and BCI research [31, 32]. Secondly, its temporal resolution is good. The EEG electrodes can directly pick up event-related potentials (ERP's) as they occur in the brain. Besides, the EEG measurement produces time series data for each electrode, which can be easier to handle in calculations in comparison to the large pool of data in an MR image. Moreover, the equipment is small compared to fMRI and thus more generally applicable, especially in BCI set-ups [32].

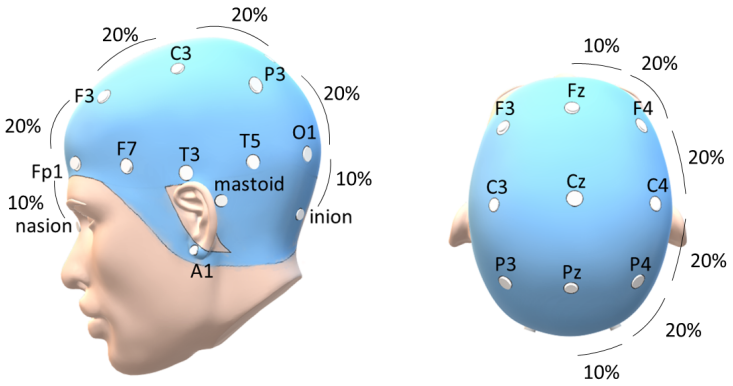

Figure 1.5: The 10-20 system with reference electrode locations [26].

A disadvantage of this technique is the relatively low signal-to-noise ratio present in raw EEG measurements. The recorded brain responses also involve neural activity from for example movement and eye blinking. Also surrounding electrical signals, such as powerline noise (50-60Hz) could disturb the measurement [32]. To minimize such disturbances in practical set-ups, the subject is therefore often sitting in a shielded room and is asked to keep still while focusing on a certain mark [36]. It is however never possible to fully prevent artifacts and noise during the recording. Therefore, specialized preprocessing steps (Section 3.1) often need to be carried out before the actual analysis to further minimize the influence of artifacts and noise on the results.

Apart from these problems, the EEG modality also has a worse spatial resolution. The electrodes pick up brain signals from every point beneath the placement area, without knowledge of the exact origin of the signal [32]. This is a clear disadvantage in comparison to fMRI, which focuses on specific points and their material composition [30]. Moreover, this fuzzy spatial resolution does also mean that nearby electrodes pick up similar signals. Also, this might pose some problems, for example in the choice of reference electrode (Section 3.1.1) [20].

To alleviate both the problems regarding EEG and fMRI measurements, a combination of modalities is used [21], or sometimes also a magnetoencephalogram. Moreover, in rare cases, electrodes can also be used in invasive techniques for the tracking of neural responses. This can be done when an electrode mesh on the brain itself is present, for example for patients who require neural monitoring and/or deep brain stimulation [32]. These techniques are of course not readily available nor possible on a large scale.

1.3 Music processing in the literature

Now that a little light was shed on the processing pathway of music and its recording procedures, how can this biological signaling chain be further explored in research? First and foremost, this can be done by developing a model for this neural pathway. We can design a set of calculations, which are holding certain assumptions, to mimic the processing of music in the ear and brain [10, 5]. These calculations thus form the bridging between the stimulus on the one hand and the corresponding brain response on the other hand.

There are of course various ways to build this model. It should be noted that this step highly dictates the final result, as it determines which assumptions are made on the processing pathway. Apart from the model selection, also the input and output representations are an important choice: they control which features of the stimulus and brain response are accounted for in the model. In this following section, the most predominant ways to build a model for music processing are given [5]. Afterwards, we continue with a list of possible representations of the stimulus in studies, which can be used in these models together with the recorded brain response.

1.3.1 Linear modeling

To linearize, or not to linearize...

As seen before, the neural pathway of music is a nonlinear process that forms complex connections to many different areas in the brain. It would thus be a logical choice to opt for a nonlinear technique to model this response, which can include the desired high complexity [42]. For these kinds of problems, (recurrent) neural networks are frequently used [5, 9]. However, as we will discuss in this section, the availability of biomedical data is a bottleneck in these applications. In literature, a trend in the use of linear techniques is therefore seen. This approximation has a few advantages and disadvantages (Table 1.1).

Table 1.1: Advantages of linear techniques w.r.t. nonlinear (neural) networks.

Linear techniques	
Advantages	**Disadvantages**
Interpretability	Limited modeling capacity
Less training examples required	
Smaller computational complexity	
Less parameters (overfitting)	

The first advantage of linear techniques is the interpretability of the coefficients [5]. The explanatory power of parameters in a neural network gets easily lost because of the amount and complexity of connections. With a linear technique however, input and output are linearly related in the form of parameters or weights. With these weights, one can easily see a relative influence or importance.

Moreover, there are also a couple of advantages with respect to the training of these models. Neural networks tend to have many parameters, so a lot of examples and time are required to optimally train these kinds of models [5]. These amounts of training examples might not always be available, it is dependent on the number of participating subjects and duration of testing in an experiment. In comparison, with fewer parameters, linear techniques have a smaller computational complexity [5].

Besides that, overfitting (see later) is also a great risk for neural networks. This effect, which can highly skew the perception of our results, is a point of attention for all models and techniques. However, this risk gets larger when the number of parameters in a model increases [5]. Neural networks thus usually require a substantial amount of regularization in training to reduce this danger of overfitting.

An obvious drawback of linear techniques, is that this is only an approximation of the real neural processing. Only a linear relationship between the brain responses and stimuli can be incorporated, but more complex calculations are not included [5]. At first glance, this seems like a very crude assumption. However, in literature, interesting results are found with linear techniques, which suggests that its modeling capacity might be large enough to capture the basic music processing information and to describe a significant relation between a stimulus and its brain response. Some of these results in music and speech literature will described further in this section.

Modelling the relation with linear techniques

In the previous subsection, it was seen that linear techniques have some significant advantages over nonlinear techniques when it comes to the implementation of neural processing models. Therefore, linear techniques will be used to define the musical processing pathway in the remainder of this work. Once the appropriate calculation technique is chosen, it is also necessary to think about the desired relation to model. Before translating the musical processing pathway to calculations, this relation needs to be carefully defined. The main approaches can be divided into three categories: encoding, decoding and 'hybrid encoding-decoding' techniques (see Figure 1.6) [16, 10].

Encoding approaches, also called forward modeling, provide a direct and intuitive mimicking of the neural pathway [14]. In this strategy, a model tries to transform a musical stimulus into its corresponding brain response. Linear regression is used to model the encoding approach. This technique provides linear mapping weights between the stimulus and brain response, which were learned from examples in a training phase. When these weights a are applied to a new stimulus $s(t-\tau)$ (delayed by a lag τ), an estimate of each channel j of the corresponding brain response $\hat{b}_j(t)$ can be calculated [10, 13]:

$$\hat{b}_j(t) = \sum_{\tau} s(t-\tau)a_j(\tau). \tag{1.1}$$

Decoding techniques or backward models take the opposite approach to look at the mapping relation: these strategies try to transform the neural response of a song back

to its original musical stimulus. Also in this approach, linear regression is used, but in the opposite mapping direction. Fitted weights g, learned from previous training data, can now be used to reconstruct a stimulus estimate $\hat{s}(t)$ out of (often) time lagged neural response channels $b_j(t)$ (with time lag τ) [10, 31, 41]:

$$\hat{s}(t) = \sum_j \sum_\tau b_j(t-\tau) g_j(\tau). \tag{1.2}$$

As seen on Figure 1.6 on the left, these mapping transformations optimize a specific objective function. The encoding or decoding weights are determined so that there is a minimal error in the model. This means that in the training phase, weights are chosen that lead to a maximal similarity (minimal error) between the real and reconstructed stimulus or brain response in the training examples [3, 5]. Common objective functions in these linear strategies are the minimalization of the mean squared error (MMSE) or maximization of the Pearson correlation. Biesmans et. al. [3] demonstrated that these approaches lead to equal results.

It is shown that the decoding approach is usually more reliable to define the neural processing relation of music and speech than encoding approaches. This can be explained by the following reason. When ascending the neural pathway starting from the stimulus, the complexity and amount of information passing in neurons increases. When these signals arrive in the brain, millions of other processes are active, besides the musical encoding. They require activity in many of the same regions where the response to music happens, leading to a brain recording that contains 'additional neural information' with respect to the stimulus [44, 14].

When relating a stimulus to its brain response via the encoding strategy, this technique will thus always encounter the problem of 'missing information' with respect to all other brain processes in the neural recording. The missing knowledge cannot be recovered from the stimulus alone, usually leading to an influence on the eventual outcome and significance of results. In decoding techniques on the other hand, it is far more practical to learn which components in the neural response are important for music processing, for example by a training phase for the weights of a linear regression model, and to filter the unrelated data out of the brain response [44, 14].

A drawback of the decoding strategy is that the weights of the model cannot be directly interpreted as an activation pattern in the brain related solely to the stimulus. The filter weights, learned during training to minimize the mapping error, could also exploit a part of the brain response that is not stimulus-related to optimize its calculations. With encoding techniques on the other hand, weights are a direct reflection of how and in which amounts a stimulus is represented in a brain response channel [14]. This is the so-called temporal response function (TRF), which can be seen as the transfer function between a stimulus signal and its neural output [9].

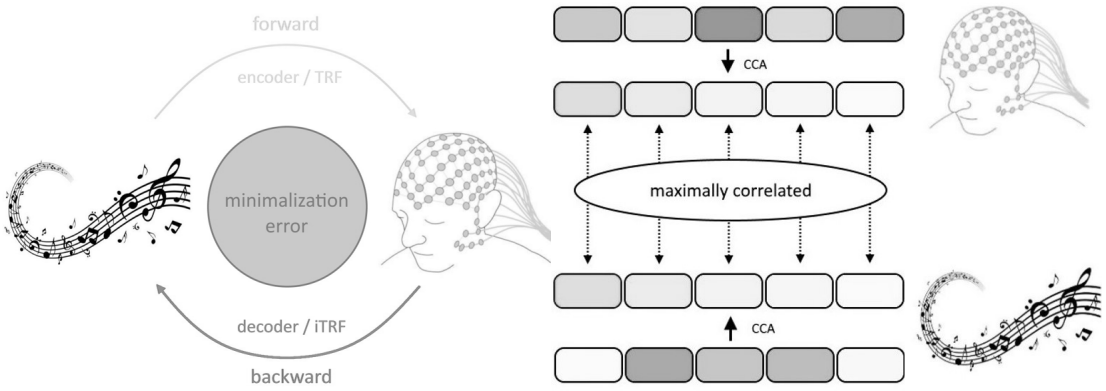

Figure 1.6: Left: In forward modeling, the EEG is predicted from the stimulus, while backward modeling, the stimulus is reconstructed from the EEG. Right: CCA combines these techniques. (EEG and music drawing adopted from [28] and [6])

Apart from entirely mapping a stimulus to a brain response or vice versa, we could also think of a combination between the two preceding strategies, a 'hybrid encoding-decoding strategy' [16]. Canonical correlation analysis (CCA) is a common linear technique to achieve this strategy in literature. Here, we transform both the stimulus and brain response with linear weights $g_{s,j}$ and $g_{b,j}$ respectively (and time lags τ), resulting in two new intermediate representations ($\hat{c}_s(t)$ for the stimulus, $\hat{c}_b(t)$ for the brain response) that can be made maximally relatable, for example by maximizing Pearson correlation [16, 10]:

$$\hat{c}_s(t) = \sum_{\tau} s(t-\tau) g_{s,j}(\tau) \qquad \hat{c}_b(t) = \sum_{j} \sum_{\tau} b_j(t-\tau) g_{b,j}(\tau). \tag{1.3}$$

Comparing these equations to the ones of encoding and decoding, it can be clearly seen why this technique is sometimes called a 'hybrid encoding-decoding strategy' [16]. In literature, slight variations of the above equations can be found, for example depending on the applied temporal filtering (lags τ) [10]. A schematic representation of CCA is shown in Figure 1.6 on the right.

1.4 Representation of the musical stimulus

With the choice of a model, also comes the choice of the appropriate representation of the stimulus, which accompanies the neural recordings in the model. In this subsection, an overview of the most common stimulus representations is given. A musical sound can be represented in various ways in research. Each of these representations focuses on (the combination of) certain aspects of the song. With the choice of representation, comes thus another choice: which parts of a stimulus are relevant for the study in question? The representation of a stimulus can thereby have a large impact on the final result. Since the field of music processing is still relatively small, representational choices are often influenced by good results in the larger neighboring domain of speech processing. A summary of the most predominant choices is listed below.

1.4.1 Envelope

An envelope signifies the overall, low frequency activity in a signal. The envelope can be a good choice when high frequency changes do not provide the bulk of information [32]. In speech processing, the envelope is a frequent choice of representation. O' Sullivan et al. [31] showed that an attended speech envelope can be discerned from a simultaneously played unattended envelope with a linear regression technique. This type of study is called auditory attention detection (AAD). de Cheveigné et al. [10] compared the results of linear regression and CCA techniques when decoding speech envelopes. Ciccarelli et al. [9] also used the envelope as stimulus representation in their comparison of linear and nonlinear techniques for speech decoding. Apart from these examples, many more can be found in literature [13, 14, 43, 44].

A reason for this popularity in speech processing literature is the following. For speech stimuli, the envelope follows a linear relation with an EEG measurement (see later) between frequencies of roughly 1 and 10 Hz [31, 10]. This is of course an interesting advantage when using linear techniques. In music processing studies, no consensus on a frequency range is described. Some studies define a larger frequency band than explained above. For example, Di Liberto et al. [11] use a range of 0.5-45 Hz in their research. On the contrary, the study of Schaefer et al. [33] defines envelopes with filtered ranges from 1-14 Hz, resembling the values used in speech processing. This divergence of ranges makes it an interesting open question for research.

There are different ways to obtain the envelope of a signal. First, the absolute value of the signal is a frequently used option. This calculation can be extended by using a log or power law relation to accommodate for the nonlinear auditory processing [3]. Alternatively, the 'mathematical envelope' can be obtained by calculating the magnitude of a complex signal using the Hilbert transform. In this method, the stimulus is seen as a modulating signal imposed onto a sine. Another calculation of the envelope is found by squaring and low pass filtering the signal [31, 10, 3].

Besides, one could ask the question if these calculations should be carried out on the raw or a processed version of the stimulus. In the latter case, a gammatone filter bank is often used, dividing the raw stimulus into several frequency bins. An advantage of adding this step to the calculations is its resemblance to the filtering in the basilar membrane, taking the natural processing in the ear into account [3]. All these different methods were compared by Biesmans et al [3]. in a linear speech processing set-up. In this study, the power law method used on the gammatone filtered subbands of the stimulus performed best. This will therefore also be the method used in this work for envelope extractions.

One last note on the envelope representations is the use of a variant in a music processing study by Sturm et al. [37]. In this publication, power slope representations are implemented, which are calculated as the differentiated envelopes of the stimuli. This choice of input for the model is chosen based on its inclusion of rapid changes in the intensity of the sound, possibly indicating note onsets.

1.4.2 Spectrogram

The spectrogram, the magnitude of the short-time Fourier transform (STFT), is also used as representation of a stimulus. Contrary to the envelope, the spectrogram includes both time and frequency information [32]. Cantisani et al. [7] implemented this representation in a stimulus reconstruction (decoding) set-up using linear regression. In comparison to the envelope, using a spectrogram as model in-/output has both its advantages and disadvantages. One advantage is that the influence of both high and low frequencies can be studied extensively. Also, the dual time-frequency information can be exploited [32]. A drawback, however, is the increase in amount of model parameters with this representation, which leads to a higher risk of overfitting.

1.4.3 Various combined features

Apart from these previously mentioned representations, a stimulus can also be defined by a variety of its properties. Our brain records numerous features of a song, e.g., pitch, timbre, intensity... As mentioned before, the combination of these features is what makes a certain song unique and recognizable. This is the reason why, in practice, a wide combination of features is used.

Gang et al. [16] extracted 20 different features[1] from stimuli, after which they were combined using PCA. Finally, they were used in a CCA approach to relate them to the brain responses of the same stimulus. Treder et al. [41] extracted spatio-temporal features of all the electrodes in their EEG experiment over three different time intervals. Afterwards, these features were used to train a classifier to perform AAD on polyphonic music. In the study of Sturm et al. [37], 9 musical properties were deduced from stimuli, exploring their relation to the 'goodness of fit' (Pearson correlation) of the implemented linear response-to-stimulus mapping.

1.5 Research questions

Now that we have seen the possible ways of setting up a linear model, along with the ways to represent stimuli and record brain responses, it is time to think about what we would like to achieve with this model. First and foremost, we pose the following question:

'Is it possible, using a linear decoding model, an EEG response and a stimulus envelope, to find a good (i.e., ideally optimal) stimulus reconstruction for perception and imagination experiments?'

If so,

- *'Is it possible to discern between different musical stimuli based on the resulting reconstructed envelopes (by implementing a classifier) in both perception and imagination experiments?'*

[1] e.g., zero-crossing rate for pitch, spectral flux for timbre...

- *'Are there any observed group effects within the results of the perception and imagination experiments (musicians vs. non-musicians, song categories, imagination techniques...)?'*
- *'Are there any effects on the results when changing the filtering range of the envelopes for the perception and imagination of music?'*

In the next chapters, we will try to find the answers to these questions. First, the used data sets and their properties will be discussed. Afterwards, an overview of the preprocessing steps is given.

Then, a first version of our stimulus reconstruction model will be defined and the obtained results will be discussed. Secondly, an extended version of this model will be implemented, after which the effects on the results will be studied. In a final chapter, different stimulus classification approaches will be discussed. All these analyses were carried out using Matlab R2019b [22].

Chapter 2

Data

In this study, two different prerecorded data sets are used. These data sets contain EEG brain recordings of several subjects listening to/imagining a selected stimulus. Each of these stimuli has a sampling frequency of 44.1 kHz. In the EEG experiments in both data sets, 64+2 (mastoid) channels were recorded at 512 Hz.

The first set, published by Stober et al. [36], contains EEG recordings in both MIR and MIIR settings. The second data set by Di Liberto et al. [12] contains only recordings of subjects listening to a stimulus and can therefore only be used in a MIR study. Apart from this, different song categories or musicality levels within subjects are also included in both data collections, making them especially useful in subsequent group analyses. In the next sections, a more detailed description of the used data is given. The comparison between both data sets can be found in the overview of Table A.2.

2.1 OpenMIIR data set - Stober et al.

In the OpenMIIR data set, published by Stober et al. [36], 12 well known stimuli (Table A.1) are presented to 10 subjects. While listening to or imagining these musical excerpts, EEG responses of the subjects are recorded. This happens in four different conditions. The experiments for each condition are repeated five times [36] per song and subject. In total, this leads to 10 subjects x 4 conditions x 12 stimuli x 5 trials = 2400 trials, or 60 per condition and subject. A visual scheme of this set-up and the conditions can be seen in Figure 2.1.

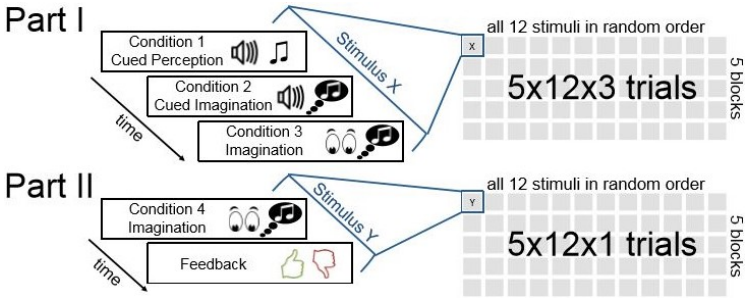

FIGURE 2.1: Set-up in the OpenMIIR data set per subject. Adopted from [36].

2.1.1 Conditions

MIR and MIIR studies require different listening and imagination conditions respectively. When two or more of these conditions are present, comparative studies could also be carried out. As shown in Figure 2.1, the four conditions in this data set are [36]:

1. Stimulus perception, with a few seconds of cue clicks preceding the stimulus to indicate the tempo;
2. Stimulus imagination, also preceded by cue clicks;
3. Stimulus imagination without cue clicks;
4. Stimulus imagination without cue clicks, but with a feedback questionnaire for the self-assessment of each subject.

The last condition is also recorded separately from the previous three [36]. This leads to an interesting question: might tempo information be retained from the previous cases, i.e., is there an effect of training with an increasing number of trials?

2.1.2 Stimuli

As mentioned before, 12 stimuli are present in this data set. These stimuli can also be classified into three groups. The first group (stimulus 1-4) contains songs with lyrics. In the second set (stimulus 5-8), the same songs are presented, but here the lyrics are removed. The third group (stimulus 9-12) involves musical pieces that are purely instrumental [36].

2.1.3 Subjects

The EEG recordings of 10 subjects are present in this data set [36]. These subjects will be denoted as P01-P10 further on in this work and can be divided according to several criteria, as listed below.

Musical training might be a factor in the performance of MIR and MIIR methods. Therefore, this data set introduces subjects with varying degrees of musical experience (Table A.2). Eight of the ten subjects had formal musical training. Four of them (P03-P05, P09) still regularly played at least one instrument, the other four subjects (P01, P06-P08) did not play music anymore. The two remaining subjects (P02, P10) were not musically trained [36].

A different way of categorizing these participants is the manner in which they imagine songs. This can be different for a music with or without lyrics. For music with lyrics, the data set allows to investigate the effects between two imagination techniques. One half of the participants imagined themselves singing (P01, P02, P05-P07), while the other half simply 'heard' the lyrics inside their heads (P03, P04, P08-P10) [36].

A parallel categorization can be made for imagining musical pieces without lyrics. Half of the participants (P04, P05, P07, P09) had some kind of visualization within these 'wordless' imagination experiments, half of them solely 'heard' the stimulus

inside their heads (P01-P03, P06, P08, P10) [36]. As mentioned in the introduction, these visualizations could arise because of shared brain regions between the visual and musical pathways. With these divisions in the participants, one could investigate if there is a difference in music imagination performance between these groups or if these aspects play any role during stimulus perception.

2.2 Bach data set - Di Liberto et al.

In this second data set, published by Di Liberto et al. [12], 10 musical excerpts from sonatas of J.S. Bach (Table A.1) are presented to 20 subjects. While listening to these pieces, the EEG responses of the participants are recorded. Tests for each stimulus are repeated three times [12, 11]. In total, this leads to 20 subjects x 1 condition x 10 stimuli x 3 trials = 600 trials.

2.2.1 Conditions

In this data set, there is only one condition: stimulus perception [12]. With this one condition, this data set can thus only be used for MIR research. One extra thing to note is the following: before listening to the stimulus, there are no cue clicks in this data set. This means that no tempo information is given before the onset of the musical pieces.

2.2.2 Stimuli

As mentioned before, 10 stimuli are present in this data set. All these stimuli are purely instrumental, originally flute and violin pieces of J. S. Bach. in this data set however, a part of the participants are expert pianists. To take into account any influence regarding the knowledge of this specific instrument, Di Liberto et al. [11, 12], changed the original instruments into piano tracks in their studies.

2.2.3 Subjects

The EEG recordings of 20 participants are present in this data set. These subjects will be denoted as P01 - P20 further on in this work. These participants can be divided by the following criterion. The first half of the participants (P01-P10) is not musically trained. The second half (P11-P20) are expert pianists, which are listening to stimuli with their preferred instrument [12]. If this difference within subjects has an effect on the decoding performance can thus be investigated. Unfortunately, since there are no imagination experiments, no information about visualizations is given in this data set.

Chapter 3

Preprocessing

The data described in the previous section cannot be directly implemented into our model. First, the extraction of the desired envelope representation and the EEG preprocessing need to be carried out. In this section, the details of these preprocessing steps are discussed. In short, the EEG recordings go through the following steps: re-referencing, artifact removal, filtering and resampling. For the stimuli, an envelope extraction is followed by a filtering and resampling step.

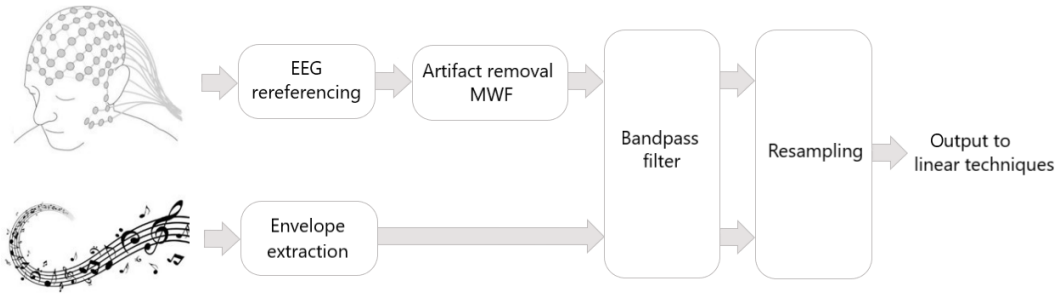

FIGURE 3.1: Schematic of the preprocessing steps. (EEG and music drawing adopted from [28] and [6] respectively)

3.1 Preprocessing of the EEG

3.1.1 Re-referencing

As mentioned before, the potential differences in an EEG are measured relative to a reference, which can be one electrode or a combination of measuring points. This reference, and thus also the relative measurements, can be freely chosen and changed by subtracting it from all data. This re-referencing step can have a few advantages.

First, the original referencing might introduce noise into the data, if these noise sources are not equally represented in the reference and other electrodes. This leads to an overall worse performance in algorithms. Opting for a new reference which captures this noise leads to a higher signal-to-noise ratio after subtraction. Secondly, neighboring EEG channels measure similar potential differences. If the original reference is positioned closely to the region of interest, a large part of the useful

signal might be gone after subtraction. However, the relative potential difference in the EEG is not changed by re-referencing, because the new reference is subtracted from all channels in the same way [20, 32].

Two common references that are subtracted from all data are:

- the average of the two mastoid channels (Figure 1.5)
- the average signal over all channels

For comparability to the two original studies of the used data sets, the first method is chosen in this work. In order to perform this re-referencing properly, the measurements of mastoid channels need to be available for all subject. In the OpenMIIR data set, these channels were not recorded for the first half of the participants due to an oversight. For the second half, they were recorded [36]. Therefore, the standard reference (common mode sense active electrode - CMS) of the EEG data acquisition was kept. This CMS electrode is traditionally placed near Cz (Figure 1.5) at the center of the head [4]. For the Bach data set, the mastoid channels were available and the above re-referencing method could thus be carried out [11].

3.1.2 Artifact removal

To counteract any noise and artifacts that are left in the signal, an artifact removal step is done. The method used for this is a multi-channel Wiener filter (MWF), via an algorithm implemented by Somers et al. [35]. In this filter, the input EEG signal $\boldsymbol{y}(t)$ is seen as a superposition of the clean EEG signal $\boldsymbol{s}(t)$ and the multi-channel artifact $\boldsymbol{n}(t)$ [35]. With the assumption that these signals are zero-mean and uncorrelated, also their covariance matrices (denoted as $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_n$ respectively) have an additive behavior [35]:

$$\begin{aligned} \boldsymbol{y}(t) &= \boldsymbol{s}(t) + \boldsymbol{n}(t) \\ \boldsymbol{\Sigma}_y &= \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_n. \end{aligned} \tag{3.1}$$

The artifact signal $\boldsymbol{n}(t)$ can be approximated by a linear combination of the original EEG channel matrix $\boldsymbol{Y}$, with weight matrix $\boldsymbol{W}$. This can be extended by adding time lagged EEG channels, making it a spatio-temporal filter. When minimizing the MMSE of the real artifact $\boldsymbol{n}(t)$ and the estimate $\hat{\boldsymbol{n}}(t)$, this results in the following equation (using that the signals are uncorrelated):

$$\hat{\boldsymbol{n}}(t) = \boldsymbol{W}^T\boldsymbol{Y} \qquad \xRightarrow{MMSE_{n\hat{n}}} \qquad \boldsymbol{W} = \boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_n. \tag{3.2}$$

The covariance matrices of the original and clean EEG signal can be estimated by indicating S clean parts (matrix $\boldsymbol{Y}_s$) and N artifacts (matrix $\boldsymbol{Y}_n$) in the original input [35]. The artifact covariance matrix can then be estimated as the difference of these two matrices (using a generalized eigenvalue decomposition approach):

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_y = \frac{\boldsymbol{Y}_n\boldsymbol{Y}_n}{N} \qquad \hat{\boldsymbol{\Sigma}}_s = \frac{\boldsymbol{Y}_s\boldsymbol{Y}_s}{S} \\ \hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_y - \hat{\boldsymbol{\Sigma}}_s. \end{aligned} \tag{3.3}$$

This way, an estimate of the filter weights and artifact signal can be made as in (3.2), after which $\boldsymbol{n}(t)$ is subtracted from the signal $\boldsymbol{y}(n)$ to obtain the clean EEG $\boldsymbol{s}(t)$.

In practice, a custom GUI was opened when running the algorithm, where broad regions around visible artifacts (mostly eye blinks and movement artifacts) were selected, clean areas remained unmarked. In the study of Somers et al. [35], it was found that an overestimation of these artifact areas had a negligible effect in comparison to an underestimation.

To calculate the filtering weights $\boldsymbol{W}$ out of these clean EEG and artifact pieces, a spatio-temporal filter with a time delay up to 5 samples was used. These weights $\boldsymbol{W}$ then construct the artifact estimate, which is subtracted from the original signal to obtain a clean EEG output.

3.1.3 Filtering

As a next preprocessing step, it is time to choose which frequencies are most descriptive for our algorithm's purpose. As mentioned before, there is no clear consensus on which filtering ranges are best for music processing. Therefore, two different ranges are tested out:

- a bandpass filter from 1-10 Hz (as in speech processing)
- a bandpass filter from 1-30 Hz (as in the data set of Stober et al. [36])

Further analysis will show which of these two filtering ranges is the best for a certain purpose (linear regression, further group analysis...) within the two chosen datasets.

3.1.4 Resampling

Next, the EEG channels are downsampled to from their inital sampling frequency of 512Hz to 64Hz. This is mostly done for time and coding efficiency, while preserving a perfectly reconstuctable EEG. The new sampling frequency of 64 Hz is also the value chosen in the study of Stober et al. [36] and it was adopted here for two reasons.

First, this sampling frequency is high enough to avoid the main problem encountered with this downsampling: aliasing. For a perfect reconstruction of a stimulus without aliasing, we need a sampling frequency at least two times as high as the highest frequency present in the signal (Nyquist criterion). This means that for a maximal frequency at 10 Hz, a minimal sampling rate of 20 Hz needs to be chosen, for 30 Hz a minimal sampling rate of 60 Hz.

Secondly, this frequency remains a divider of the old sampling frequency. This means in this case that in the process of downsampling, 1 out of 4 samples of the original EEG are taken instead of performing interpolation. This also benefits the accuracy of the resampled EEG.

3.2 Preprocessing of the stimulus

3.2.1 Envelope extraction

The chosen stimulus representation is the envelope, which will be calculated according to the best functioning method in the study of Biesmans et al. [3], introduced in Section 1.4. As in this study, a gammatone filter bank consisting of 15 filters is made. This filterbank separates the original stimulus into 15 subbands, each with their own center frequency. In the study of Biesmans et al. [3], which uses speech stimuli, these center frequencies have a range between 150 Hz and 4 kHz. However, for our music stimuli, center frequencies ranging from 1 Hz to 5 kHz seemed suitable, based on an exploratory analysis of the power spectral density of the used stimuli.

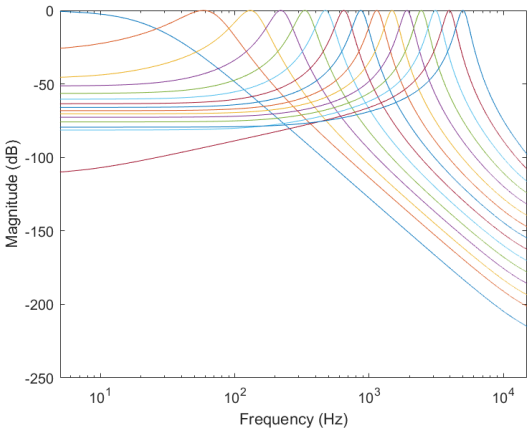

Figure 3.2: The gammatone filterbank used in this process.

From these 15 stimulus subbands, the envelope was extracted using a power law relation. For the corresponding exponent, the suggested value of 0.6 in the study of Biesmans et al. [3] was implemented. After this power law step, the subbands $a_i(t)$ are combined into one envelope signal $s(t)$, with summation weights equal to unity.

$$s(t) = \sum_{i=1}^{15} |a_i(t)|^{0.6} \tag{3.4}$$

3.2.2 Filtering

After the envelope extraction step, the music envelopes are filtered to obtain the desired range of frequencies. The bandpass filter used for this is traditionally the same as for the EEG, which is a logical step with the subsequent decoding algorithm in mind. The two bandpass filtering ranges are thus again defined as:

- 1-10 Hz (as in speech processing)
- 1-30 Hz (as in the data set of Stober et al. [36]).

3.2.3 Resampling

For the last step in the preprocessing of the stimuli, resampling, the new sampling frequency is the same as the EEG channels (from 44,1 kHz to 64 Hz). A thing to note is that here, samples need to be interpolated to obtain the downsampled envelope, since the new sampling frequency is not a divider of the old one. However, if we attempted to find a frequency which holds this property for both the stimulus and the EEG, we would be interested in the greatest common divider of 44,1 kHz and 512 Hz. This divider is equal to 4 Hz, which unfortunately does not fulfill the Nyquist criterion. In this case, aliasing would thus occur.

Chapter 4

Linear regression model

In this chapter, a stimulus reconstruction model is made with the preprocessed data. We here adopt a linear model, resulting in set of decoder weights that relate the EEG response to the stimulus. To decouple the training and test phase of this model, cross-validation is implemented. Due to the risk of overfitting in the training phase, ridge regression is included in this model. A second inner loop (apart from the outer cross-validation loop) is introduced for the estimation of the optimal regularization hyperparameter.

The calculation of the decoder weights in this double loop setting will be explained in detail in this chapter (Figure 4.1), after which the obtained results are discussed. Moreover, group effects within the results will be studied by means of linear mixed-effects models for the perception and imagination experiments.

4.1 Stimulus reconstruction model

The decoder weights in this model are used to reconstruct a stimulus from an EEG response (decoding approach) and are calculated via linear regression, similar to [31, 44]. We consider a stimulus envelope $s(t)$ and an EEG response $\mathbf{r}(t) = \begin{bmatrix} r_1(t) & \cdots & r_N(t) \end{bmatrix}^T$, consisting of N channels. These channels can be linearly combined using decoder weights $\mathbf{g}(t) = \begin{bmatrix} g_1(t) \cdots & g_N(t) \end{bmatrix}$ to produce an estimate of the stimulus envelope $\hat{s}(t)$. To increase the model capacity, this relation is extended to a spatio-temporal filter by introducing a range of time lags τ for each EEG channel.

$$\hat{s}(t) \quad = \quad \sum_{i=1}^{N} \sum_{\tau} g_i(\tau) r_i(t \ - \ \tau) \tag{4.1}$$

We choose time lags ranging from $\tau_0 = 0$ ms to $\tau_n = 250$ ms, which has shown to cover most of the neural processing of an audio stimulus [31]. At a sampling rate of 64 Hz, this range thus spans 17 samples. However, in this decoding model, we approach the neural processing pathway in the opposite sense. Envisioned on a timescale, we try to decode an earlier stimulus out of a later response. Therefore, in backward modeling, we need to induce a negative time lag in the spatio-temporal

filter to ensure the incorporation of relevant EEG events [31]. When reconstructing T time samples of the stimulus, with time lags ranging from the above mentioned τ_0 to τ_n, (4.1) generalizes to the following matrix formulation:

$$\mathbf{S} = \mathbf{Rg} \text{ with}$$

$$\mathbf{R} = \begin{bmatrix} r_1(\tau_n) & \cdots & r_1(T) & 0 & \cdots & \cdots & 0 \\ r_2(\tau_n) & \cdots & r_2(T) & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_1(\tau_0) & \cdots & r_1(T-1) & r_1(T) & 0 & \cdots & 0 \\ r_2(\tau_0) & \cdots & r_2(T-1) & r_2(T) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}^T, \quad \mathbf{g} = \begin{bmatrix} g_1(\tau_n) \\ g_2(\tau_n) \\ \vdots \\ g_1(\tau_0) \\ g_2(\tau_0) \\ \vdots \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} S(0) \\ S(1) \\ S(2) \\ S(3) \\ \vdots \\ S(T) \end{bmatrix} \tag{4.2}$$

The decoder weights $\mathbf{g}$ are optimized via the MMSE criterion, in which the mean squared error between the stimulus and its reconstruction is minimized, leading to (4.3) for a stimulus length of T samples [31, 44]:

$$\begin{aligned} \mathbf{g} &= \arg\min_{\mathbf{g}}\left(\sum_0^T (\mathbf{S}-\mathbf{Rg})^T(\mathbf{S}-\mathbf{Rg})\right) \\ \Rightarrow \mathbf{g} &= (\mathbf{R}^T\mathbf{R})^{-1}(\mathbf{R}^T\mathbf{S}) = \mathbf{C}_{\mathbf{RR}}^{-1}\mathbf{C}_{\mathbf{RS}}. \end{aligned} \tag{4.3}$$

$\mathbf{C}_{\mathbf{RR}}$ and $\mathbf{C}_{\mathbf{RS}}$ are respectively the estimated autocorrelation matrix of the time-lagged EEG channels and cross-correlation matrix between the time-lagged EEG and the stimulus envelopes. After obtaining the decoder $\mathbf{g}$, a reconstructed envelope can be found via (4.2). However, there is a risk of overfitting in this optimization. This effect gets larger when a lot of parameters need to be trained, i.e., when the number of channels $\times$ time lags is large, while the amount of training data is limited.

In this problem, we encounter a high number of time-lagged EEG channels, while only a small pool of training examples is usually available to calculate the decoder $\mathbf{g}$ in research. This limited amount of data leads to badly estimated and low rank correlation matrices in the inverse problem (4.3) [44], resulting in decoder weights that produce an accurate reconstructed envelope for the used training examples, however, by also giving attention to 'irrelevant' attributes in the model. This does unfortunately not mean that the model works equally well for new, unseen data. In other words, overfitting can lead to a biased interpretation of the results.

Therefore, ridge regression is often added to the estimation of the decoder (4.4). A hyperparameter λ is introduced into the calculations, penalizing the square magnitude of the decoder weights [44]. This drives weights down if they only marginally contribute to the decoding process, but still deviate much from the zero value. Unnecessary decoder weights are thus regulated towards a value close to zero, giving attention to the real trends in the data and thereby reducing overfitting.

$$\begin{aligned} \mathbf{g} &= \arg\min_{\mathbf{g}}\left(\sum_0^T (\mathbf{S}-\mathbf{Rg})^T(\mathbf{S}-\mathbf{Rg}) + \lambda\mathbf{g}^T\mathbf{g}\right) \\ \Rightarrow \mathbf{g} &= (\mathbf{R}^T\mathbf{R}+\lambda\mathbf{I})^{-1}(\mathbf{R}^T\mathbf{S}) = (\mathbf{C}_{\mathbf{RR}}+\lambda\mathbf{I})^{-1}\mathbf{C}_{\mathbf{RS}} \end{aligned} \tag{4.4}$$

The two covariance matrices $\mathbf{C_{RR}}$ and $\mathbf{C_{RS}}$ in (4.4) can be estimated from the available training data. However, the optimal value of the hyperparameter λ cannot be directly deduced from envelope and EEG examples, it requires its own optimization strategy. Optimizing this new hyperparameter thus introduces a trade-off between an overall better stimulus reconstruction and time complexity.

4.2 Leave-one-song-out decoding algorithm

In this study, we explore if a song can be decoded when the model is trained on every other song in the same perception/imagination condition. This is done for one subject and condition at a time. For simplicity, we assume a perception condition further on in this section. The available data then consists of 12 songs x 5 trials = 60 unique trials per subject for the OpenMIIR data set or 10 songs x 3 trials = 30 trials for the Bach data set (Table 4.1). Each of these trials contains a stimulus envelope and its EEG response.

The available data is used for training and testing of the model. To accurately measure the performance of our stimulus reconstruction, we cannot train and test the model on the same data, this would create biased results. The test and training phases need to be decoupled. A cross-validation (CV) loop is therefore implemented, which splits the available data into a test and training part in subsequent folds (Figure 4.1). In this CV loop, we withhold the envelope and EEG trials of one song per fold for testing and use the rest for model training. This results in five test trials for the OpenMIIR data set per fold, or three for the Bach data set (Table 4.1).

For the training of a decoder without regularization ($\lambda = 0$), this CV loop is the only division in the data that needs to be made. However, when including a hyperparameter λ, this value also needs to be optimized. Since this optimization is dependent on the used data, a second inner CV loop is foreseen, nested into the outer one. This inner CV loop further divides the training data of the outer CV loop into subfolds, splitting off the available trials for one other song for validation (Table 4.1). In the next subsections, the calculations in this nested loop system will be explained in detail. A schematic depiction of this can be found in Figure 4.1.

TABLE 4.1: Data in the outer and inner CV loops during the study of perception.

OpenMIIR data set		
Total	Outer loop	Inner loop
5x12 trials	→ 5x1 trials (test)	
	→ 5x11 trials	→ 5x1 trials (validation)
		→ 5x10 trials (training)
Bach data set		
Total	Outer loop	Inner loop
3x10 trials	→ 3x1 trials (test)	
	→ 3x9 trials	→ 3x1 trials (validation)
		→ 3x8 trials (training)

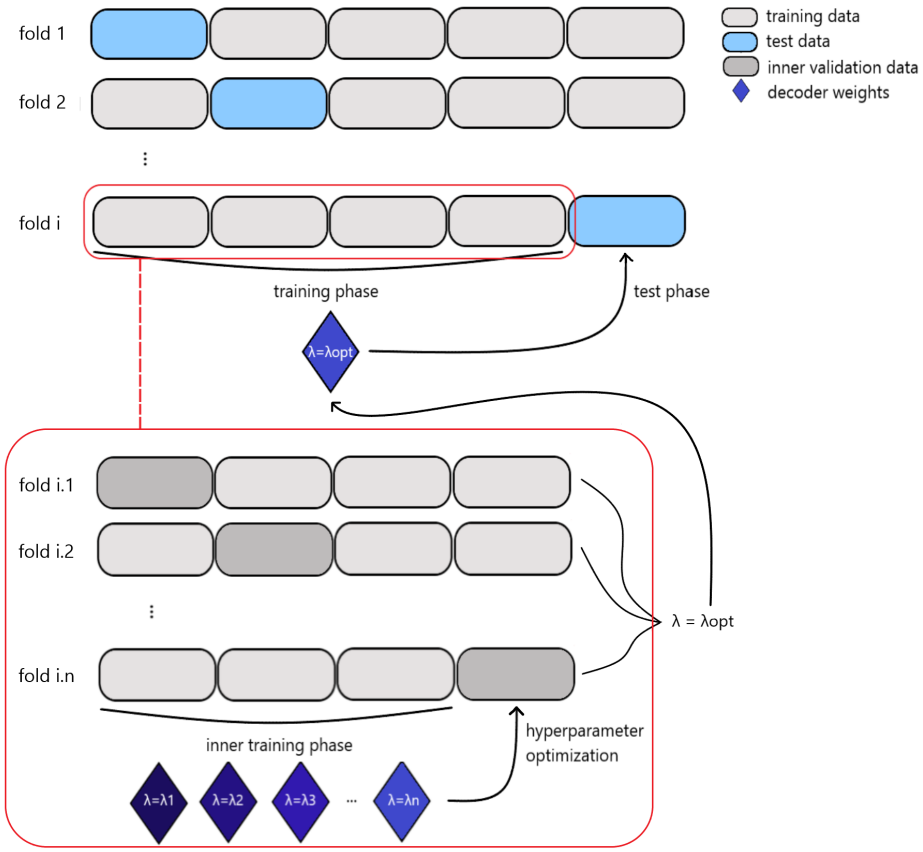

Figure 4.1: The model is trained and tested using two nested CV loops, where the inner loop (below) is used to determine the regularization parameter, while the outer loop (above) is used to test the optimized decoder.

4.2.1 Set-up of the outer CV loop

As a first step, all the available trials are standardized. This optional step makes sure that all variables in the model (i.e., the time-lagged EEG channels) are treated on the same scale. These calculations benefit the direct comparison of decoder weights and the interpretability of the model. Afterwards, the outer CV loop is created, splitting off one song per fold (Table 4.1) for testing. All other trials are training data. When including regularization into our model (4.4), the training data will be used in the following inner loop for hyperparameter optimization [44].

4.2.2 Inner CV loop - Hyperparameter optimization

As mentioned before, the optimal hyperparameter value λ cannot be deduced directly from the data. Additional optimization is thus required, which takes the form of a nested inner loop on the training data. For every new subfold in this nested loop, we again split off one song (Table 4.1). All remaining trials are used to make decoders with a range of possible hyperparameter values, which will be validated against the newly separated song. The subdivisions in the data for this inner CV loop will further be called 'inner validation data' and 'inner training data' respectively.

To make decoders with the inner training data (4.4), a few components need to be determined: the covariance matrices $\mathbf{C_{RR}}$ and $\mathbf{C_{RS}}$ and the hyperparameter value λ. For the hyperparameter, a range of values is used [44], which will each be implemented in a separate decoder:

$$\lambda = 1.848^n \times 10^{-6} \quad \text{with} \quad n = [0, 50] \tag{4.5}$$

Next, the $\mathbf{C_{RR}}$ and $\mathbf{C_{RS}}$ matrices for each inner training trial are estimated (4.4). However, to include the information of all inner training trials into a more accurate training of the decoders, a combination of these covariance matrices is made. This combination can be done in two ways:

1. Keeping the N covariance matrices separate for every inner training trial, calculating decoder weights for each and averaging the decoders in a final step.

$$\mathbf{g} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{C_{RR}}_i + \lambda\mathbf{I})^{-1}\mathbf{C_{RS}}_i \tag{4.6}$$

2. Averaging the N covariance matrices of all inner training trials as a first step, after which one decoder is directly obtained. Equivalently, one can use a concatenation of all inner training data instead of averaging [3].

$$\bar{\mathbf{C}}_{\mathbf{RR}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{C_{RR}}_i \quad \text{and} \quad \bar{\mathbf{C}}_{\mathbf{RS}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{C_{RS}}_i$$
$$\mathbf{g} = (\bar{\mathbf{C}}_{\mathbf{RR}} + \lambda\mathbf{I})^{-1}\bar{\mathbf{C}}_{\mathbf{RS}} \tag{4.7}$$

In the research of Biesmans et al. [3], it was shown that the second approach leads to a more accurate result than the first method. The matrices $\mathbf{C_{RR}}$ and $\mathbf{C_{RS}}$ in (4.6) are estimated with only a limited amount of data (one trial of relatively short duration). This leads to an ill-posed inverse problem and an inaccurate decoder [3]. However, in (4.7), the information of multiple trials is combined in the training of one single decoder, which produces much better results. The second approach will therefore also be the used method.

After making one decoder for every hyperparameter in the range (4.5) per subfold, these decoders are validated against the inner validation data. With each of these decoders, we reconstruct an envelope out of every inner validation EEG trial (4.2). Then, the Pearson correlation between the real inner validation envelope and these reconstructed envelopes is calculated.

When this calculation of Pearson correlations between the real envelopes and obtained reconstructions is continued for every subfold, the optimal hyperparameter can then be deduced as the value which leads to the highest averaged Pearson correlation over all subfolds.

4.2.3 Outer CV loop - Testing

Now that the optimal hyperparameter value λ is found, the full optimal decoder still needs to be tested against the test data that was separated in the outer loop. Therefore, the $\mathbf{C_{RR}}$ and $\mathbf{C_{RS}}$ matrices of all training data in the outer loop are combined (4.7), with the optimal value for the hyperparameter λ. With this optimal decoder, the EEG trials of the test data are mapped to their respective reconstructed envelopes (4.2). Finally, by calculating the Pearson correlation between each reconstructed and original envelope for each test fold, the performance of our algorithm is quantified.

4.2.4 Statistical significance

Permutation test

To check the statistical significance of the Pearson correlations obtained during validation, a permutation test is performed for all subjects combined. The significance level for our correlations is obtained as follows.

If an envelope is correlated with its reconstruction, ideally a high correlation would be expected. However, if the indices of the reconstructed envelope are shuffled, we would expect this value to be lower, since the relation in the data is broken up. The correlation value in the second case is dependent on the performed permutation of indices. If this random permutation is done many (for example 10000) times, we can estimate the distribution of correlation values. From this distribution, the 97.5 percentile can be deduced, which can be used as the desired significance level (on an α-level of 0.05). To see if our correlation is statistically significant, we then simply need to compare its value to this 97.5 percentile [43].

In the discussion of the results, median Pearson correlations over the (five or three, Table 4.1) available trials per song and subject will be used. The plotted significance level will therefore also be deduced from the estimated distribution of these median correlations for all subjects.

Wilcoxon signed rank test

To compare two results of for example different conditions, the Wilcoxon signed rank test is used. In this nonparametric test, two groups of observations x and y are compared by looking at the distribution of their difference. The null and alternative hypotheses of this test (implemented as a one-sided test) are: [24]

$$\begin{aligned} H_0 &: \text{Median}(x - y) = 0 \\ H_1 &: \text{Median}(x - y) > 0 \end{aligned} \tag{4.8}$$

When a certain significance level α is chosen (here, $\alpha = 0.05$ will be used), the null hypothesis can either be retained or rejected, based on the obtained the p-value.

4.3 Results and discussion

The leave-one-song-out decoding algorithm was implemented on the two available data sets, for both data filtering ranges in the preprocessing steps (1-10 Hz and 1-30 Hz). All conditions, i.e., perception or imagination, are tested separately.

For the perception case, a literal implementation of Section 4.2 was done, solely with perception trials. This implementation follows exactly the division of test and training data in Table 4.1 for both data sets. However, a slight variation was implemented for imagination experiments. Here, the perception trials of the OpenMIIR data set were used for training (Table 4.2), to avoid that possibly misaligned and distorted imagination trials lead to an inaccurate decoder. This approach removes the explicit need for a nested loop construction. Both CV loops can be decoupled, because they each use different data. This leads to a single global decoder per subject, which can be tested with each of the available imagination trials.

Table 4.2: Data in the outer and inner CV loops during the study of imagination.

OpenMIIR data set		
Total	'Outer loop'	'Inner loop'
2x5x12 trials	→ 5x12 trials (imagination, test)	
	→ 5x12 trials (perception)	→ 5x1 trials (validation)
		→ 5x11 trials (training)

4.3.1 Results for the OpenMIIR data set (bandpass filter 1-10 Hz)

The resulting correlations for a bandpass filtering between 1-10 Hz are given in Figure 4.2 and 4.3 for the perception (gray) and an imagination experiment (condition 2, red). Each of these displayed values is calculated as the median correlation over the five trials belonging to the same song and subject. Also the median value per song is depicted, along with its numerical value in the table on the right hand side of the figure. The significance level of these correlations is calculated using a permutation test. A further subdivision of the results per subject is given in Appendix B. The overall median correlations and their variance can be seen in Table 4.3.

Table 4.3: Overall results for the OpenMIIR data set (bandpass filter 1-10 Hz).

	Perception	Imagination (2)	Imagination (3)	Imagination (4)
Overall median	0.09	3.3×10^{-3}	-2.7×10^{-3}	-1.2×10^{-3}
Variance	0.014	9.9×10^{-3}	11×10^{-3}	0.013

Stimulus perception

For the perception condition with a bandpass filtering between 1-10 Hz, the resulting correlations can be seen in Figure 4.2 (gray). The overall median and variance of the results can be found in Table 4.3. A few observations can be made from these results.

The median values (gray filled 'dots') show that for all songs, at least half of the median Pearson correlations ρ are significant. Looking at Appendix B, P06 and P10 are the best performing subjects ($\rho_{P06} = 0.1761$, $\rho_{P10} = 0.1615$) in this experiment, P03 shows the lowest performance ($\rho_{P03} = 0.0182$) and is the only subject that does not meet the significance level for at least half of the songs.

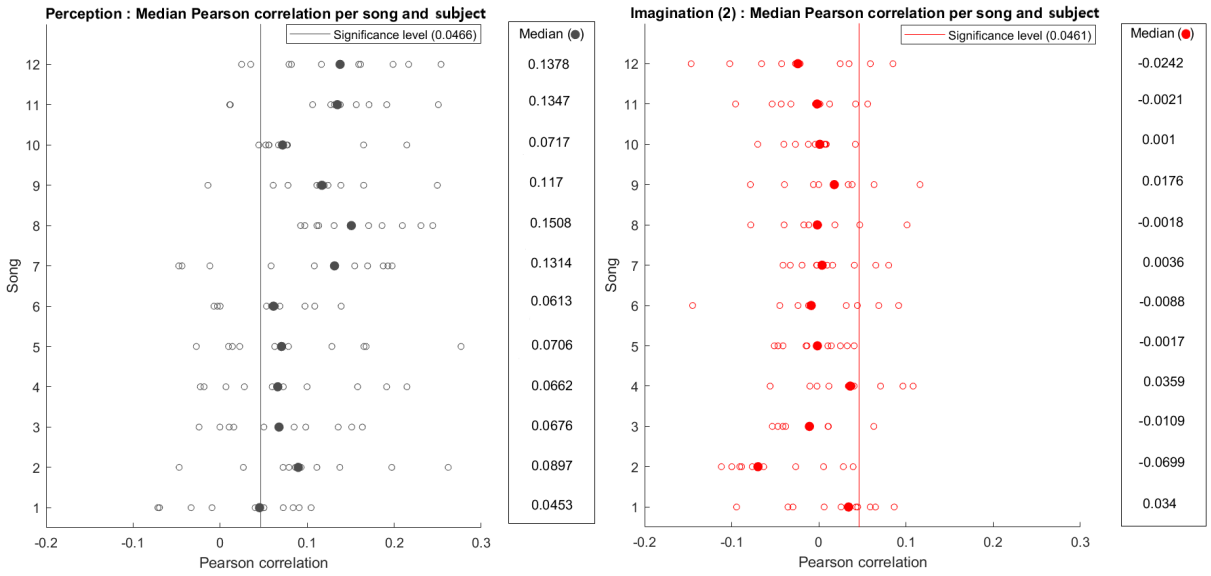

Figure 4.2: In the case of perception, all the median correlations per song surpass the significance level. For imagination, the contrary is true.

Secondly, there might be a trend visible when looking at song categories. Songs 8-12 seem to perform overall better in this algorithm than the other songs, largely overlapping with the category of instrumental songs (9-12, Table A.1). Songs in the categories 'lyrics' and 'no lyrics' seem to have slightly worse results, except for song 8. This observation will be statistically verified in Section 4.4.

Stimulus imagination: condition 2 (preceded by cue clicks)

The results for imagination experiment (2), preceded by cue clicks, can also be viewed in Figure 4.2 (red) for a bandpass filtering between 1-10 Hz. The median and variance of the correlations in this condition are far lower than for perception (Table 4.3).

In these figures, the better performance of songs 8-12 during perception is not seen during imagination and by looking at Appendix B, all subjects are scoring low. The median values per song and per subject do not surpass the significance level. This overall worse performance is confirmed by conducting a Wilcoxon signed rank test on the median Pearson correlations of both conditions in Figure 4.2 ($\rho_{perc} > \rho_{imag2}$, $p < 10^{-15}$).

A possible reason for this observation could be a distortion of the stimuli during imagination. Tempo inconsistencies, for example, could influence these results. Song features might appear at different time points in the original stimulus than in the reconstruction, decreasing their correlation. While there are cue clicks present in this experiment, which give a hint of the tempo beforehand, they are not played during imagination. Therefore, this distortion is still highly likely to be present in the EEG trials for testing.

Moreover, we are uncertain about the onset of the imagination trials. All EEG trials are recorded starting from a certain point in time, but this instance does not necessarily align with the actual onset of imagination. A strategy to find the optimally reconstructed envelopes will be explained in the next chapter.

Stimulus imagination: condition 3 and 4 (without preceding cue clicks)

In the OpenMIIR data set, two imagination experiments without preceding cue clicks were available (Section 2.1.1). The third experiment was recorded together with the first two (Figure 2.1), the fourth experiment was conducted separately. The results for these conditions can be seen in Figure 4.3. Their overall median correlation and variance are given in Table 4.3.

We see similar result as for the previous imagination condition: the song medians (filled 'dots') are all situated at lower values than the significance level, indicating that more than half of the depicted correlations are not significant. For both conditions, the results are much lower than for perception (Wilcoxon signed rank test, $\rho_{perc} > \rho_{imag3}$, $p < 10^{-16}$, $\rho_{perc} > \rho_{imag4}$, $p < 10^{-15}$). For this, the above reasons (condition 2) apply.

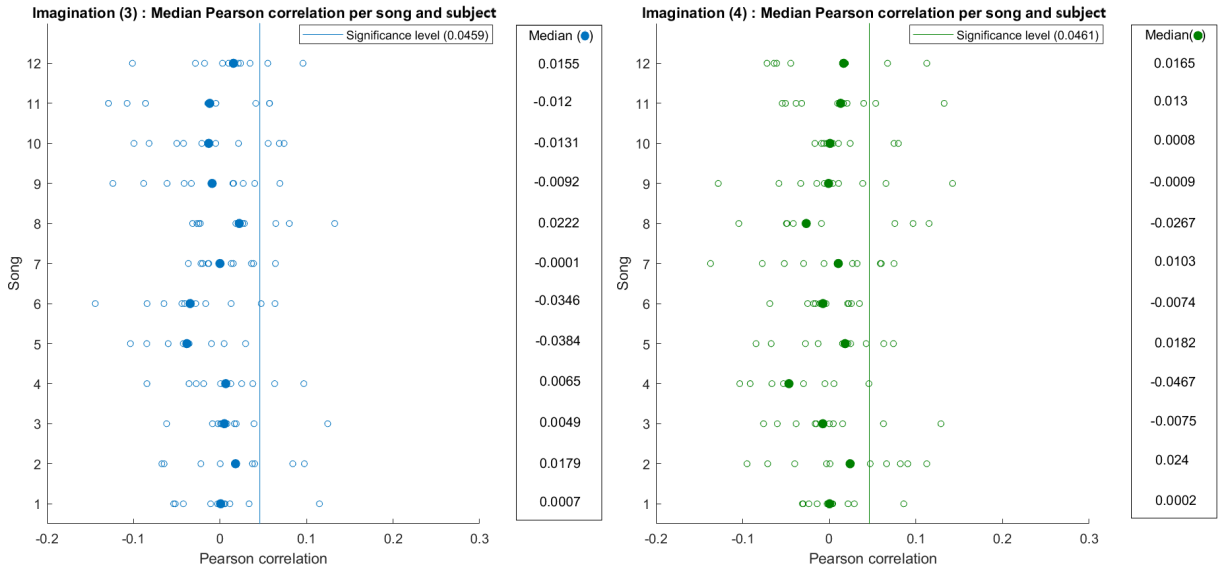

FIGURE 4.3: For the other two imagination experiments, the same conclusions as for the second imagination condition can be drawn.

All three imagination conditions have similar results: they produce significant correlations for less than half of the subjects per song. This does also mean that, from these observations, no significant difference could be found between the experiments with and without preceding cue clicks (Wilcoxon signed rank test, $p = 0.91$ when comparing ρ_{imag2} to ρ_{imag3}, $p = 0.82$ when comparing ρ_{imag2} to ρ_{imag4}).

In Section 2, it was mentioned that the fourth condition, recorded separately from all other experiments, could be used to spot possible effects of song memorization (i.e., subject training). Based on the observations and previously conducted tests, there is no statistically significant training of subjects between experiments.

4.3.2 Results for the OpenMIIR data set (bandpass filter 1-30 Hz)

Similarly, the results for the data with bandpass filter between 1-30 Hz can be analyzed. The overall median Pearson correlations for these conditions can be found in Table 4.4. Both the overall median and variance of our results are lower compared to those for a bandpass filtering between 1-10 Hz. From Appendix B, the same observations can be made about the best/worst performing songs and subjects, but with overall lower correlations than for the previous filtering range.

Table 4.4: Overall results for the OpenMIIR data set (bandpass filter 1-30 Hz)

	Perception	Imagination (2)	Imagination (3)	Imagination (4)
Overall median	0.08	-1.1×10^{-3}	-3×10^{-3}	-6.35×10^{-4}
Variance	0.011	8.3×10^{-3}	9.2×10^{-3}	0.01

A formal Wilcoxon signed rank test between the results of the two filtering ranges gives us an important insight. When comparing the median correlations of the two frequency ranges in the case of perception, a significant difference ($p = 1.3 \times 10^{-5}$) is found, rejecting the null hypothesis. In other words, the bandpass filtering between 1-10 Hz produces significantly better correlations than the filtering between 1-30 Hz. For the three imagination conditions, this difference is not found. As in the previous experiments, these imagination conditions produce low results.

This result suggests that the neural processing of music holds its most predominant information in the same frequency range as for speech processing, namely the delta (<4Hz), theta (4-8 Hz) and alpha (8-10 Hz) bands present in the EEG response [18]. The effect of filtering range will be investigated further as we extend our stimulus reconstruction model in the next chapter.

4.3.3 Comparison to the original study of Stober et al. [36, 34]

In the original study of Stober et al. [36, 34], a direct implementation of the linear model in the study of O'Sullivan et al. [31] was made, in the case of perception. The average correlation value for trial-specific decoders was 0.11, with a very high variance of 0.52. For this algorithm, no statistical significance of the correlations

was found. After obtaining these results, an algorithm using neural networks was proposed in the work of Stober et al. [36, 34].

In our study, the median value of 0.09 stays in line of this average correlation value. Our variance is on the other hand much lower. A proposed reason for this outcome given in the study of Stober et al. [34] was the instability of the decoder weights induced by the use of short trials (of a few seconds, Table A.1). The decoders were also highly dependent on the chosen time lags τ. While it is not literally mentioned in the study, this proposed reason might subsume that a combination of decoders as in Equation 4.6 was used, leading to an unstable decoder.

Moreover, the study of Stober et al. [34] does not include regularization into the model, which will definitely influence the variance in the obtained results. As a solution, dimension reduction with PCA was done to decrease overfitting, but the amount of time lags included in the model was high (τ = 0 - 350 ms), leading to a large amount of parameters for training.

4.3.4 Results for the Bach data set (bandpass filter 1-10 Hz)

The single perception condition in the Bach data set produces the results in Figure 4.4 on the left for a bandpass filtering between 1-10 Hz. Comparing these perception results to the ones for the OpenMIIR data set ($\rho_{perc,O}$), we see that the Bach data set ($\rho_{perc,B}$) has a significantly better performance (Wilcoxon rank sum test, $\rho_{perc,B} > \rho_{perc,O}$, $p < 10^{-9}$).

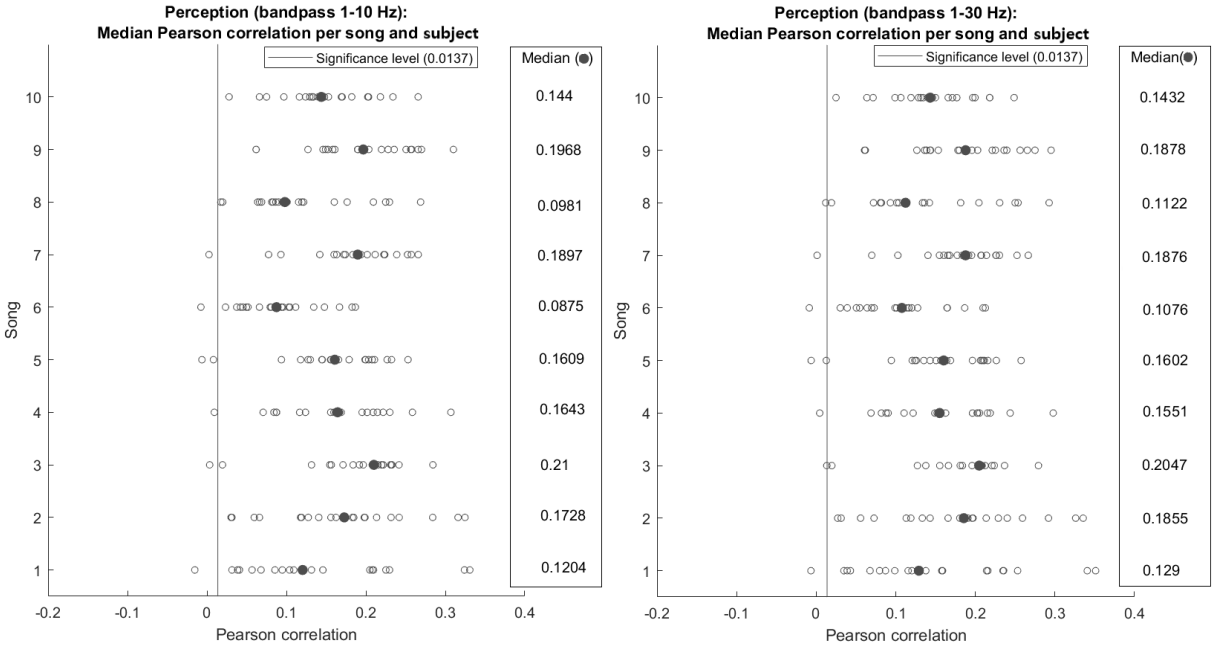

FIGURE 4.4: The results for the Bach data set for both filtering ranges show a significance for more than half of the subjects per song.

A reason for this could be the difference in song categories. By visual inspection of the results in the OpenMIIR data set, it was seen that the instrumental songs seemed to outperform the other two categories. The Bach data set contains only instrumental music, which could influence the relative performance of the model. Moreover, trial length could play a role in the difference between these results. Contrary to the OpenMIIR data set, these stimuli each have a duration of a few minutes (Table A.1). This incorporates more information into the training phase than short stimuli, hence more accurately trained decoders.

4.3.5 Results for the Bach data set (bandpass filter 1-30 Hz)

The results for a bandpass range of 1-30 Hz are depicted in Figure 4.4 on the right. A visual inspection of the results for both filtering ranges does not show an eye-catching difference between these correlations, both preprocessing approaches present good results. A Wilcoxon signed rank test between the median Pearson correlations of both filtering ranges indicates no significant difference on an $\alpha = 0.05$ level ($p = 0.086$).

4.3.6 Comparison to the original study of Di Liberto et al. [11, 12]

In the original study of Di Liberto et al. [11], similar values for the correlation were found. In this study, larger filter ranges (0.5-45 Hz) and higher sampling rates are used (128 Hz). Our results thus show a similar performance for smaller frequency ranges (even only up to 10 Hz), thus achieving equal results while incorporating less frequency information into our model.

Table 4.5: Overall results for the Bach data set

	Bandpass 1-10 Hz	Bandpass 1-30 Hz
Overall median	0.158	0.156
Variance	6.8×10^{-3}	6.6×10^{-3}

4.4 Investigating effects on groups within the data

In this section, we want to study possible effects within the correlation results. To do this, one needs to consider the hierarchical structure of the used data sets, which leads to many possible groupings. Trials belong to a certain subject and stimulus. These songs can for example contain lyrics or be purely instrumental. Subjects can be musically trained or not. All possible categorical groupings can be found in Table A.2. One can then pose an interesting question: are there certain groups in the available data which perform significantly better than others?

We will investigate these effects in the data using a linear mixed-effects model (LME) [23]. Here, our obtained Pearson correlations ρ will be fitted using categorical predictor variables, which each can have multiple grouping levels. This relation (4.9)

is based upon two different types of variables: fixed and random effects. Fixed effects (with design matrix X and fixed-effects vector β) globally reveal tendencies in the observations and can be used to study trends in which we are interested. Random effects (with design matrix Z and random-effects vector b) cause random variations within the data, which are not explained by the fixed effects. The fitting error is represented by ϵ, which is normally distributed with a variance σ^2 [23].

$$\boldsymbol{\rho} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} = \mathcal{N}(0, \sigma^2) \tag{4.9}$$

For the perception experiments in both data sets, possible variables of interest (fixed effects) are song categories and musicality of the subjects. Since there are two ways to describe musicality in the data sets (musical training and instrument practice, Table A.2), both effects will be investigated. For the imagination experiments in the OpenMIIR data set, it would additionally be interesting to investigate if different techniques of imagining a song (how do they 'hear' the lyrics inside their heads, possible visualizations) influence the results.

With these variables, several candidate models can be made. The optimal LME model will be chosen by computing the Akaike criterion (AIC) for each and selecting the model with the lowest value. Furthermore, plots of the residuals were made to assert their normality assumption and homoscedasticity of the model. In the next subsections, the contents of the best performing models for perception and imagination will be discussed, along with their respective observations.

4.4.1 Results for the OpenMIIR dataset - Perception

For perception experiments, we find the best performing model in Table 4.6. This model uses instrument practice to define the musicality of subjects. After computing the Akaike criterion for all candidate models using the description 'musical training', these models were suboptimal. This description of musicality also never showed statistical significance between its grouping levels.

The perception LME model in Table 4.6 has the lowest AIC value of all candidate models (AIC = -957.03). A formal comparison between the perception LME model in Table 4.6 and its equivalent candidate model using 'musical training' shows that the former is significantly better (Likelihood Ratio Test, DF1 = 7, AIC1 = -957.03, DF2 = 8, AIC2 = -955.16, p = 0.049).

When fitting this LME model to our correlations of all subjects, it was found that correlations are significantly higher for instrumental music than for other song categories (estimated coefficient =0.0395, t=1.998, p=0.046), which confirms the visual inspections we did in Figure 4.2. Furthermore, musician who are still playing instruments have significantly worse (coefficient estimate=-0.0549, t=2.17, p=0.03) correlations in comparison to subjects who never played an instrument or have stopped (Section 5.5). For a preprocessing filter range of 1-30 Hz, the effect of instrument practice persists, however, there is no significant effect found for instrumental music (DF=595, t=1.6365 p=0.10226).

Table 4.6: Fixed and random effects for the perception and imagination experiments.

	Fixed effects	**Random effects**
Perception	Musicality(2): Instrument practice Song category (only OpenMIIR data set)	Subject (Intercept) Song (Intercept)
Imagination	Musicality(2): Instrument practice Visualizations Imagination technique (lyrics) Song category	Subject (Intercept) Song (Intercept)

4.4.2 Results for the OpenMIIR data set - Imagination

For the three imagination experiments in the OpenMIIR data set, the fixed effects are extended with imagination techniques. Subjects were reported to have two types of imagination approaches for both music with and without lyrics. For the former, people either hear themselves singing or hear the lyrics inside their heads. For music without lyrics, any presence of visualizations during imagination is also investigated (Table 4.6).

By fitting this model on data of the three imagination experiments, no significant effects were found. This result is not completely unexpected, since the median correlations for these experiments (Figure 4.2 and Figure 4.3) are largely non-significant.

4.4.3 Results for the Bach data set

In this data set, there is only one song category: instrumental music. This effect can therefore not be tested by the model. This means that only one fixed effect is present in the model for this data set: the instrument practice for a subject. By analogy to the previous data set, an LME model is fitted where no significant influence of instrument practice is found in the data (estimated coefficient =0.009, t=0.34, p=0.73).

For a preprocessing filter range of 1-30 Hz, this outcome does not change. In the original study of Di Liberto et al. [11], there was a significant effect of musicality found with a bandpass filtering range of 1-45 Hz, but this could not be replicated with the current set-up in this study.

Chapter 5

Inclusion of time shifts

Up to this point, trials are included in the reconstruction model, starting from their onset, i.e., from the instance t = 0. By doing this, we assume that a noticeable processing of a song could become visible at time t=0 in an EEG response. However, as for example seen in the imagination experiments of Chapter 4, this does not necessarily lead to an optimally reconstructed envelope.

A new global parameter ν will be added to the model of Chapter 4, which incorporates the possibility of an EEG response latency in our reconstruction model. This would mean that a possible latency in response, because of an extensive neural processing, is included. Moreover, in imagination experiments, a possible mismatch between the trial onset and actual onset of imagination can be quantified. In this chapter, an altered leave-one-song-out decoding approach, which includes this time shift parameter ν, will be explained, along with a discussion of the results with respect to the two data sets.

5.1 The importance of latency

When a stimulus is presented to a subject, signals ascend the neural pathway. The processing of these signals in the ear and brain might result in a latency of the neural response with respect to the stimulus presentation.

We consider an example EEG response, from which we want to reconstruct an envelope via (4.2), along with a trained decoder. The spatio-temporal filtering, with time lags τ, of our model then includes shifted versions of the EEG channels into the stimulus reconstruction.

However, if a possible latency between the stimulus and EEG response is not fully incorporated into the calculations, i.e., when this latency is larger than the maximal time lag τ_n, the decoding approach of Chapter 4 does not lead to an optimally reconstructed envelope. Calculations are then filled with noise and artifacts from other brain processes, which are not related to musical activity. Moreover, for imagination experiments, the exact onset of the imagined response is unknown. All EEG trials are recorded starting from a certain point in time, but this instance does not necessarily align with the actual onset of imagination.

For these reasons, a possible latency will be incorporated into an altered version of the linear model of Chapter 4. To decouple this latency from the time lags τ of the spatio-temporal filter, a new global parameter ν will be introduced [10]. In the next section, this adapted version of the stimulus reconstruction will be discussed.

5.2 Adapted stimulus reconstruction model

For this linear model, we again consider a stimulus envelope $s(t)$ and an EEG response $\mathbf{r}(t) = \begin{bmatrix} r_1(t) & \cdots & r_N(t) \end{bmatrix}^T$, consisting of N channels. However, when reconstructing a stimulus, not only a spatio-temporal filtering of the EEG response is done, also a global time shift parameter ν is now introduced. The stimulus reconstruction with the trained decoder weights $\mathbf{g}(t) = \begin{bmatrix} g_1(t) \cdots & g_N(t) \end{bmatrix}$ from (4.1) and the reconstructed envelope $\hat{s}(t)$ now becomes:

$$\hat{s}(t) = \sum_{i=1}^{N} \sum_{\tau} g_i(\tau) r_i(t - \tau + \nu). \tag{5.1}$$

When reconstructing T time samples of the stimulus, with time lags ranging from τ_0 = 0 ms to τ_n = 250 ms and with a time shift ν, (5.1) generalizes to the following matrix formulation:

$$\mathbf{S} = \mathbf{R}\mathbf{g} \quad \text{with}$$

$$\mathbf{R} = \begin{bmatrix} r_1(\tau_n + \nu) & \cdots & r_1(T + \nu) & 0 & \cdots & \cdots & 0 \\ r_2(\tau_n + \nu) & \cdots & r_2(T + \nu) & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_1(\tau_0 + \nu) & \cdots & r_1(T - 1 + \nu) & r_1(T + \nu) & 0 & \cdots & 0 \\ r_2(\tau_0 + \nu) & \cdots & r_2(T - 1 + \nu) & r_2(T + \nu) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}^T,$$

$$\mathbf{g} = \begin{bmatrix} g_1(\tau_n) \\ g_2(\tau_n) \\ \vdots \\ g_1(\tau_0) \\ g_2(\tau_0) \\ \vdots \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} S(0) \\ S(1) \\ S(2) \\ S(3) \\ \vdots \\ S(T) \end{bmatrix}. \tag{5.2}$$

As can be seen in (5.2), no inclusion of the time shift is implemented in the decoder weights $\mathbf{g}$. This means that the decoder can be found in the same way as before (4.4), using a time-lagged EEG matrix $\mathbf{R}$ that does not include a time shift ν.

However, in the leave-one-song-out decoding algorithm, a new parameter ν now appears, which needs to be optimized. The adaptation of this algorithm and its double loop mechanism will be discussed in the next section.

5.3 Adapted leave-one-song-out decoding algorithm

In the adaptation of the leave-one-song-out decoding algorithm, a new type of parameter needs to be optimized: the time shift ν, which incorporates the possibility of a latency into the reconstruction model. For this parameter, a range of values implemented, from which the optimal latency will be deduced.

In a study performed by de Cheveigné et al. [10], optimal time shifts ν of approximately one second after the presentation of a speech stimulus were found. We choose time shifts ranging from $\nu = 0$ - 2 s for trials in the OpenMIIR data set, which broadly incorporates the latency found in this study. At a sampling rate of 64 Hz, this range thus spans 129 samples. For the Bach data set, only one second of additional EEG data is provided, which naturally forms the maximal time shift for the implementation of the model. For the Bach data set, this range thus spans 65 samples.

The training of the decoder weights follows the same double loop implementation as before. In the outer CV loop, the trials of one song are split off to use as test data, all other songs will be used in the training phase. The same implementation of the inner CV loop, without time shifts ν, then follows for the hyperparameter optimization.

However, the test phase will be altered slightly. Previously, we simply tested the optimized decoder weights by reconstructing the envelopes of each test trial, after which the Pearson correlation between these reconstructed envelopes and the real song envelope was calculated. Now, the test EEG responses will be shifted over the specified range for the parameter ν, testing the decoder weights on each of these shifted trials. For every test example in the outer CV loop, a range of Pearson correlations is thus obtained, one for each time shift. In comparison, the single correlation per trial obtained in the approach of Chapter 4 is equal to the correlation found for $\nu = 0$ s.

Finally, the optimal time shift ν_{opt} per trial is found as the value linked to the highest Pearson correlation in the obtained range. Note that the time shift is thus optimized per individual trial, so that for each the reconstructed envelope is obtained, which produces the maximal Pearson correlation with respect to the corresponding real envelope.

5.4 Results and discussion

For the implementation of this adapted algorithm on the different perception and imagination conditions, the same procedures as in Section 4.3 are followed.

For the perception experiments, only perception data is used for the entire decoding process. For imagination experiments, perception data is again used for the training of the decoder.

5.4.1 Results for the OpenMIIR data set

The resulting correlations for a bandpass filtering between 1-10 Hz are given in Figure 5.1 for all four conditions in the OpenMIIR data set. The significance level of these correlations is calculated using a permutation test. The overall median correlations, their variance and median optimal time shift are given in Table 5.1.

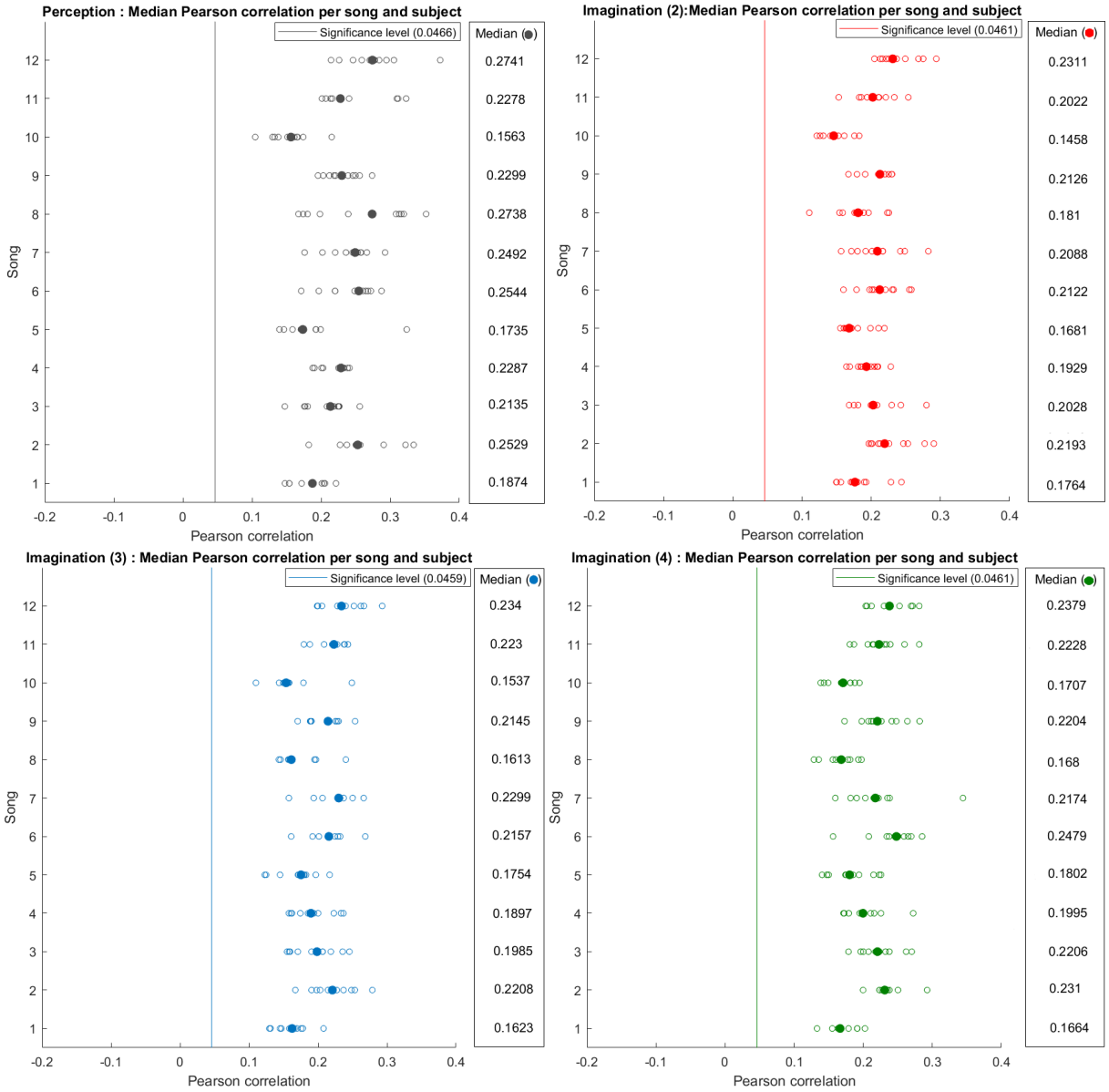

Figure 5.1: Comparing these results with implemented time shift for the OpenMIIR data set to the results of Chapter 4, a significant improvement in the decoding of the stimuli is seen.

A visual inspection of the median Pearson correlations for all subjects per song learns that the inclusion of the time shifts ν into the model adds a large improvement to our decoding approach. All median Pearson correlations per song and subject are significant. This is confirmed by a Wilcoxon signed rank test between the results with and without time shift ν for the perception experiment ($\rho_{perc,\nu} > \rho_{perc,\nu=0}$, $p < 10^{-20}$). For the imagination experiments, similar conclusions can be drawn.

As can be seen on Figure 5.1, the perception experiment still outperforms the imagination experiments in this set-up of the leave-one-song-out decoding algorithm. For all imagination conditions, performing a Wilcoxon signed rank test confirms this outperformance ($p < 10^{-3}$ for all imagination experiments). Moreover, the fourth imagination experiment might show signs of subject training, since this experiment performs significantly better than the other two ($\rho_{imag4} > \rho_{imag2}, p = 0.02, \rho_{imag4} > \rho_{imag3}$, $p = 4 \times 10^{-3}$).

Furthermore, the median optimal time shift for the perception case in Table 5.1 closely matches the results found by de Cheveigné et al. (around one second). These results thus indicate that the EEG response to a song maximally includes relevant musical information approximately one second after the stimulus onset. Also for the imagination experiments, an optimal time shift of approximately one second is found. These values are however slightly higher than for the perception case, a possible reason for this could be the varying onsets of imagination over all trials.

For the preprocessing filter range of 1-30 Hz, the results can be found in Appendix C. Parallel to the model without time shifts, the filter range of 1-10 Hz leads to a significantly better performance. This time however, this influence is seen for all four perception and imagination conditions (Wilcoxon signed rank test, $p < 10^{-10}$ for all pairwise comparisons per condition).

TABLE 5.1: Overall results for the OpenMIIR data set (bandpass filter 1-10 Hz)

	Perception	Imagination (2)	Imagination (3)	Imagination (4)
Overall median	0.21	0.20	0.19	0.21
Variance	5.5×10^{-3}	3.3×10^{-3}	3.4×10^{-3}	3.7×10^{-3}
Median optimal time shift (s)	0.99	1.02	1.03	1

5.4.2 Results for the Bach data set

For the single perception case of the Bach data set (Table 5.2), similar conclusions can be drawn. Also in this case, the inclusion of the time shift ν into the model results in the significance of all median correlations. By visual inspection, no clear difference in performance between both preprocessing filtering ranges can be seen in Figure 5.2 (Wilcoxon signed rank test, $p = 0.1$).

In comparison to the OpenMIIR data set, the influence of the time shift seems slightly lower on these stimuli. The overall median correlation has only increased slightly in comparison to the model without time shift, but the variance has decreased.

For the perception experiment in this data set, the optimal time shift is often equal to zero, resulting in a overall median result of 0 s (Table 5.2). These results thus seem to indicate that the EEG response already contains relevant musical information right after the stimulus onset. This observation is different from the OpenMIIR data set, although the same algorithm was performed. This difference in outcome can have many reasons, also practical (for example a slight skew in the onset annotations of one data set).

Table 5.2: Overall results for the Bach data set

	Bandpass 1-10 Hz	Bandpass 1-30 Hz
Overall median	0.16	0.16
Variance	5×10^{-3}	5.7×10^{-3}
Median optimal time shift (s)	0	0

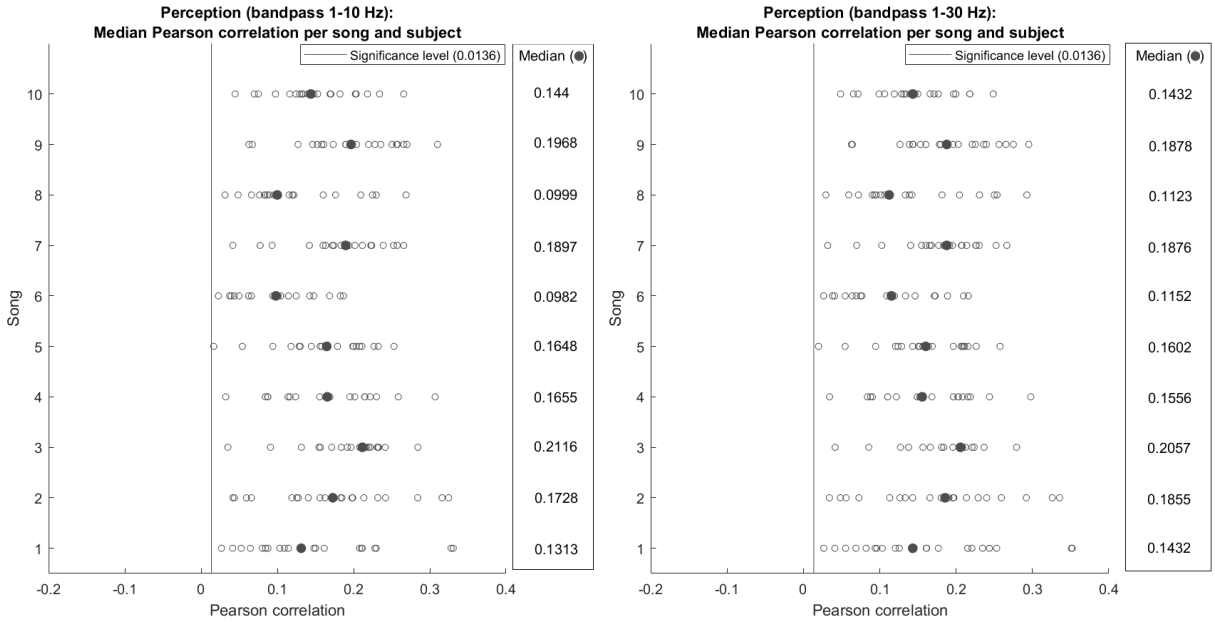

Figure 5.2: Comparing these results with implemented time shift for the Bac data set to the results of Chapter 4, an improvement in the decoding of the stimuli is seen.

5.5 Investigating effects on groups within the data

Using these results with the optimal time shift ν, the LME models of Section 4.4 were fitted for all perception and imagination experiments. In the next subsections, the observations made from the best performing models in Table 4.6 will be discussed.

5.5.1 Results for the OpenMIIR data set - Perception

When fitting this LME model to our correlations of all subjects, there is no significant effect for instrumental music (t=0.357, p=0.72), which is different from the case where

the time shift $\nu = 0$. Furthermore, the effect of instrument practice (Section 4.4) is now just non-significant on a $\alpha = 0.05$ level (estimate = -0.0234, t=-1.8395, p=0.06). For a preprocessing filter range of 1-30 Hz, no significant effects were found.

5.5.2 Results for the OpenMIIR data set - Imagination

By fitting the model of Table 4.6 on the data of the three imagination experiments for an optimal time shift ν, we obtain a significant effect for an imagination technique in condition two and four of the data set. Subjects who 'hear' the lyrics inside their head produce significantly better overall correlations (over all song categories) than those who imagine themselves singing (condition 2: estimated coefficient = 0.011, DF=593, t=2.29, p=0.022; condition 4: estimated coefficient = 0.010, DF=594, t=2.01, p=0.044). For the third condition, no significant effects were found.

Moreover, a slight negative influence of instrument practice is found on the fitted correlations (estimated coefficient =-0.012, DF=594, t=-2.28, p=.023) in the fourth imagination condition, as was the case in the perception experiment of Section 4.4. Musicians are known to have a different cortical activation pattern than non-musicians, an example of brain plasticity. For example, it was previously discovered that musicians, who regularly play on their instruments, actually show less activation of their motor cortex during a simulation of instrument practice, because of a more intense recruitment of a lower number of neurons [25]. Hypothetically, this difference might also be registered in the EEG response, possibly leading to these lower results. With a preprocessing filter range of 1-30 Hz, this influence of musicality persists for the third and fourth imagination condition.

5.5.3 Results for the Bach data set

Also for the Bach data set, the model of Table 4.6 was built. However, for both preprocessing filter ranges, no significant effects within the results were found.

5.6 Correlations in function of the time shift ν

Apart from discussing the results for an optimal time shift ν_{opt}, we can also investigate the obtained correlations per subject over all time shifts ν in the chosen range. We will therefore focus on the results for the filtering range with the best performance (bandpass 1-10 Hz). The correlations in function of time shift ν for the OpenMIIR data set can be seen in Figure 5.3 and Appendix C.3, for subject P01 (representative for all subjects). The results for the first subject of the Bach data set are presented in Figure 5.4.

In these figures, a periodical course can be observed: the reconstructed envelopes seem to have a maximal and minimal correlation with their real envelope at distinct, evenly spaced time points. The optimal time shift ν_{opt} corresponds to the highest value of these local maxima (multicolored 'dots' in Figure 5.3 and Figure 5.4). However, there exist other time shifts, at higher and lower values of ν, with similar (but slightly lower) results for the correlation. This observation raises the question

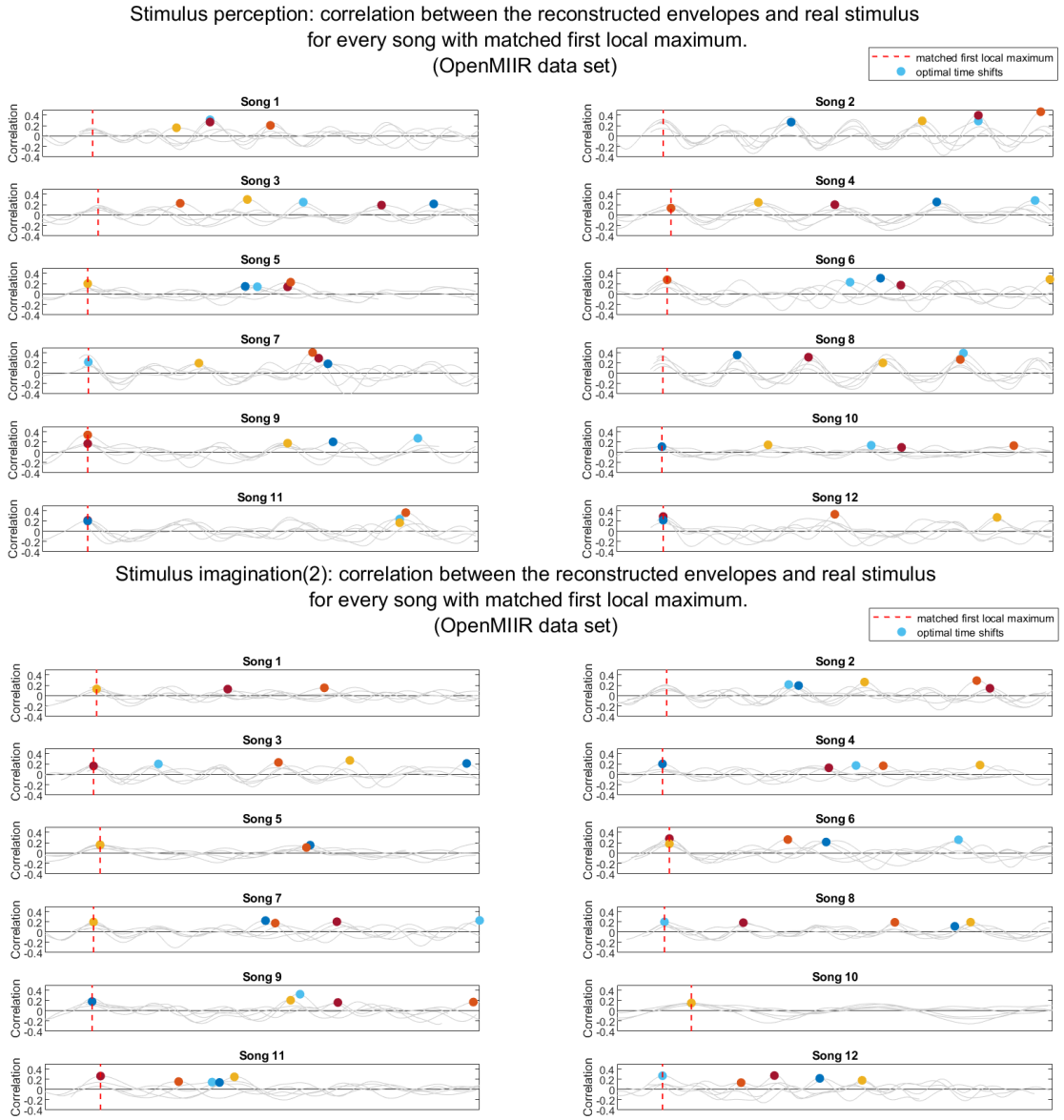

Figure 5.3: When looking at the correlation in function of time shift for each of the five reconstructed envelopes per song, seemingly periodical patterns arise in both perception and imagination experiments.

if the time shifts ν actually need to be as large as the optimal shift ν_{opt} in our stimulus reconstruction model to extract relevant musical information out of the EEG responses. Can we achieve similar results for smaller time shifts ($\nu < \nu_{opt}$) for the reconstruction of envelopes? In other words, the optimal time shift ν_{opt} is linked to the highest correlation in Figure 5.3 and Figure 5.4, but does it also represent the possible minimal latency before the EEG response?

To gather more information about the periodical behavior of the correlations in function of time shift ν, we first investigate if this periodicity has a link with beat tracking (Section 1.1.2). Afterwards, we will try to estimate the minimal time shift needed for each trial to cover a possible EEG response latency.

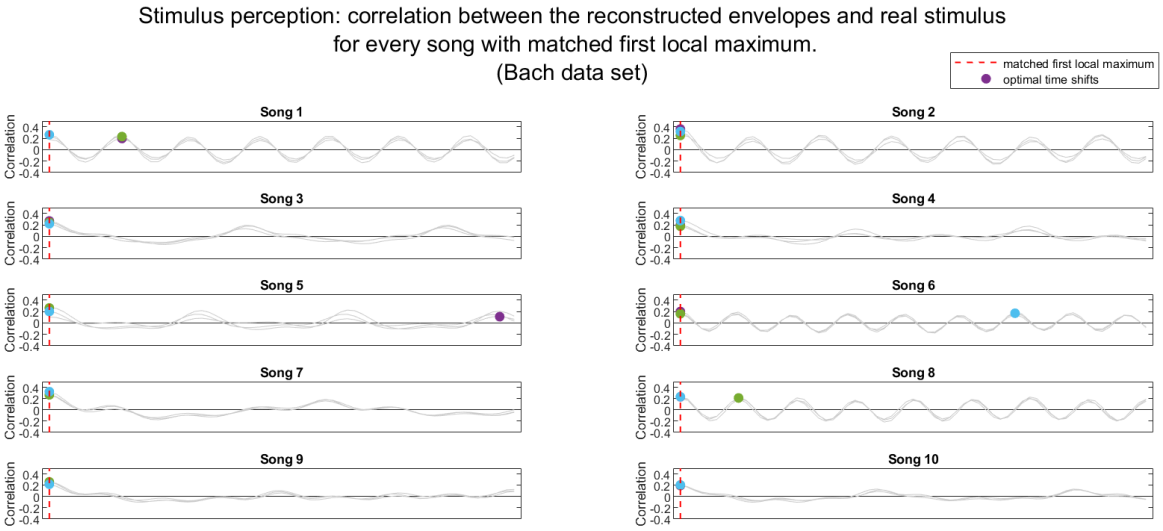

FIGURE 5.4: When looking at the correlation in function of time shift for each of the three reconstructed envelopes per song, seemingly periodical patterns arise.

5.7 Musical periodicity

As mentioned in Section 1.1.2, music is known for its periodical structure in time. This periodicity is represented via oscillatory signals in the brain, which facilitate neural beat tracking. This phenomenon is also interesting to keep in mind while dealing with our stimulus reconstruction model.

As visual study of Figure 5.3 and Figure 5.4 learns that the obtained correlations in function of the time shift ν also seem to behave in a periodical manner for each stimulus and subject. The correlation for each song seems to be maximal or minimal at regular intervals of the time shift parameter ν. One could then ask the question if such periodical behavior has a link with neural beat tracking, i.e., if the frequency connected to this periodical behavior is similar to the tempo of the used stimuli.

To investigate this hypothesis, the autocorrelation of the results for each trial in Figure 5.3 and Figure 5.4 is calculated. One example of this autocorrelation can be seen in Figure 5.5, for subject P01 in the OpenMIIR data set. Here, we get a clear view on the periodical behavior of these results. Distinct local maxima and minima at regular time shift intervals are seen in this plot of the autocorrelation.

To extract the tempo out of this autocorrelation, the median time shift interval between the subsequent peaks of the autocorrelation is calculated for every song. The median tempo over all subjects is then compared to (i.e., correlated with) the

Figure 5.5: An example of autocorrelation of the results in Figure 5.3 sheds more light on the periodical behavior of the results.

tempo of the used songs. Tempo information for all songs can be found in Table A.1. The resulting median tempo estimates over all subjects and the real tempo of the used songs for both data sets can be found in Tables 5.3 and 5.4.

For the OpenMIIR data set, Table 5.3 shows that a possible neural tracking of the tempo might in fact be present in the results of Figure 5.3. This tempo tracking can be confirmed by calculating the Pearson correlation (and its p-value via a t-test) between the estimated and real tempo of all songs. For all conditions, except for the fourth imagination experiment, the tempo estimates are thus significantly correlated with the real tempo of the songs.

Table 5.3: Tempo estimates derived from the autocorrelation of the results of the OpenMIIR data set in Figure 5.3 are correlated with the real tempo of the songs.

	Tempo estimates (BPM)				
Song	Perception	Imagination (2)	Imagination (3)	Imagination (4)	Real tempo (BPM)
1	213.3	210.5	213.3	210.5	212
2	187.3	189.7	187.3	182.9	189
3	192.1	196.9	194.5	196.9	200
4	158.4	156.7	160	160	160
5	109	128	133.6	99.9	212
6	187.3	185.1	187.3	180.7	189
7	196.9	189.9	172.8	183.3	200
8	160	160	160	160	160
9	172.6	176.6	172.6	174.5	178
10	121.9	146.3	122	114.6	166
11	109.7	114.7	124	119.2	104
12	139.7	124	149.1	143.7	140
	0.580	0.702	0.612	0.463	(correlation)
	0.048	0.011	0.034	0.129	(p-values)

For the Bach data set (Table 5.4), very high tempo estimates are found. When dividing these estimates by the real tempo, it is seen that they are (up to a few hundredths) the two-, four- or six-fold of the real tempo. This suggests that beat tracking is present for the stimuli in this data set, along with harmonics of the beat.

TABLE 5.4: Tempo estimates for the Bach data set are an X-fold of the real tempo.

Song	Tempo estimates (BPM)	Real tempo (BPM)	X-fold (-)
1	404.21	100	4.04
2	404.21	100	4.04
3	144.91	70	2.07
4	320	80	4
5	192	47	4.09
6	512	125	4.09
7	295.82	50	5.92
8	480	120	4
9	247.74	-	-
10	284.44	140	2.03
	0.642	(correlation)	
	0.062	(p-value)	

5.8 Extracting a minimal EEG response latency

From our observations made in Figure 5.3 and Figure 5.4, we will now try to extract a possible minimal latency for the EEG response. At the optimal time shift ν_{opt}, we assume that the stimulus reconstruction model is presented with relevant musical information, since the maximal correlation is found in this point. In other words, at this time instance, the minimal response latency should have been reached. The same is then true for the following local maxima at higher time shifts, which have somewhat lower correlation values. However, these similar, slightly lower correlation values can also be found at lower time shifts than the optimal one.

Therefore, the local maxima at time shifts $\nu < \nu_{opt}$ are investigated. By visual inspection and a peak analysis, where a lower limit on the peak correlations ($\rho > 0.3 \times \rho_{opt}$) was imposed, the first local maxima respecting this lower limit showed correlation values similar to the magnitude found for local maxima at and beyond the optimal time shift.

This thus suggests that the EEG responses at these lower time shifts also carry relevant musical information for our stimulus reconstruction model. These first local maxima were used to create Figure 5.3 and Figure 5.4: the results for all trials per song were overlaid while matching these local maxima on the figure.

For the Bach data set, the results in Table 5.2 are still applicable, only a slightly higher variance of 7.5×10^{-3} is seen. For the OpenMIIR data set, the overall results for these first local maxima per condition and their corresponding time shifts ν

are presented in Table 5.5. The time shifts in the perception experiment are now all lower than the previously found median value of 1 s. The three imagination conditions also generally follow this rule, only a handful of exceptions exist where this is not the case. The overall median time shift for the perception experiment now also much more resembles the low value found for the Bach data set. For the imagination experiments, this remains higher, possibly because of a mismatch between the trial and imagination onsets. In Figure 5.6 and Appendix C, the median Pearson correlations are presented for the perception and imagination experiments in the OpenMIIR data set. All median correlations are significant, asserting that the stimulus reconstruction model still works on relevant musical information.

Table 5.5: Overall results for the OpenMIIR data set (bandpass filter 1-10 Hz)

	Perception	Imagination (2)	Imagination (3)	Imagination (4)
Overall median	0.16	0.13	0.13	0.14
Variance	2.9×10^{-3}	1.2×10^{-3}	1.7×10^{-3}	1.9×10^{-3}
Median time shift(s)	0.06	0.20	0.19	0.20

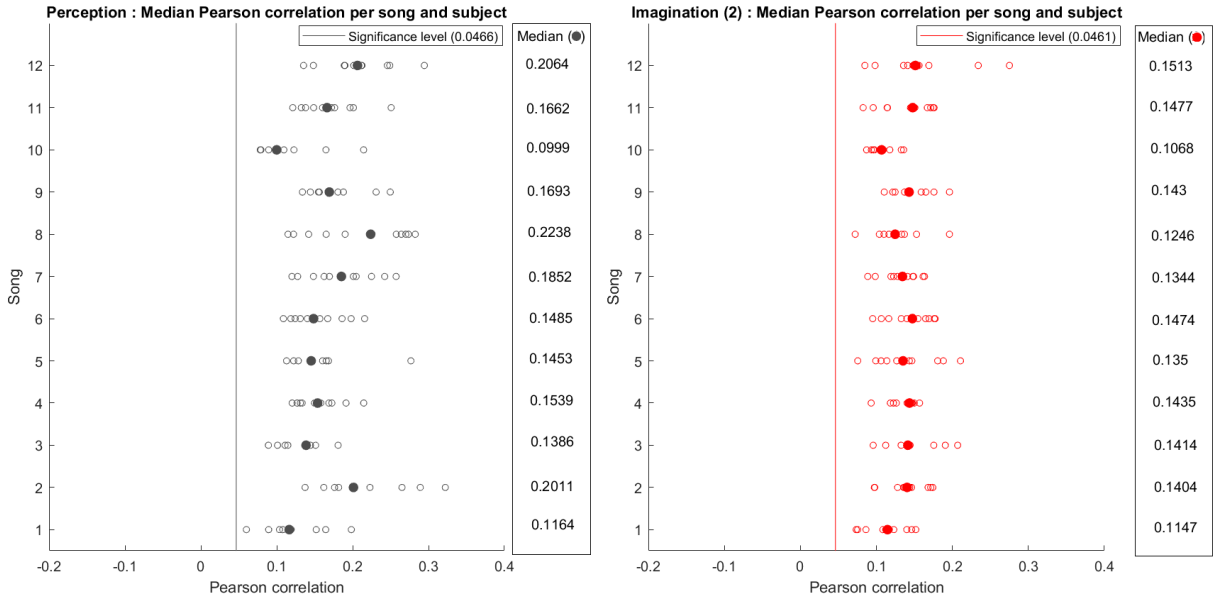

Figure 5.6: In these figures for the estimated minimal time shift, all the median correlations are significant (OpenMIIR data set).

Chapter 6

Stimulus classification

In this chapter, we investigate if it is possible to discern between musical stimuli based on their obtained reconstructed envelopes, for the reconstruction model with and without time shifts ν. Different classifiers will therefore be made. We will implement a global classifier, which will assign one song label out of all the available songs in a data set to every reconstructed envelope. Afterwards, classifiers distinguishing between two songs and song categories will be implemented. To address their performance, the accuracy of each classifier will be calculated.

6.1 Implemented classification strategies

In this chapter, each of the classifiers label their instances based on a maximal correlation strategy, which is carried out as follows. Each instance, i.e., a reconstructed envelope of a certain subject, is correlated with each available real stimulus envelope in the classification process. From the resulting Pearson correlations, the highest is selected and the corresponding label of the real stimulus envelope is assigned to the instance to classify.

The classification strategies, used in this study, are:

1. A global classifier, implemented per subject. This classifier assigns one label out of all the possible (10 for the Bach data set or 12 for the OpenMIIR data set) song labels to each reconstructed envelope of a subject.
2. A two-song classifier, to investigate the classification process of every possible combination of two songs per data set. For each of these combinations, a classifier per subject assigns one of the two available song labels to their reconstructed envelopes.
3. A song category classifier, in which we try to classify reconstructed envelopes based on song categories. This classification strategy is only possible for the OpenMIIR data set, the Bach data set contains only instrumental music. In this classifier, the same procedure as in the global classifier is carried out, but only with three possible labels instead of 12.

For each of the classifiers, the accuracy is computed to investigate the classification performance. To conclude if this accuracy is statistically significant, McNemar's

test [34, 5] is performed, in which we compare our obtained accuracy to the result for a classifier which always predicts the majority class in the data. Since our two data sets are perfectly balanced for every song (five trials per song and subject in the OpenMIIR data set, three in the Bach data set), this majority class is taken to be any of the available songs during classification.

6.2 Results and discussion

6.2.1 Results for the global classifier

For the global classifier, the obtained accuracy per subject is plotted in Figure 6.1 for all perception/imagination conditions (bandpass 1-10 Hz), without and with the optimal time shift ν_{opt} of the stimulus reconstruction model in Chapter 5. In this figure, the significant classification results are shaded with gray clusters per condition.

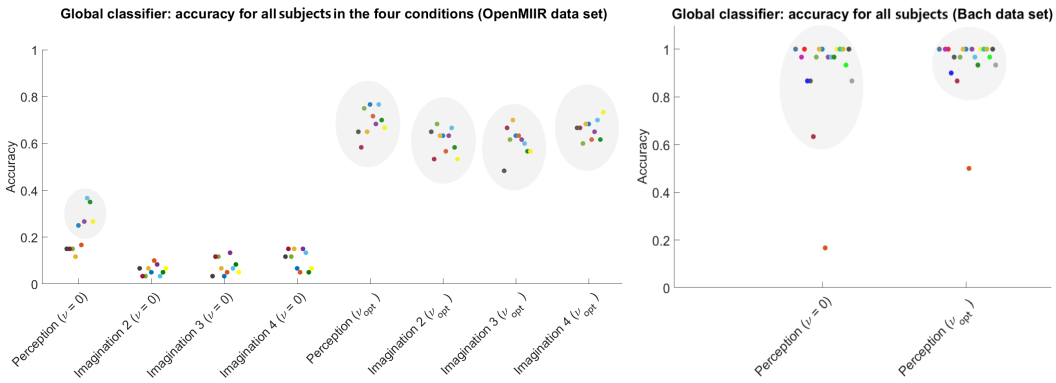

Figure 6.1: When looking at the results of both data sets, the improvement of including a time shift $\nu = \nu_{opt}$ into the model is clearly seen.

For the OpenMIIR data set, we see a significant improvement in the accuracy by including the optimal time shift ν_{opt} into our stimulus reconstruction model. For the Bach data set, including a time shift into the model increases the accuracy per subject, but does not change the amount of statistically significant results. The classification results for perception in this data set are also higher than for the OpenMIIR data set. We also see a high number of perfect classifications (accuracy equal to one) for this data set. In Chapter 4 and 5, this data set showed higher correlations with respect to their real envelopes than the OpenMIIR data set. However, to obtain a good classification of our reconstructed envelopes, their correlation with other stimulus envelopes should also be lower, leading to a lower amount of misclassifications. In Section 6.3, we inspect possible influences for the misclassification in this global classifier.

For a preprocessing filter range of 1-30 Hz, the accuracy per subject shows similar results (Figure D.1) for both data sets. One extra subject in the perception experiment (without time shifts) of the OpenMIIR data set now reaches a significant accuracy. However, generally, a slightly lower accuracy over all subjects is found for this filter range.

6.2.2 Results for the two-song classifier

For this classification strategy, we make subdivisions in the data for every possible combination of two songs. The reconstructed envelopes per subdivision are then labeled by a two-song classifier. The mean accuracy over all subdivisions is presented in Figure 6.2, per condition (one 'dot' represents one subject). Comparing this to Figure 6.1, the effect of studying two songs in a classifier instead of 12 is clearly seen (higher accuracy). To test the combined performance of these classifiers, their predictions for each reconstructed envelope were fused using majority voting, which showed a significant performance of the ensemble classifier in all cases, also for the 1-30 Hz preprocessing filter range (Figure D.2).

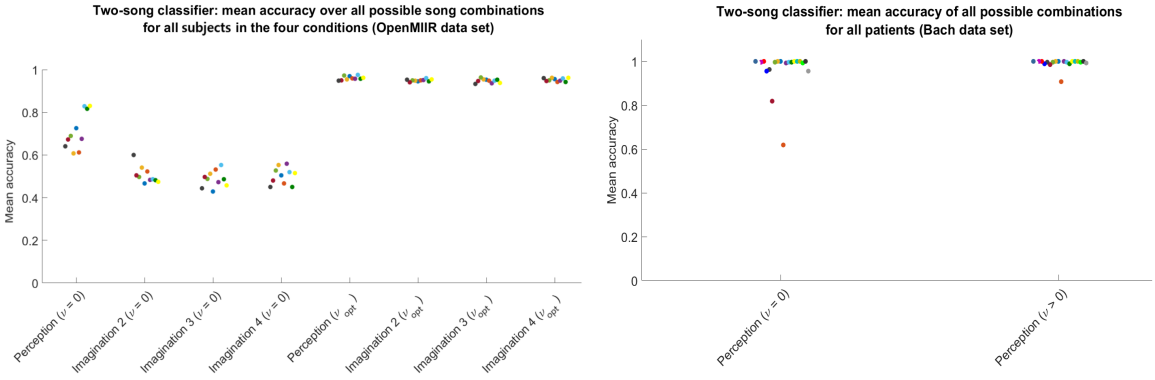

FIGURE 6.2: Mean accuracy per subject for the two-song classifier.

6.2.3 Results for the song category classifier

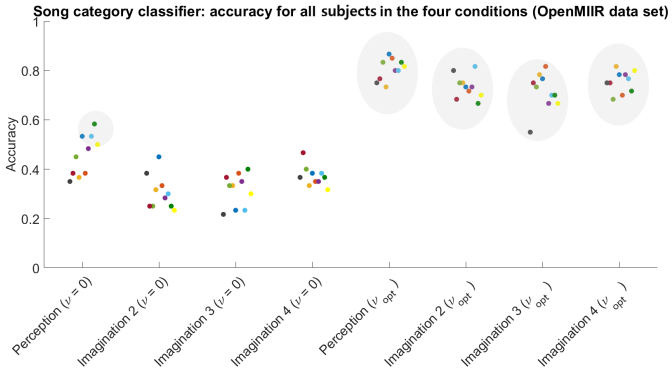

FIGURE 6.3: The improvement of including a time shift $\nu = \nu_{opt}$ into the stimulus reconstruction model can also be seen for this classifier.

As mentioned before, this classifier can only be implemented for the OpenMIIR data set. In Figure 6.3 (bandpass filter 1-10 Hz), we again see an improvement in accuracy when time shifts ν_{opt} are included in the stimulus reconstruction model. As with the global classifier, the imagination conditions, which produced non-significant results for a time shift $\nu = 0$, are now significant for all subjects. Also, for this classifier, no significant influence between the two preprocessing filtering ranges was found.

6.3 Misclassification by the global classifier

We investigate the misclassifications by the global classifier by computing the confusion matrix for every studied condition over all subjects. When visually comparing these misclassifications to the information about our stimuli (Table A.1), possible interesting effects to study are the following. In the OpenMIIR data set, the same songs with and without lyrics are present. Are these song pairs more mistaken for each other by the global classifier than two other random songs? Moreover, we will investigate the effect of tempo on the misclassification count. Are stimuli with a similar tempo misclassified more than stimuli with diverging tempo's?

To observe the effect of the lyrics/no lyrics song pairs in the OpenMIIR data set per condition, the total misclassification count (i.e., the sum of misclassifications) over these 4 stimulus pairs is compared to the total misclassification count of 4 other randomly selected pairs of songs. We repeat the selection of random song pairs 10000 times, to get an estimate of the distribution of misclassification counts for a specific condition. When the total misclassification count of the lyrics/no lyrics song pairs exceeds the 97.5 percentile of this distribution, the effect of these song pairs is seen as significant. This leads to a significant total misclassification count for the perception experiment without time shifts (for filter range 1-10 Hz) and with time shifts (filter range 1-30 Hz) in the OpenMIIR data set.

To investigate a possible influence of tempo difference, the total misclassification count in function of tempo difference between song pairs is presented in Figure 6.4 for a filtering range from 1-10 Hz. For the OpenMIIR data set, no significant overall trend in function of tempo difference is seen. The peak at 0 BPM of tempo difference stands for the lyrics/no lyrics song pairs described above. For the Bach data set, a clear, high peak at zero tempo difference is seen for both perception experiments (with and without time shifts). Similar results are found for 1-30Hz filtering range. The two different stimuli with the same tempo (Table A.1) in this data set are thus mistaken for each other relatively often, while at a higher tempo difference, less misclassification of stimuli takes place.

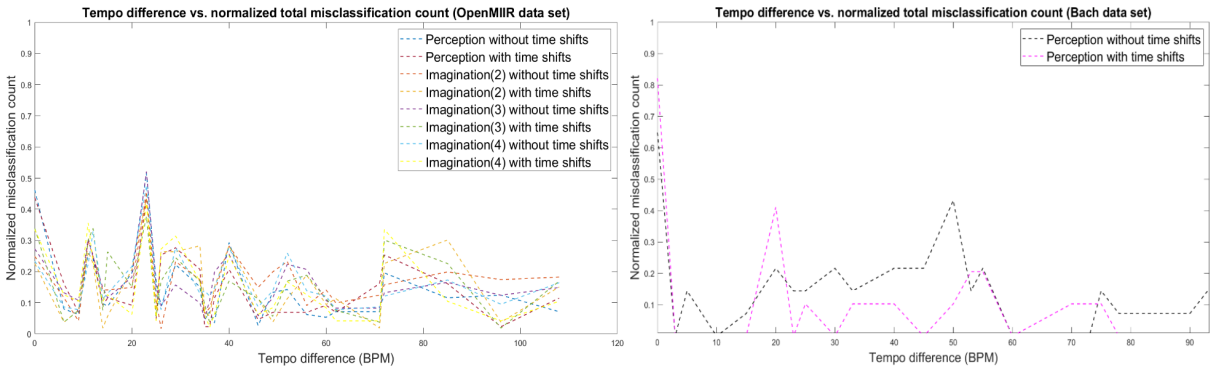

Figure 6.4: For the effect of tempo differences in general, there is no significant trend in the misclassification count of the OpenMIIR data set. For the Bach data set, a clear peak at zero tempo difference is seen.

Chapter 7

Conclusions and future work

In this study, we decoded stimulus envelopes out of their corresponding EEG responses. For this, we used a linear decoding algorithm, which was extended to incorporate a possible EEG response latency. With the information obtained in the previous chapters, our research questions can now be answered, based on the observed results. Our general research question was the following:

'Is it possible, using a linear decoding model, an EEG response and a stimulus envelope, to find a good (i.e., ideally optimal) stimulus reconstruction for perception and imagination experiments?'

The answer to this question is 'yes': it is possible to decode an EEG response into a reconstructed envelope, which achieves a significant correlation with the real envelope. However, the results obtained for each condition highly depend on the designed model. For our stimulus reconstruction model of Chapter 4, varying results were found over all studied experiments. For perception, in which subjects listen to a fixed stimulus, significant results were found for this model. However, for the imagination experiments, possible distortions due to tempo inconsistencies and a mismatch between the trial and imagination onsets make it difficult for this model to achieve good results.

For the extended stimulus reconstruction model with the inclusion of time shifts, which made it possible to explicitly include these EEG response latencies, the results vastly improved for all experiments. Even after extracting a minimal time shift from our periodical course of correlations, the median Pearson correlations of all experiments remained significant.

Next, we have shown that it is possible to discern between songs, based on the correlation of their reconstructed envelopes with the real song envelopes. For these classifiers however, the same conclusions can be drawn. For the model without time shifts, only the perception experiment achieves a significant accuracy for a number of subjects in the global and song category classifiers. When including time shifts into our model, these classifiers have a vastly improved performance, also for the imagination experiments in the OpenMIIR data set.

Moreover, a few effects within the obtained correlations were found, for both the perception and imagination experiments. For the reconstruction model during perception, without time shifts, significantly higher correlations were found for the category of instrumental music than for the other categories. Multiple experiments, with and without time shifts, also showed lower correlations for musicians who regularly play an instrument in comparison to subjects with other levels of instrument practice. Additionally, models for the imagination experiments with inclusion of time shifts showed a significant effect for the used imagination technique: subjects who 'hear' the lyrics inside their heads produce significantly better results than those who imagine themselves singing.

Lastly, the influence of the preprocessing filter range on the results of our models was studied. In general, we found that the 1-10 Hz filtering range significantly outperformed the 1-30 Hz range for the OpenMIIR data set. For the Bach data set, generally no difference between the results for both filtering ranges is seen.

With this study in mind, one can think about possible future music processing topics and problems which can be tackled. The workings of encoding and hybrid 'encoding-decoding' techniques can be tested on both perception and imagination experiments. Our stimulus reconstruction model could also still be extended, by for example including a response latency into the training of the decoder. Different types of classifiers could be added to specialize on the classification of certain songs. A further study on the found effects within our results can be conducted for both perception and imagination experiments.

Moreover, as seen in Section 1.4, its was found that musical stimuli in perception experiments could be discerned from each other in an auditory attention detection set-up. Would it be possible to explicitly make such a functioning model for imagination experiments as well? Every study about the perception and imagination of stimuli teaches us more about the workings of the brain. Would it one day be possible to reconstruct an imagined song at real-time, creating a 'Shazam for the brain'?

Appendices

Appendix A

Data

A.1 Stimuli used in both datasets

Table A.1: Stimuli used in both datasets. For each, the duration, tempo and song category is given.

	OpenMIIR dataset (Stober et al.)[36]			
	Musical stimulus	Duration (s)	Tempo (BPM)	Category
1	Chim chim cheree	13.3	212	Lyrics
2	Take me out to the ballgame	7.7	189	
3	Jingle Bells	9.7	200	
4	Mary had a little lamb	11.6	160	
5	Chim chim cheree	13.5	212	No lyrics
6	Take me out to the ballgame	7.7	189	
7	Jingle Bells	9.0	200	
8	Mary had a little lamb	12.2	160	
9	Emperor Waltz	8.3	178	Instrumental
10	Hedwig's Theme (Harry Potter)	16.0	166	
11	Imperial March (Starwars Theme)	9.2	104	
12	Eine kleine Nachtmusik	6.9	140	
	Bach dataset (Di Liberto et al.)[12][17]			
	Musical stimulus	Duration (s)	Tempo (BPM)	Category
1	Partita in A minor for Solo flute Allemande	160.9	100	Instrumental
2	Partita in A minor for Solo flute Courante	156.0	100	
3	Partita in A minor for Solo flute Sarabande	122.2	70	
4	Partita in A minor for Solo flute Bouree Anglaise	137.8	80	
5	Partita No. 2 in D minor Allemande	167.7	47	
6	Sonata No. 1 in G minor Presto	201.3	125	
7	Partita No. 1 in B minor Allemanda	175.2	50	
8	Partita No. 2 in D minor Gigue	184.2	120	
9	Partita No. 3 in E major Loure	136.2	-	
10	Partita No. 3 in E major Gavotte en Rondeau	180.1	140	

A.2 Comparison between the datasets of this study.

Table A.2: Comparison between the datasets of this study.

	OpenMIIR (Stober et al.)[36]	**Bach (Di Liberto et al.)[12]**
Subjects	10	20
Stimuli	12	10
Trials per unique setting	5	3
Conditions		
Perception	cued	not cued
Imagination	cued	-
	not cued (x2)	-
Song categories		
Lyrics	1-4	-
No lyrics	5-8	-
Instrumental	9-12	1-10
Musicality(1):		
Musical training		
Non-musicians	2 (P02, P10)	10 (P01-P10)
Musicians	8 (others)	10 (P11-P20)
Musicality(2):		
Instrument practice		
Current	4 (P03-P05, P09)	10 (P01-P10)
Past	4 (P01, P06-P08)	-
Never	2 (P02, P10)	10 (P11-P20)
Imagination technique (lyrics)		
Imagine themselves singing	5 (P01, P02, P05-P07)	-
Hear the lyrics inside their heads	5 (P03, P04, P08-P10)	-
Visualizations (no lyrics)		
Yes	4 (P04, P05, P07, P09)	-
No	6 (P01-P03, P06, P08, P10)	-
Total amount of trials	**2400**	**600**

Appendix B

Linear regression model: Results

B.1 Results OpenMIIR dataset: Bandpass 1-10 Hz

B.1.1 Stimulus perception

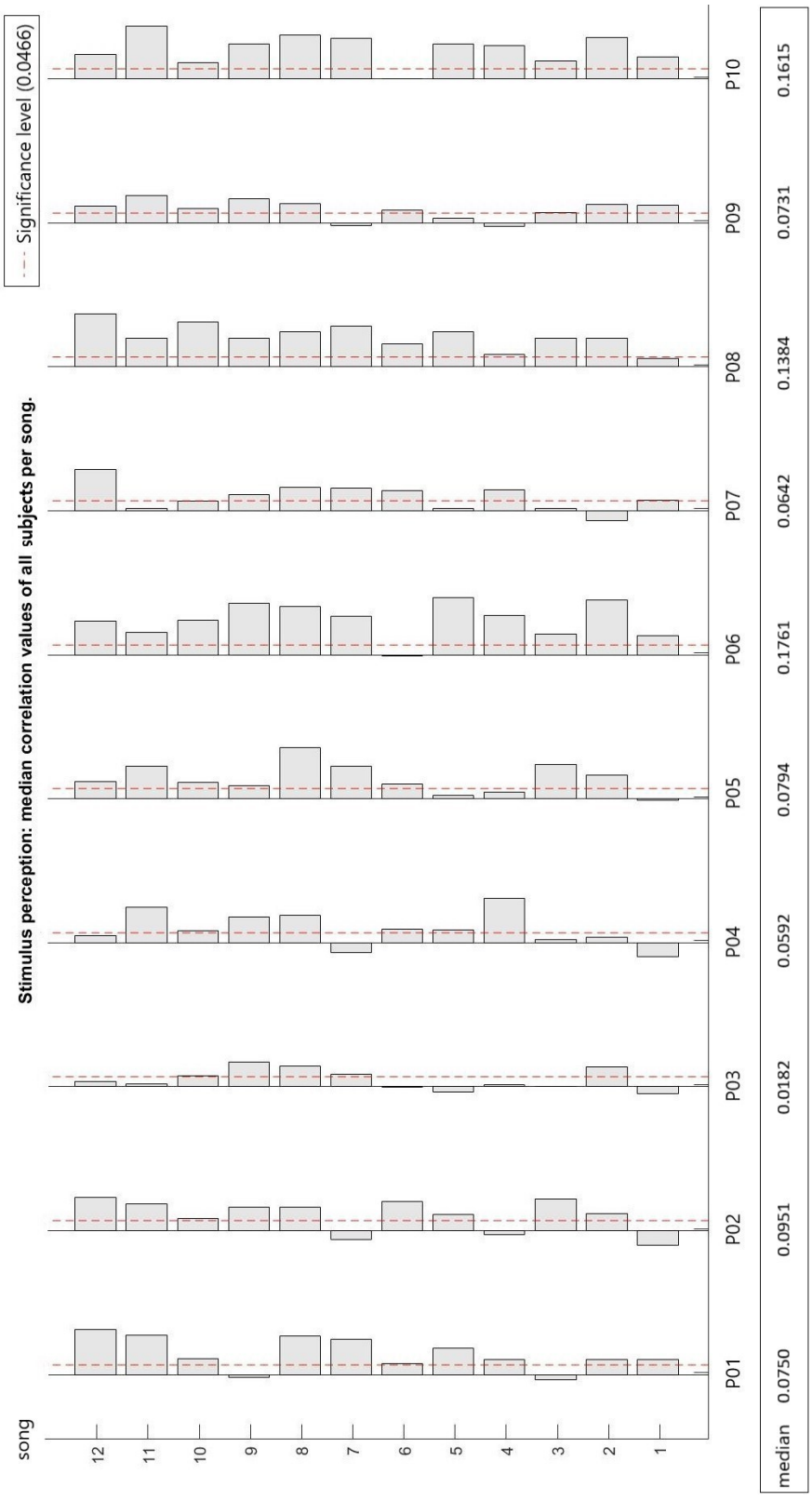

Figure B.1: A subdivision per subject shows the worst and best performing participants.

B.1.2 Stimulus imagination (condition 2)

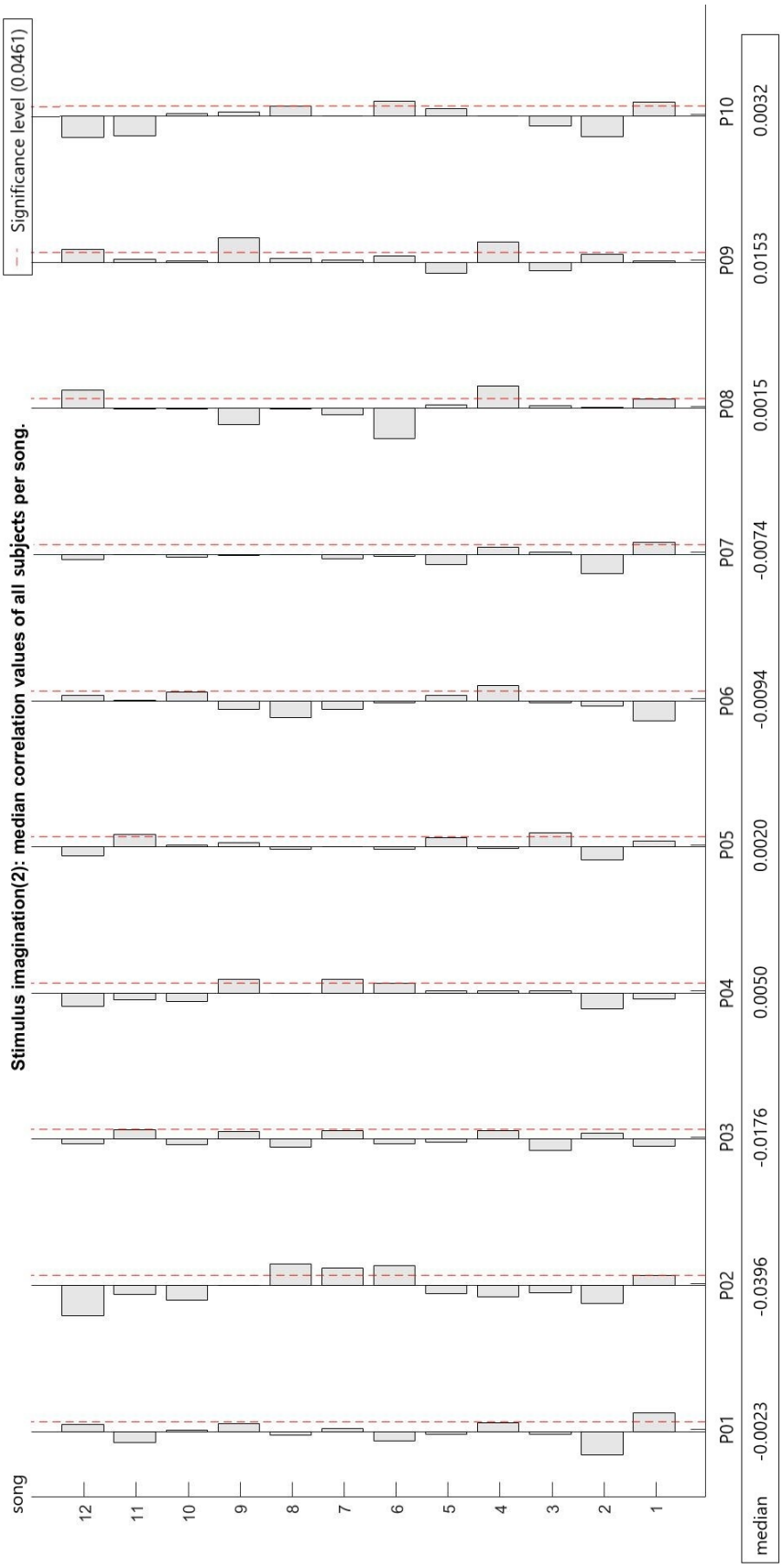

FIGURE B.2: A subdivision per subject shows an overall low performance.

B.1.3 Stimulus imagination (condition 3)

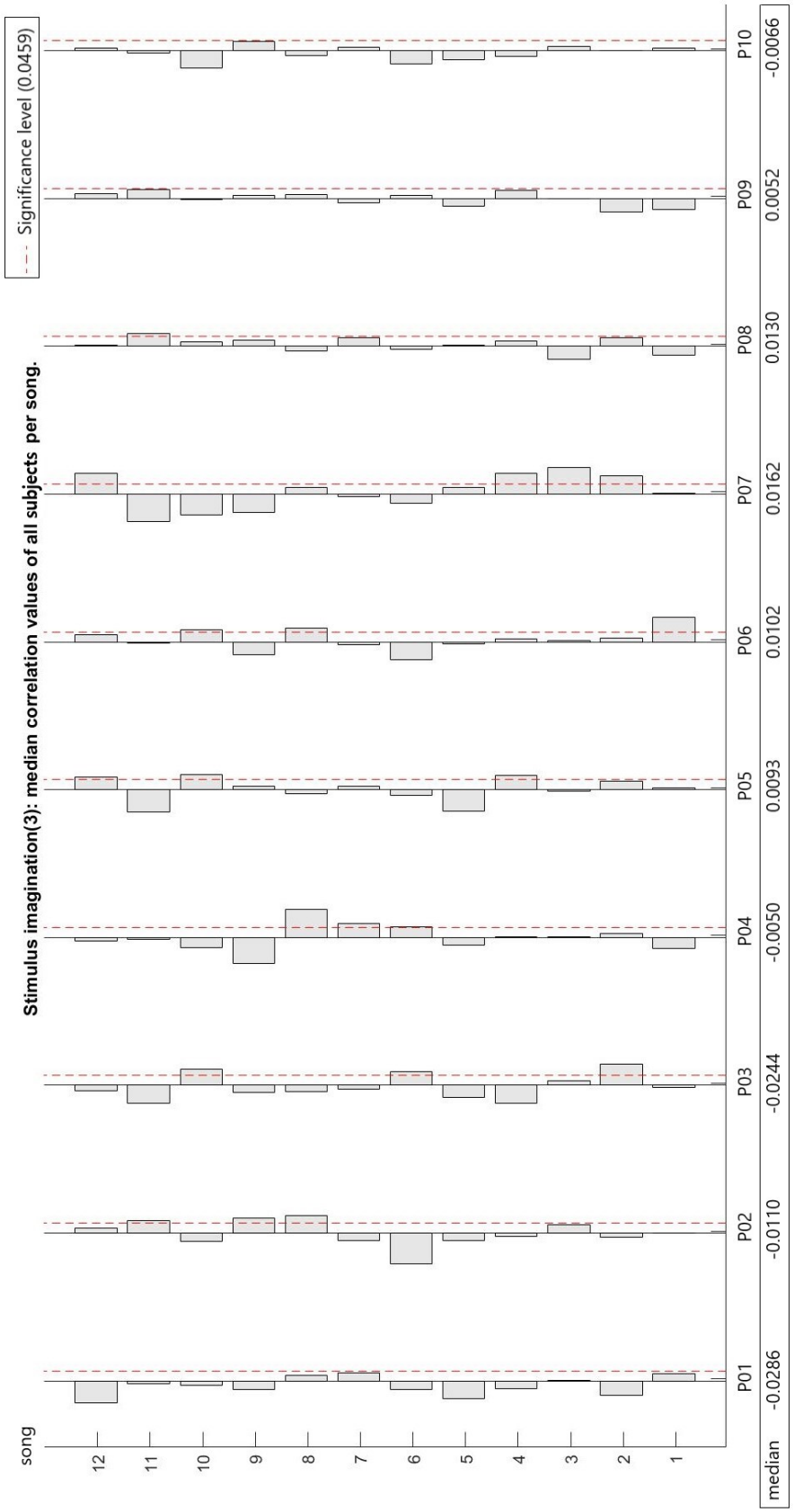

Figure B.3: A subdivision per subject shows an overall low performance.

B.1.4 Stimulus imagination (condition 4)

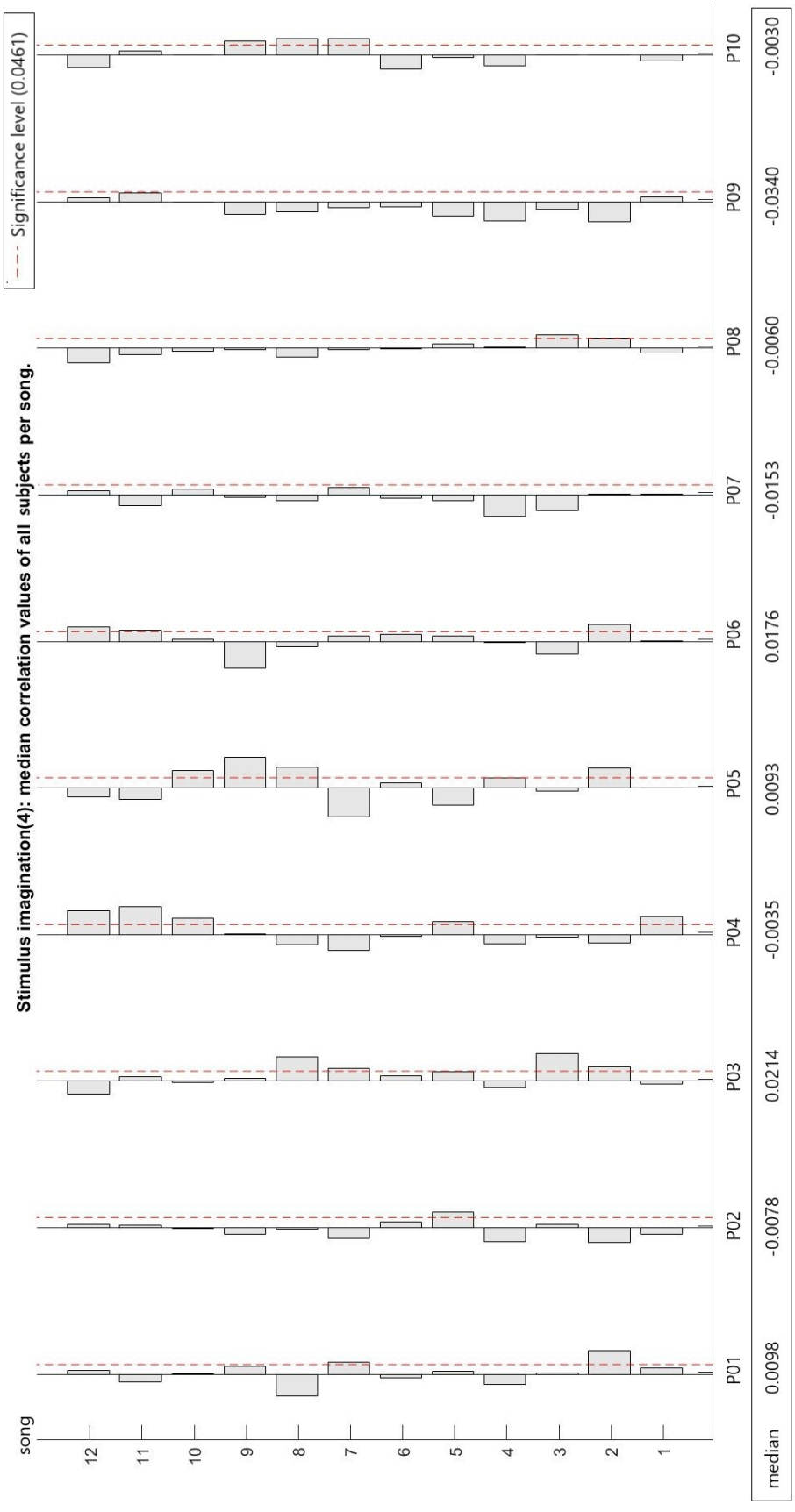

FIGURE B.4: A subdivision per subject shows an overall low performance.

B.2 Results OpenMIIR dataset: Bandpass 1-30 Hz

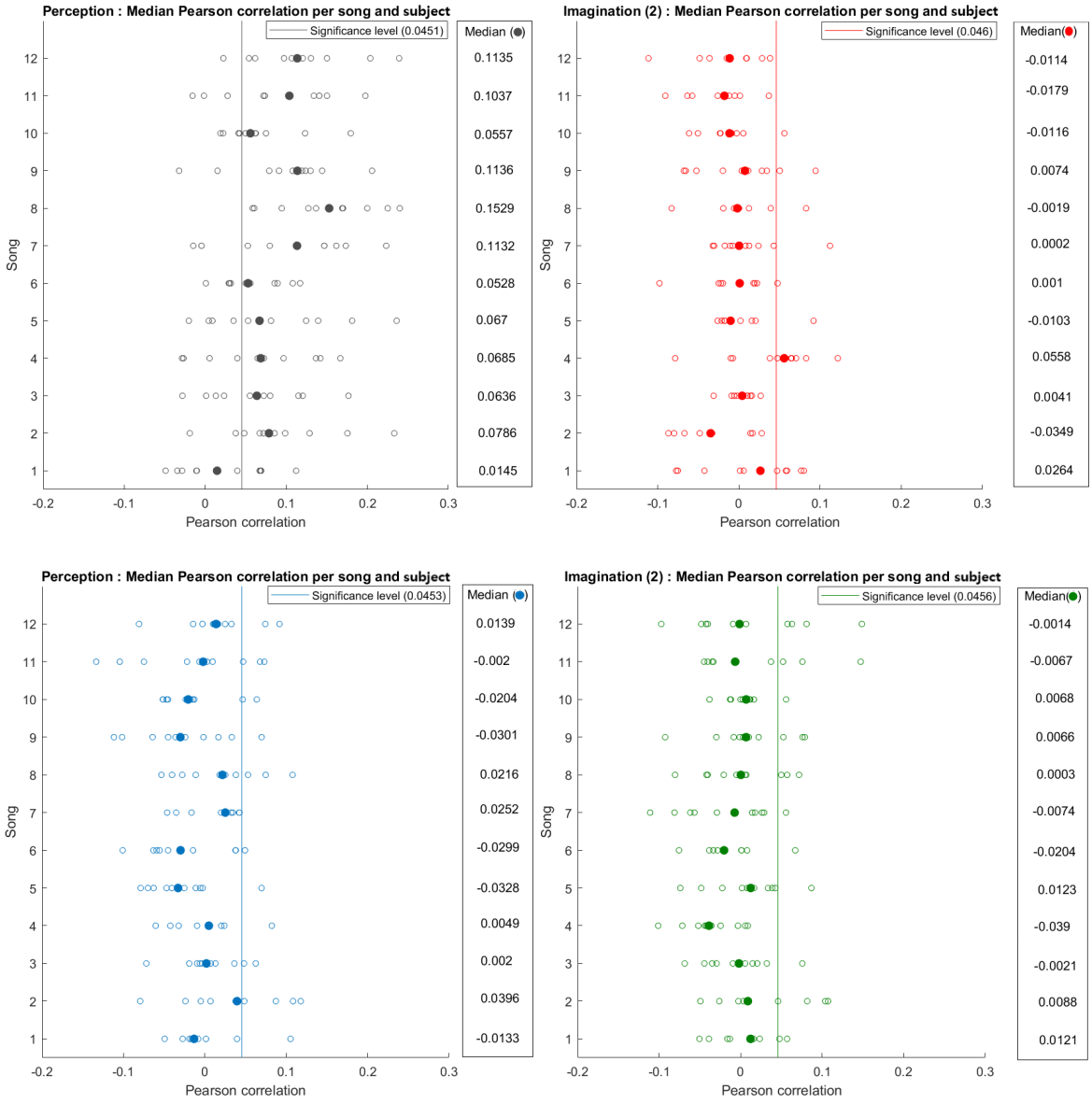

Figure B.5: The figures for this preprocessing filter range show overall slightly lower median correlations, the previous range had a significantly better performance.

B.2.1 Stimulus perception

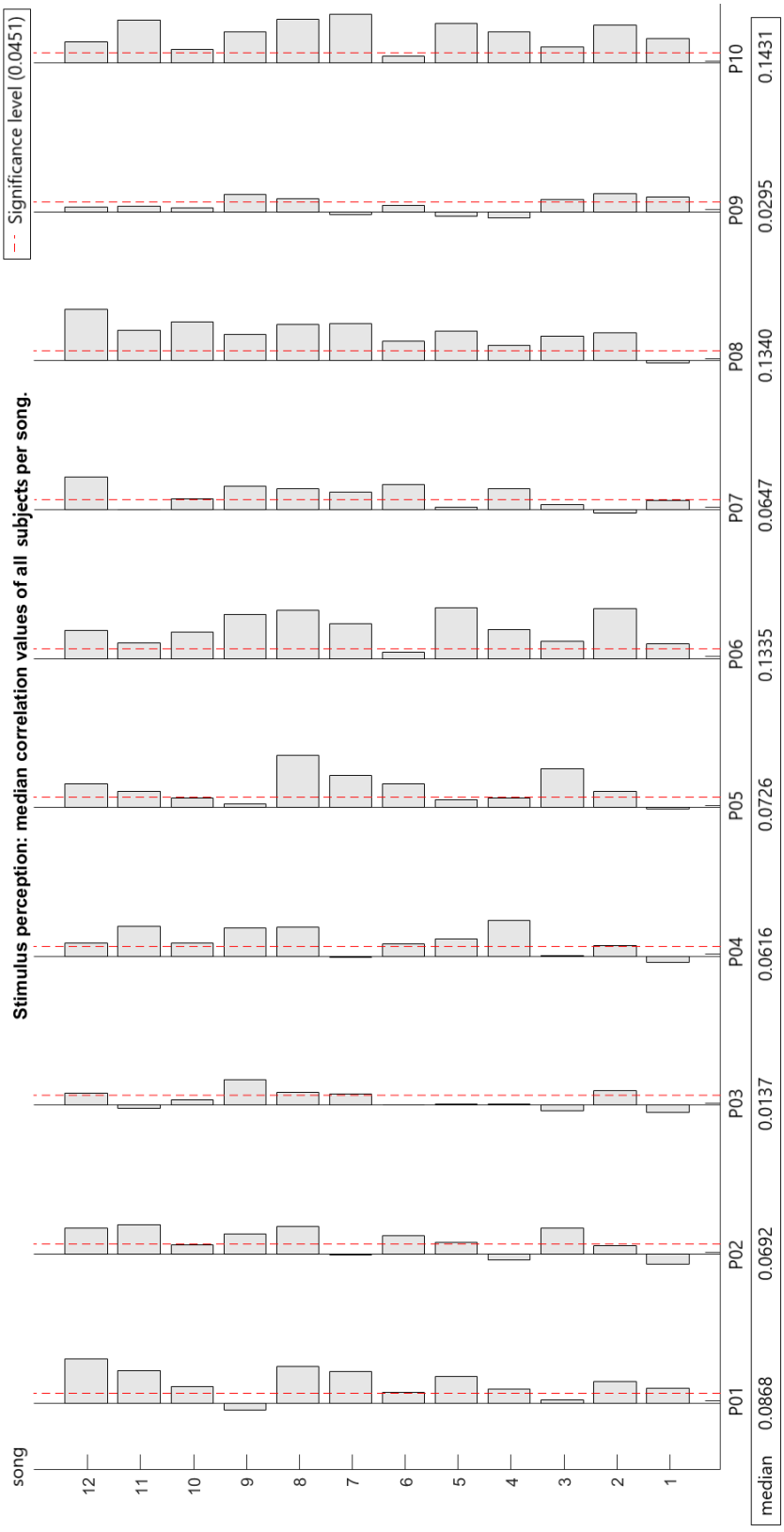

FIGURE B.6: A subdivision per subject shows the best and worst performing participants.

B.2.2 Stimulus imagination (condition 2)

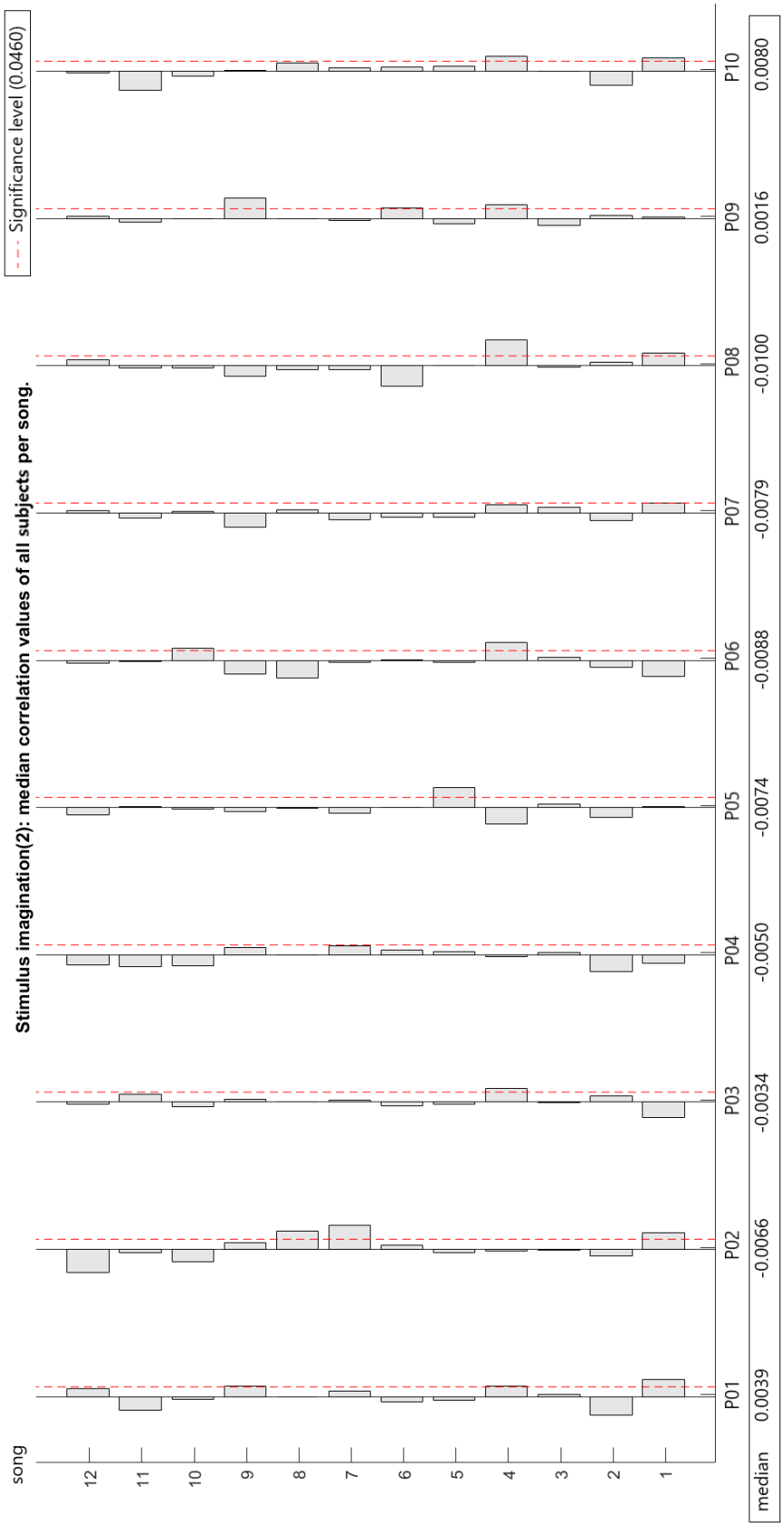

Figure B.7: A subdivision per subject shows an overall low performance.

B.2.3 Stimulus imagination (condition 3)

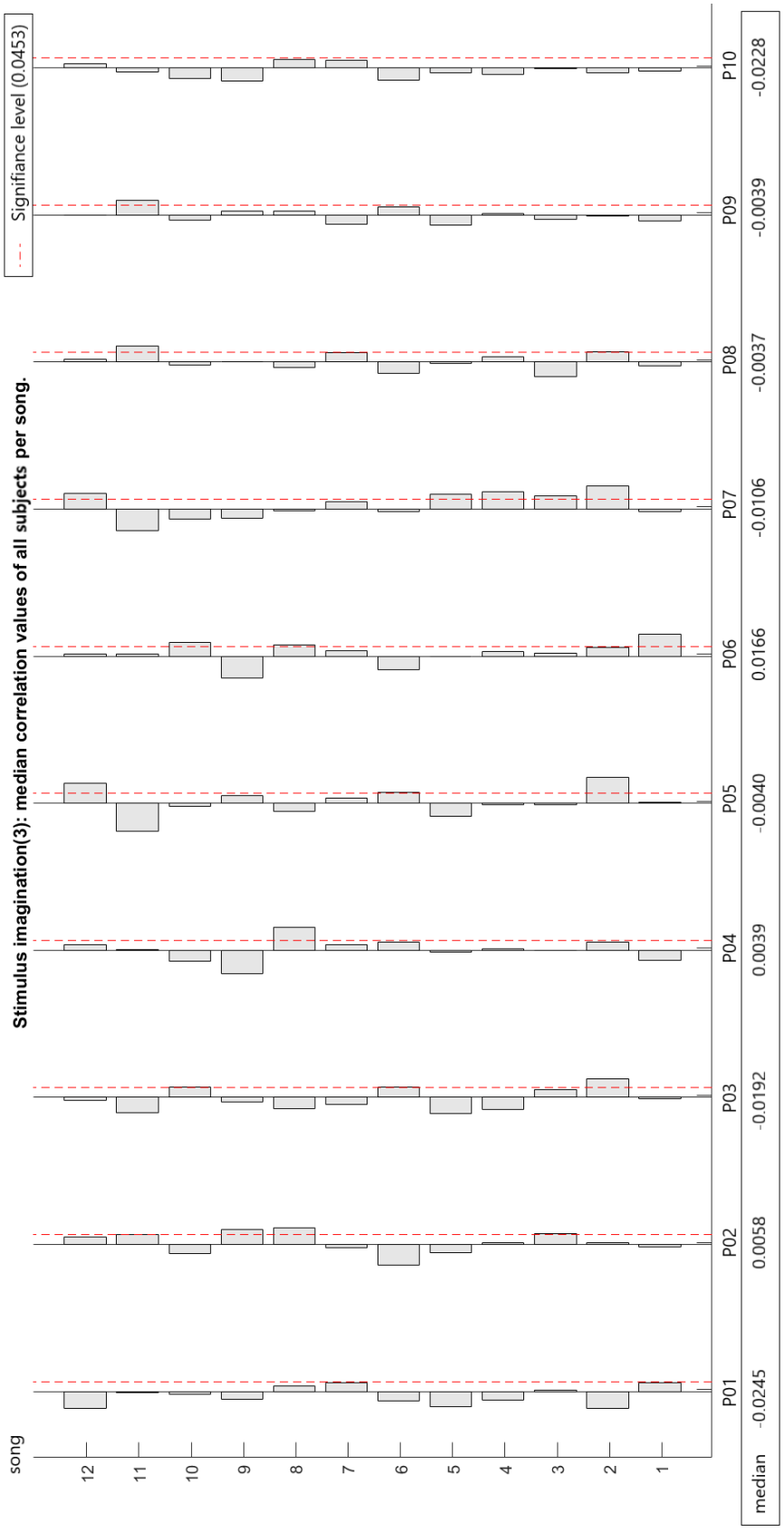

FIGURE B.8: A subdivision per subject shows an overall low performance.

B.2.4 Stimulus imagination (condition 4)

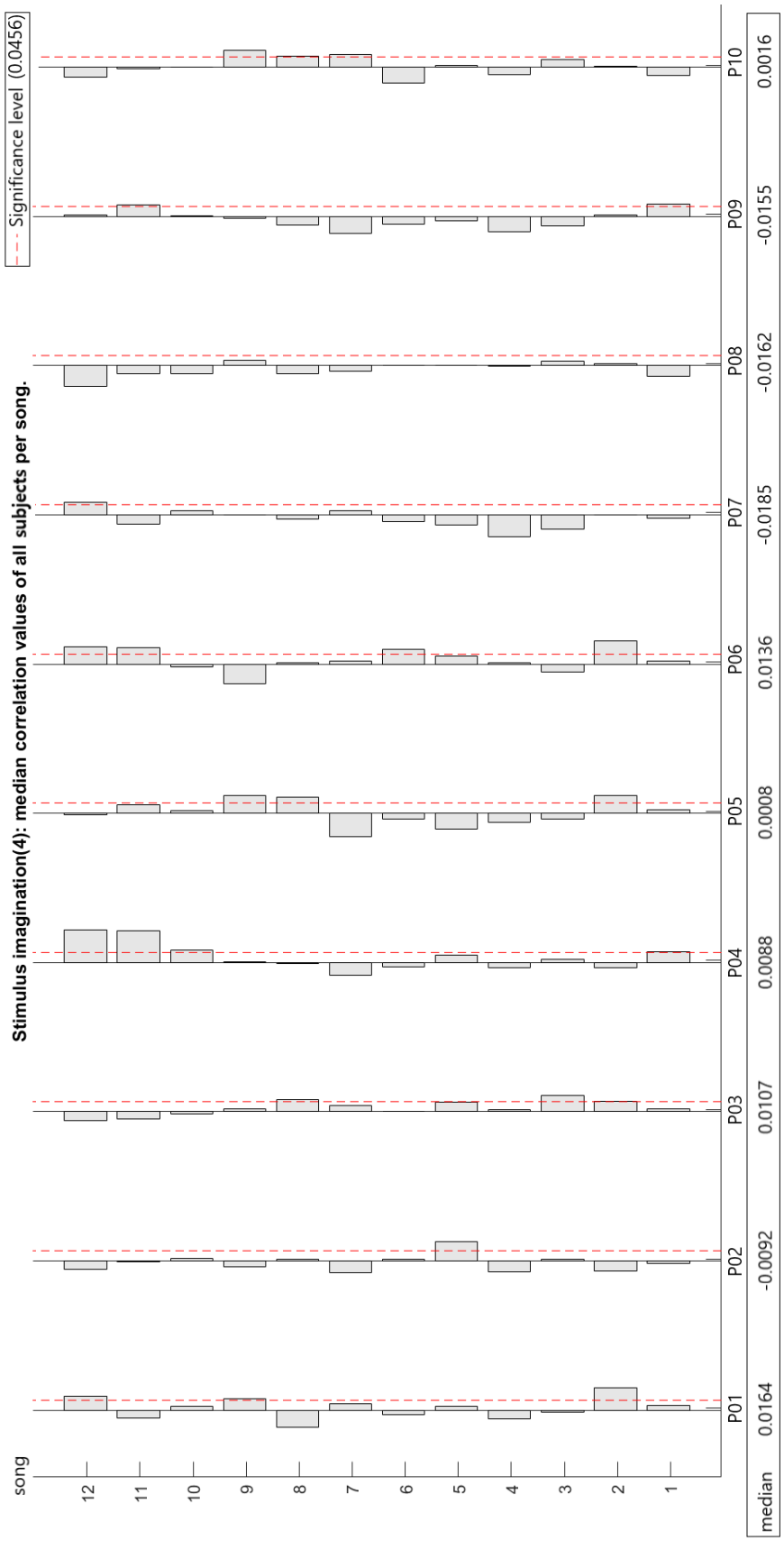

Figure B.9: A subdivision per subject shows an overall low performance.

B.3 Results Bach dataset: Bandpass 1-10

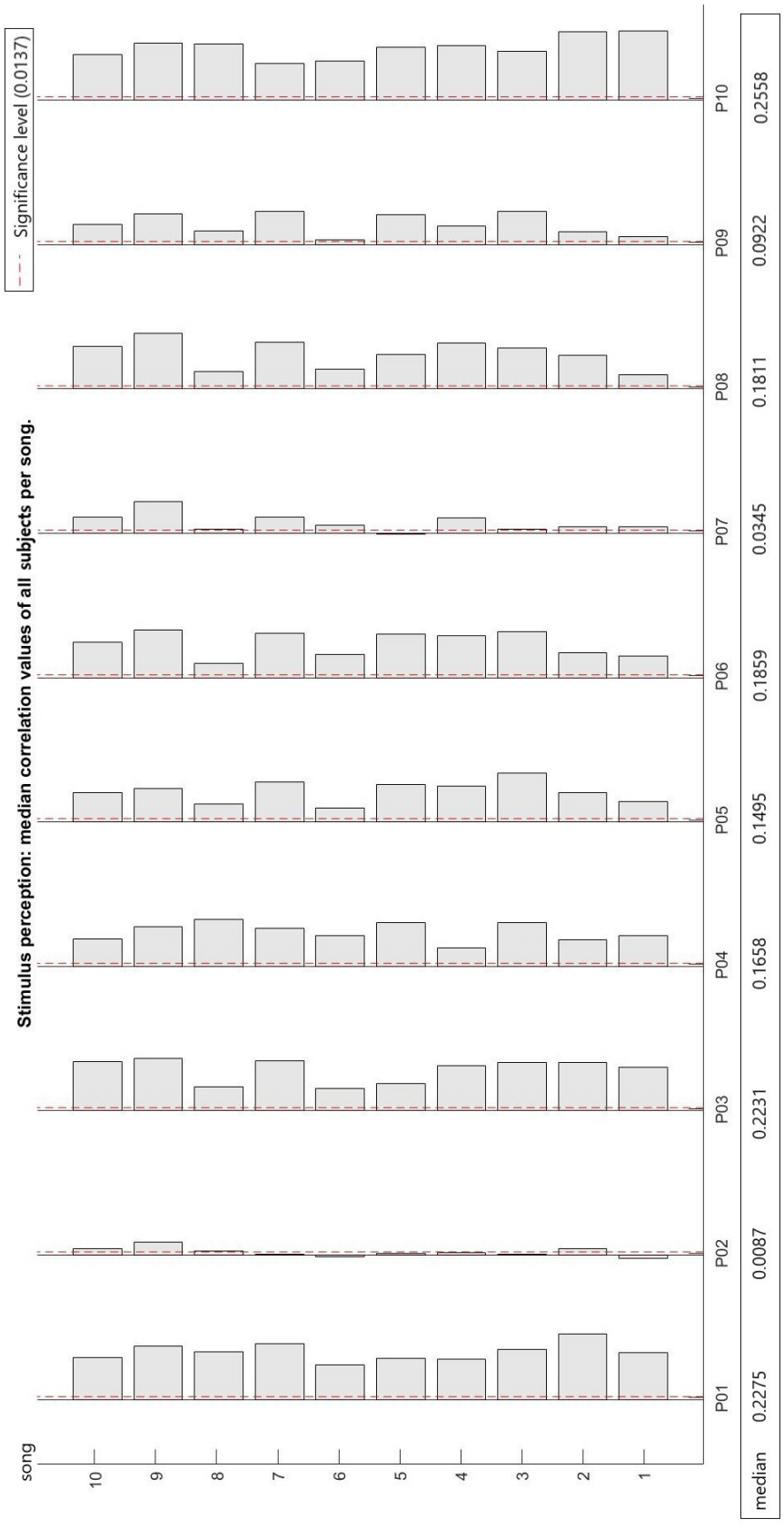

FIGURE B.10: A subdivision per subject shows the best and worst performing participants.

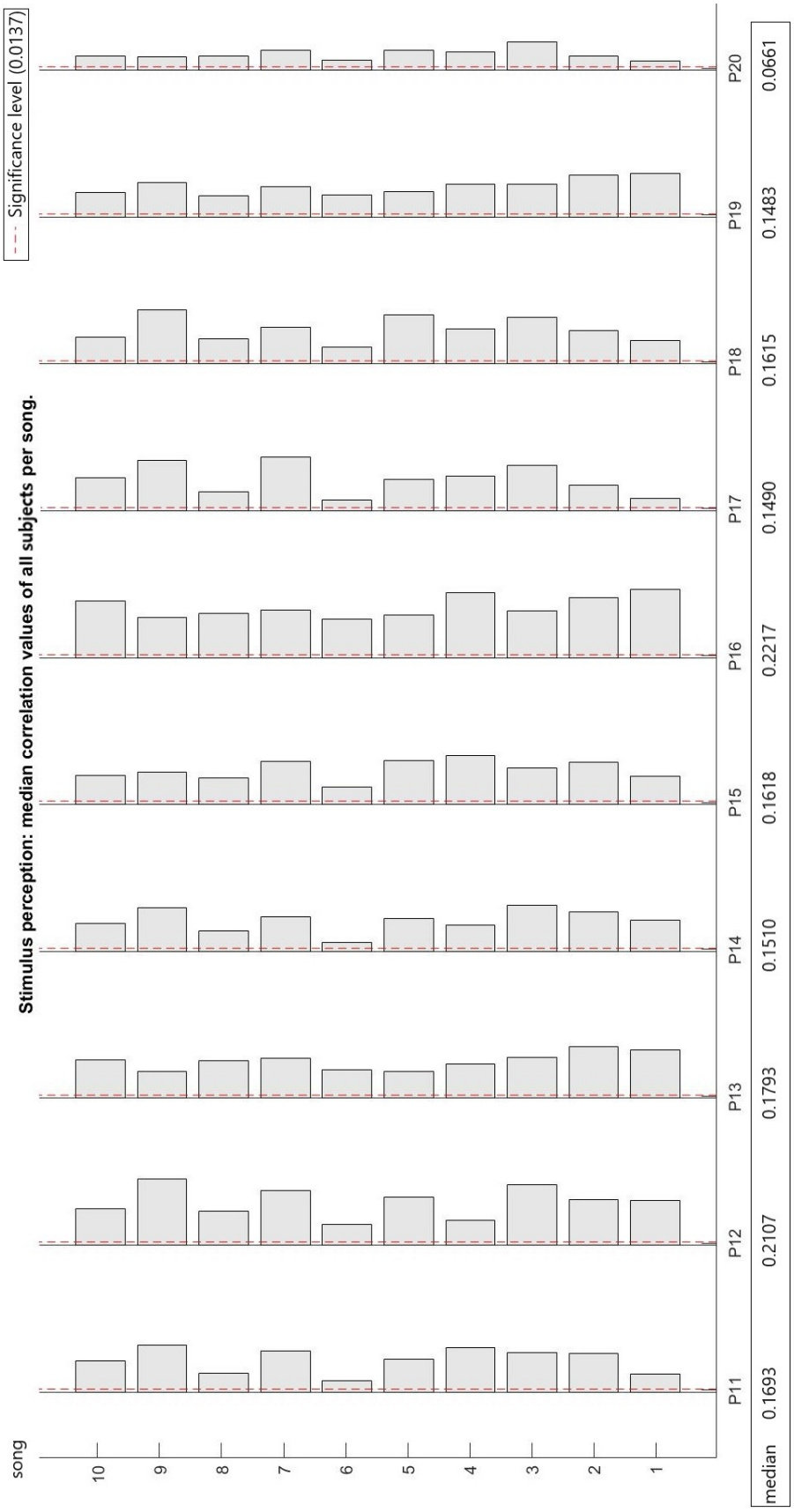

Figure B.11: A subdivision per subject shows the best and worst performing participants.

B.4 Results Bach dataset: Bandpass 1-30

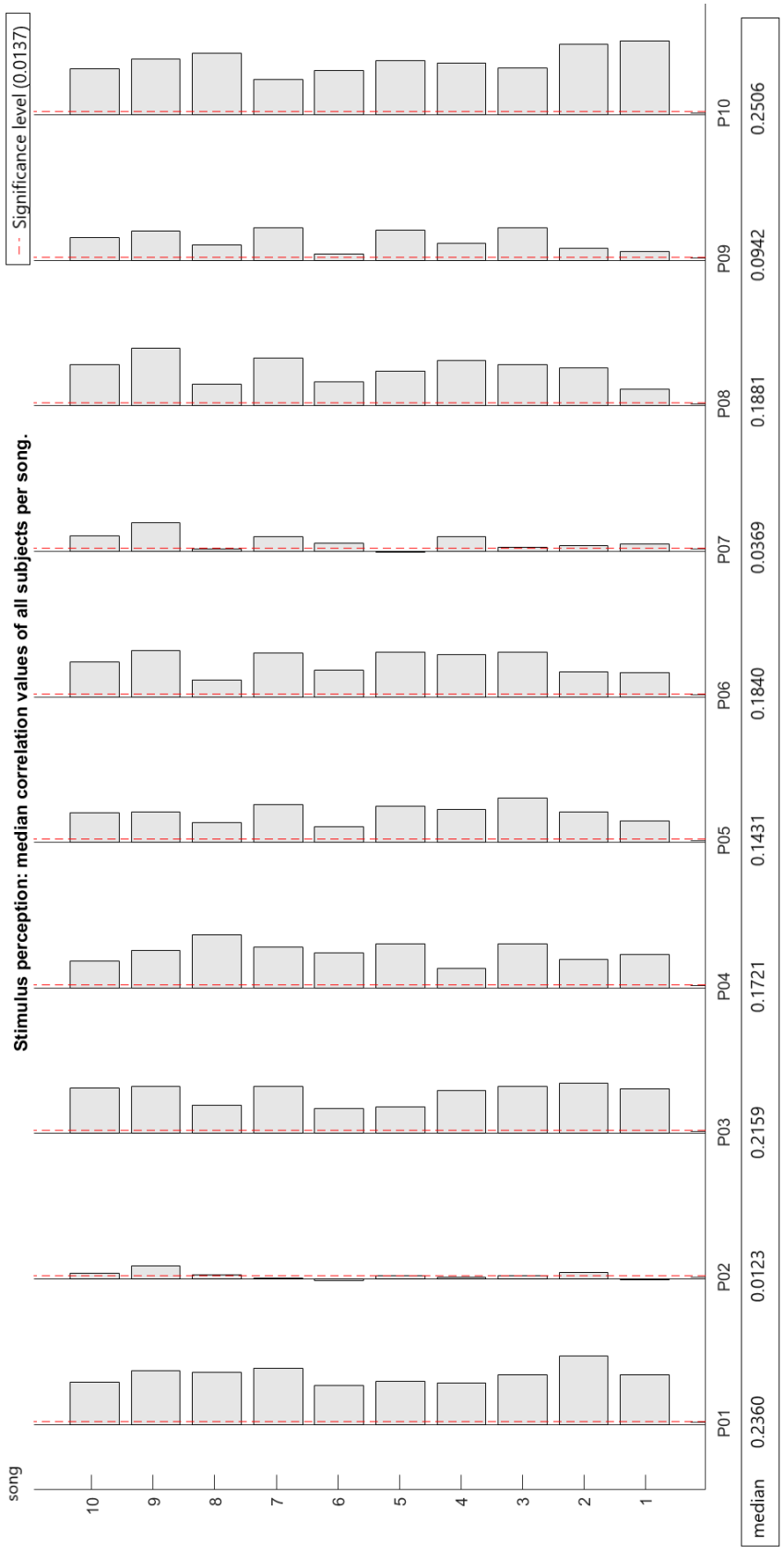

FIGURE B.12: A subdivision per subject shows the best and worst performing participants.

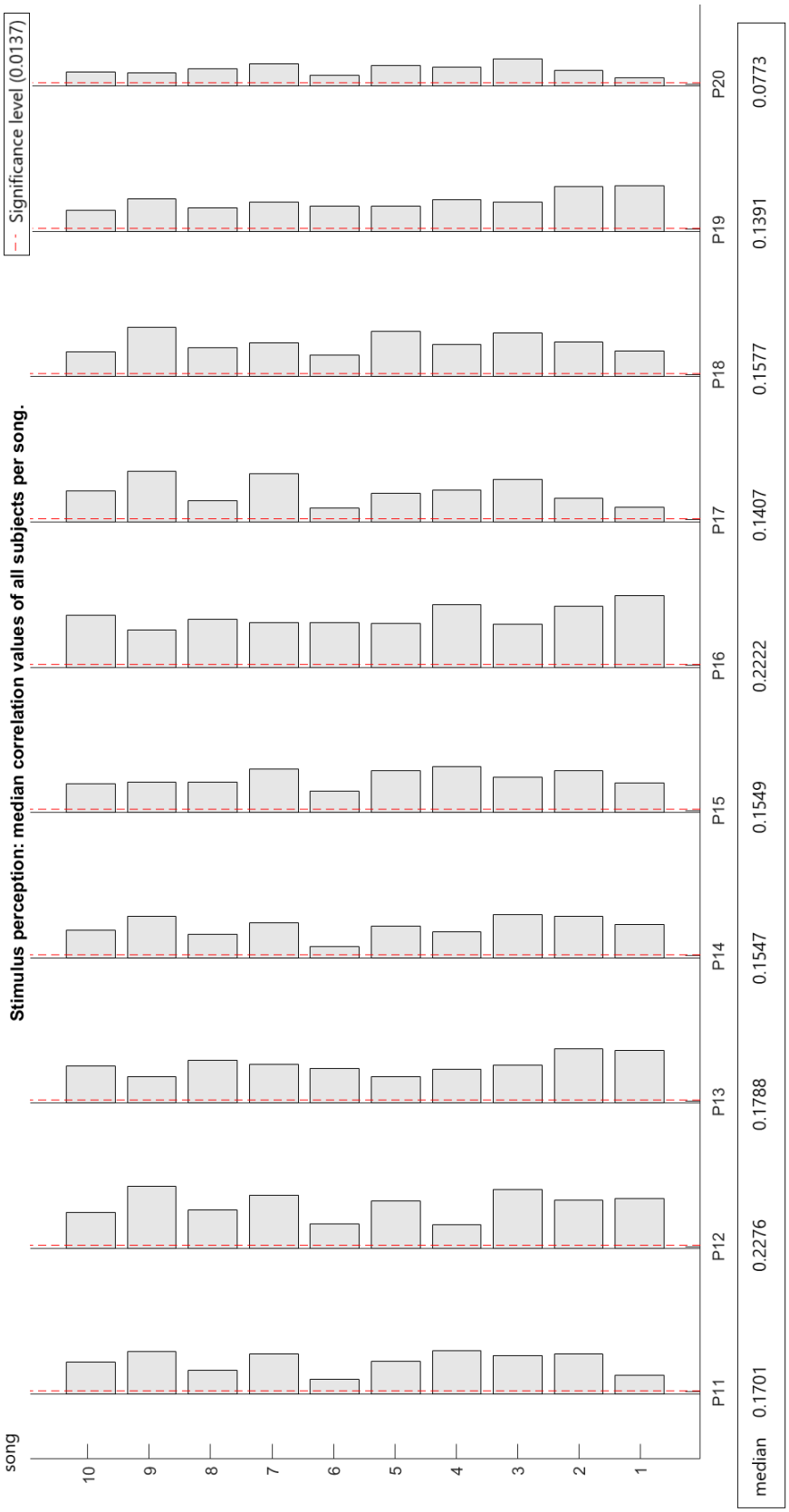

Figure B.13: A subdivision per subject shows the best and worst performing participants.

Appendix C

Including time shifts

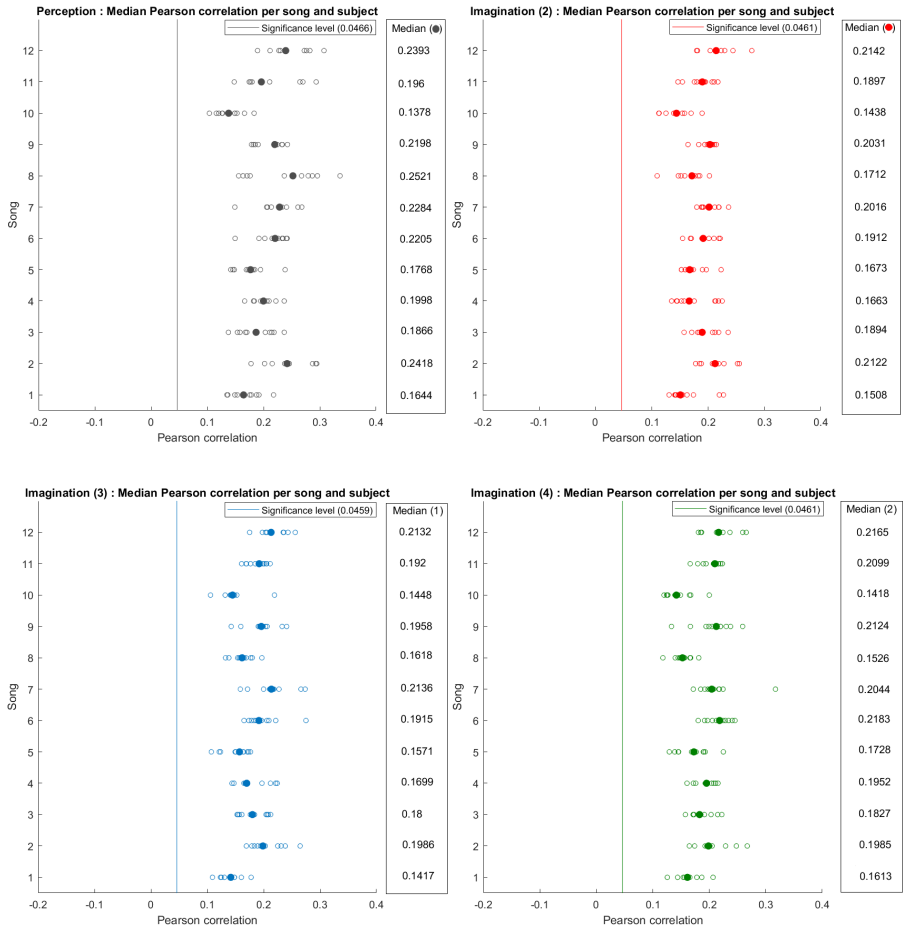

FIGURE C.1: For the OpenMIIR data set, the filtering range between 1-30 Hz performs significanty worse than the 1-10 Hz filter range.

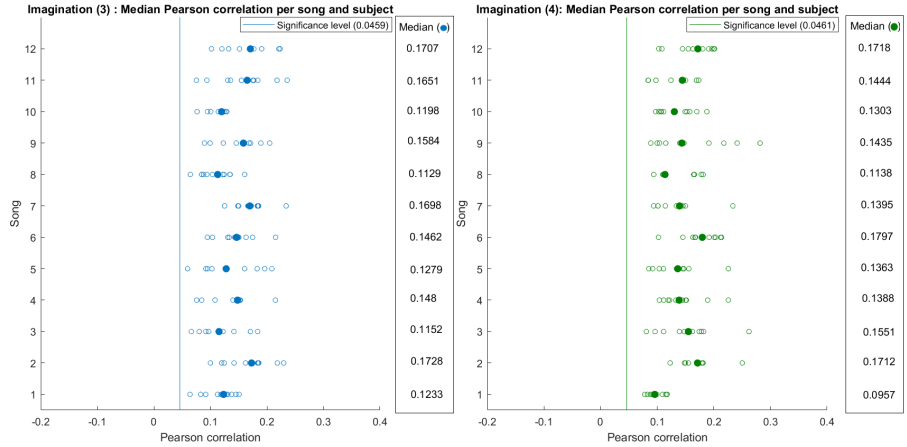

FIGURE C.2: In these figures for the estimated minimal time shift during the last two imagination experiments (OpenMIIR data set), all the median correlations are significant.

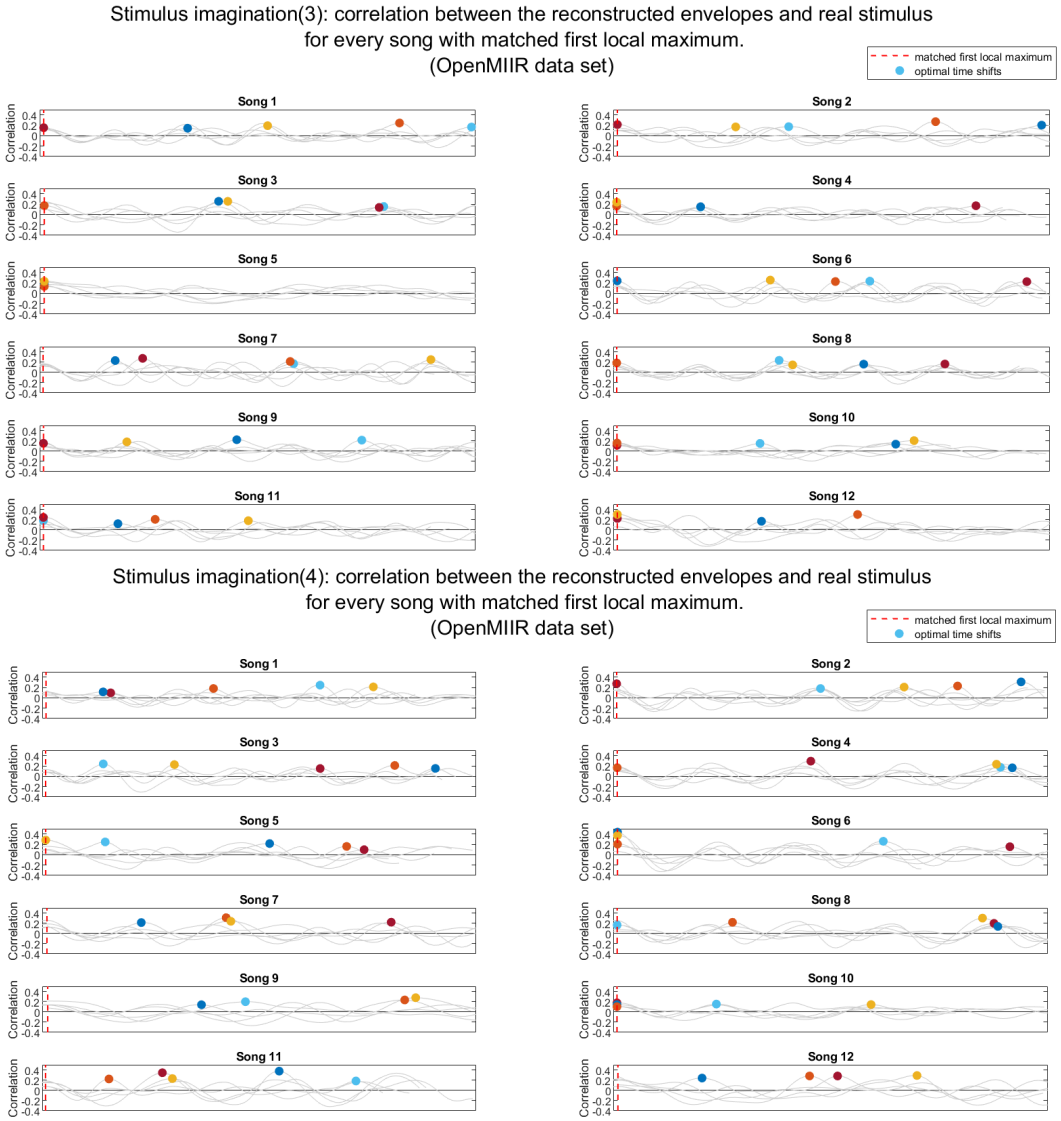

FIGURE C.3: When looking at the correlation in function of time shift for each of the five reconstructed envelopes per song, seemingly periodical patterns arise in these imagination experiments.

Appendix D

Stimulus classification

D.1 Global classifier

Table D.1: Results for the OpenMIIR data set (bandpass 1-10Hz).(* indicates a significant result ($\alpha = 0.05$))

	Global classifier							
	Perception		Imagination (2)		Imagination (3)		Imagination (4)	
	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$
P01	0.250*	0.767*	0.467	0.633*	0.033	0.633*	0.067	0.683*
P02	0.167	0.717*	0.523	0.945	0.050	0.633*	0.050	0.617*
P03	0.117	0.650*	0.541	0.950	0.067	0.700*	0.150	0.683*
P04	0.267*	0.683*	0.483	0.949	0.130	0.617*	0.150	0.650*
P05	0.150	0.750*	0.497	0.952	0.117	0.617*	0.117	0.600*
P06	0.367*	0.767*	0.486	0.950	0.067	0.600*	0.133	0.700*
P07	0.150	0.583*	0.505	0.961	0.117	0.667*	0.150	0.667*
P08	0.350*	0.700*	0.482	0.583*	0.083	0.567*	0.050	0.617*
P09	0.150	0.650*	0.600	0.650*	0.033	0.483*	0.117	0.667*
P10	0.267*	0.667*	0.474	0.533*	0.050	0.567*	0.067	0.733*

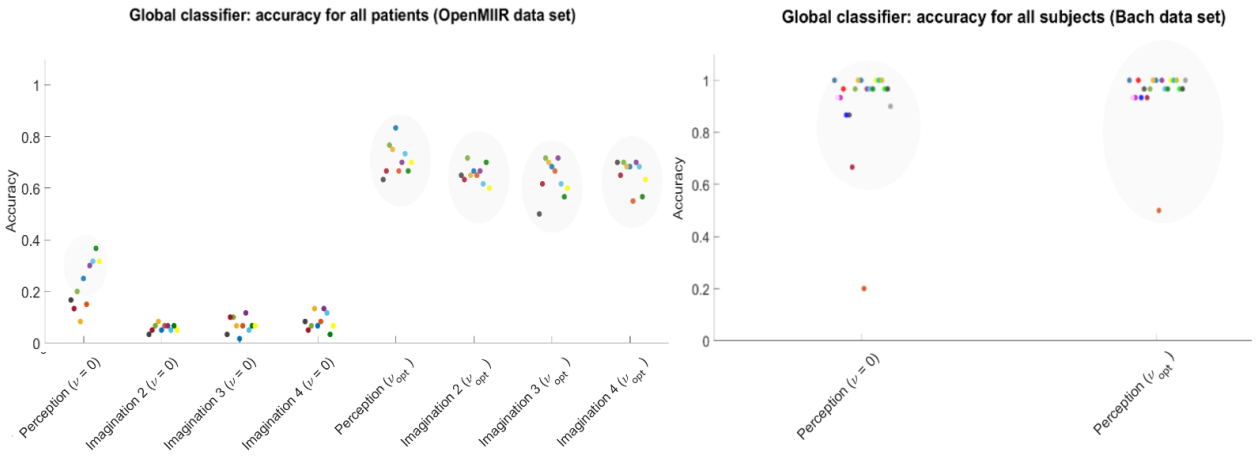

Figure D.1: Results for the global classifier (bandpass filter 1-30 Hz)

D.2 Two-song classifier

TABLE D.2: Mean accuracy for every subject for the two-song classifier (bandpass 1-10Hz).

Two-song classifier								
	Perception		Imagination (2)		Imagination (3)		Imagination (4)	
	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$
P01	0.7258	0.9697	0.4667	0.9454	0.4287	0.9530	0.5045	0.9560
P02	0.6121	0.9591	0.5227	0.9500	0.5318	0.9484	0.4667	0.9424
P03	0.6076	0.9545	0.5409	0.9485	0.5121	0.9545	0.5530	0.9621
P04	0.6757	0.9575	0.4833	0.952	0.4727	0.9363	0.5591	0.9484
P05	0.6893	0.9727	0.4969	0.9500	0.4879	0.9636	0.5273	0.9500
P06	0.8287	0.9757	0.4863	0.9606	0.5530	0.9484	0.5197	0.9590
P07	0.6727	0.9500	0.5045	0.9409	0.4970	0.9469	0.4803	0.9470
P08	0.8267	0.9575	0.4818	0.9455	0.4864	0.9530	0.4500	0.9424
P09	0.6409	0.9484	0.6000	0.9530	0.4439	0.9333	0.4500	0.9606
P10	0.8303	0.9621	0.4742	0.9545	0.4576	0.379	0.5151	0.9621

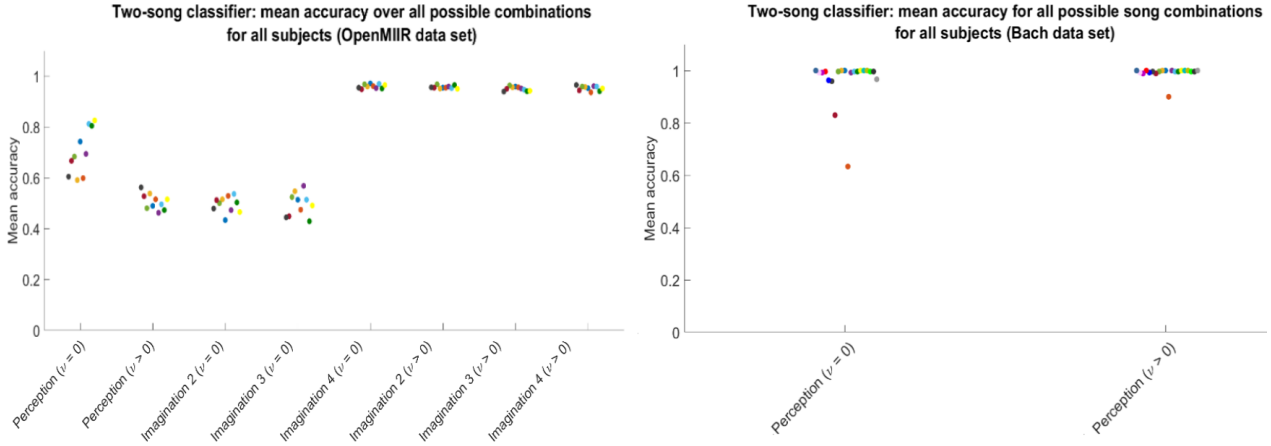

FIGURE D.2: Results for the two-song classifier (bandpass filter 1-30 Hz)

D.3 Song category classifier

Table D.3: Mean accuracy for every subject for the two-song classifier (bandpass 1-10Hz).

	Two-song classifier							
	Perception		Imagination (2)		Imagination (3)		Imagination (4)	
	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$	$\nu = 0$	$\nu > 0$
P01	0.5333	0.8667	0.4500	0.7333	0.2333	0.7667	0.3833	0.7833
P02	0.3833	0.8500	0.3333	0.7167	0.3833	0.8167	0.3500	0.700
P03	0.3667	0.7333	0.3167	0.7500	0.3333	0.7833	0.3333	0.8167
P04	0.4833	0.8000	0.2833	0.7333	0.3500	0.6667	0.3500	0.7833
P05	0.4500	0.8333	0.2500	0.7500	0.3333	0.73333	0.4000	0.6833
P06	0.5333	0.8000	0.3000	0.8167	0.2333	0.7000	0.3833	0.7667
P07	0.3833	0.7667	0.2500	0.6833	0.3667	0.7500	0.4667	0.7500
P08	0.5833	0.8333	0.2500	0.6667	0.4000	0.7000	0.3667	0.7167
P09	0.3500	0.7500	0.3833	0.8000	0.2167	0.5500	0.3667	0.7500
P10	0.5000	0.8167	0.2333	0.7000	0.3000	0.6667	0.3167	0.8000

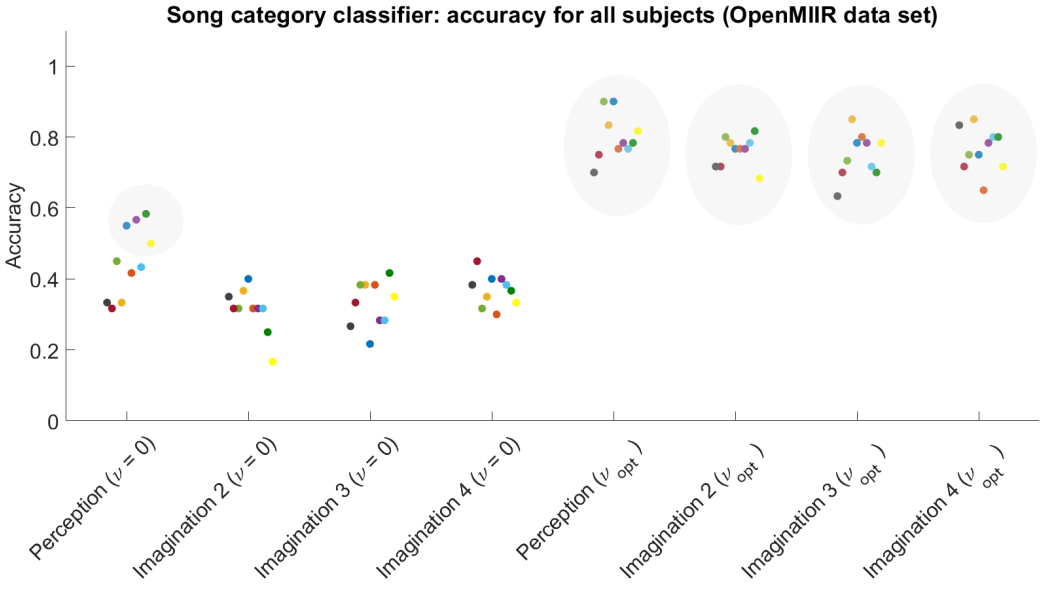

Figure D.3: Results for the Song category classifier (bandpass filter 1-30 Hz)

Bibliography

[1] Anderson, D. J., Rose, J. E., Hind, J. E., and Brugge, J. F. Temporal position of discharges in single auditory nerve fibers within the cycle of a sine-wave stimulus: frequency and intensity effects. *The Journal of the Acoustical Society of America 49*, 4B (1971), Suppl 2:1131+.

[2] Bear, M. F., Connors, B. W., and Paradiso, M. A. *Neuroscience: exploring the brain*, 4th ed. ed. Philadelphia : Wolters Kluwer, 2016.

[3] Biesmans, W., Das, N., Francart, T., and Bertrand, A. Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 25*, 5 (2017), 402–412.

[4] Biosemi. Faq. https://www.biosemi.com/faq/cms&drl.htm.

[5] Blockeel, H. *Machine learning and inductive inference*, 5de, herz. uitg. ed. Acco, Leuven, 2019.

[6] Brusheezy. Moderne muziek notities kwast.

[7] Cantisani, G., Essid, S., and Richard, G. EEG-based decoding of auditory attention to a target instrument in polyphonic music. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2019), vol. 2019-, IEEE, pp. 80–84.

[8] Chittka, L., and Brockmann, A. Perception space-the final frontier (primer). *PLoS Biology 3*, 4 (2005), e137.

[9] Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., Mesgarani, N., Quatieri, T. F., and Smalt, C. J. Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. *Scientific Reports 9*, 1 (2019).

[10] de Cheveigné, A., Wong, D. D. E., Di Liberto, G. M., Hjortkjaer, J., Slaney, M., and Lalor, E. Decoding the auditory brain with canonical component analysis. *Neuroimage 172* (2018), 206–216.

[11] Di Liberto, G., Pelofi, C., Shamma, S., and de Cheveigné, A. Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoustical Science and Technology 41*, 1 (2020), 361–364.

[12] Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., de Cheveigné, A., Shamma, S., and Mesgarani, N. Cortical encoding of melodic expectations in human temporal cortex. *eLife 9* (2020).

[13] Etard, O., Kegler, M., Braiman, C., Forte, A. E., and Reichenbach, T. Decoding of selective attention to continuous speech from the human auditory brainstem response. *NeuroImage 200* (2019), 1–11.

[14] Fuglsang, S. A., Dau, T., and Hjortkjaer, J. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage 156* (2017), 435–444.

[15] Gallun, F., Lewis, M., Folmer, R., Diedesch, A., Kubli, L., Mcdermott, D., Walden, T., Fausti, S., Lew, H., and Leek, M. Implications of blast exposure for central auditory function: A review. *Journal of Rehabilitation Research and Development 49*, 7 (2012), 1059–74.

[16] Gang, N., Kaneshiro, B., Berger, J., and Dmochowski, J. P. Decoding neurally relevant musical features using canonical correlation analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pp. 131–138.

[17] Grossman, D. J. Dave's j. s. bach page. http://www.jsbach.net/.

[18] Kučikienė, D., and Praninskienė, R. u. The impact of music on the bioelectrical oscillations of the brain. *Acta medica Lituanica 25*, 2 (2018), 101–106.

[19] Landegger, L. D., Vasilijic, S., Fujita, T., Soares, V. Y., Seist, R., Xu, L., and Stankovic, K. M. Cytokine levels in inner ear fluid of young and aged mice as molecular biomarkers of noise-induced hearing loss. *Frontiers in Neurology 10* (2019).

[20] Leuchs, L. Choosing your reference - and why it matters. *Brain Products* (03-05-2019).

[21] Lu, J., Wu, D., Yang, H., Luo, C., Li, C., and Yao, D. Scale-free brain-wave music from simultaneously EEG and fMRI recordings (scale-free brain-wave music). e49773.

[22] The Mathworks, Inc. *MATLAB version R2019b*. Natick, Massachusetts, 2019.

[23] Matlab. Linear mixed-effects models. https://www.mathworks.com/help/stats/linear-mixed-effects-models.html.

[24] Matlab. Wilcoxon signed rank test (signrank). https://www.mathworks.com/help/stats/signrank.html.

[25] Meister, I., Krings, T., Foltys, H., Boroojerdi, B., Müller, M., Töpper, R., and Thron, A. Effects of long-term practice and task complexity in musicians and nonmusicians performing simple and complex motor tasks: Implications for cortical motor organization. *Human Brain Mapping 25*, 3 (2005), 345–352.

[26] Microsoft Corporation. Paint3D, 3D Library.

[27] Mueller, K., Mildner, T., Fritz, T., Lepsien, J., Schwarzbauer, C., Schroeter, M. L., and Möller, H. E. Investigating brain response to music: A comparison of different fMRI acquisition schemes. *NeuroImage 54*, 1 (2011), 337–343.

[28] Neuroscientifically Challenged. 2-minute neuroscience: Electroencephalography (EEG). https://www.youtube.com/watch?v=t$ZcKT4l_JZk$.

[29] New York Times. Dr. Georg von Bekesy, 73, dies; '61 Nobel Laureate in Medicine.

[30] Oldendorf, W., and Oldendorf, W. *Advantages and Disadvantages of MRI. In: Basics of Magnetic Resonance Imaging.Topics in Neurology (Oldendorf, Basics of Magnetic Resonance Imaging), vol 1. Springer, Boston, MA.*

[31] O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex 25*, 7 (2015), 1697–1706.

[32] Rangayyan, R. M. *Biomedical signal analysis*, 2nd ed. ed. IEEE Press series in biomedical engineering. Wiley, Hoboken, New Jersey, 2015.

[33] Schaefer, R. S., Farquhar, J., Blokland, Y., Sadakata, M., and Desain, P. Name that tune: Decoding music from the listening brain. *Neuroimage 56*, 2 (2011), 843–849.

[34] Sebastian, S. Toward studying music cognition with information retrieval techniques: Lessons learned from the OpenMIIR initiative. *Frontiers in Psychology 8* (2017).

[35] Somers, B., Francart, T., and Bertrand, A. A generic EEG artifact removal algorithm based on the multi-channel wiener filter. *Journal of Neural Engineering 15*, 3 (2018).

[36] Stober, S., Sternin, A., Owen, A. M., and Grahn, J. A. Towards music imagery information retrieval: Introducing the OpenMIIR dataset of EEG recordings from music perception and imagination. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, pp. 763–769.

[37] Sturm, I., Dahne, S., Blankertz, B., and Curio, G. Multi-variate EEG analysis as a novel tool to examine brain responses to naturalistic music stimuli. *Plos One 10*, 10 (2015).

[38] Suetens, P. *Fundamentals of medical imaging*, 3rd ed. ed. Cambridge University Press, Cambridge, 2017.

[39] Tal, I., Large, E. W., Rabinovitch, E., Wei, Y., Schroeder, C. E., Poeppel, D., and Zion Golumbic, E. Neural entrainment to the beat: The "missing-pulse" phenomenon. *The Journal of neuroscience : the official journal of the Society for Neuroscience 37*, 26 (2017), 6331–6341.

[40] Tempel, B. L. Nub502: The auditory system.

[41] Treder, M. S., Purwins, H., Miklody, D., Sturm, I., and Blankertz, B. Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *Journal of Neural Engineering 11*, 2 (2014), 10.

[42] van der Mee, T., Reijntjens, M., Schaefer, R., and Scherder, E. Algemeen Dagblad. Dit is wat muziek met ons brein doet. https://www.ad.nl/wetenschap/dit-is-wat-muziek-met-ons-brein-doet a80871d6/?referrer=https://www.google.com/, 20-01-2019.

[43] Verschueren, E., Somers, B., and Francart, T. Neural envelope tracking as a measure of speech understanding in cochlear implant users. *Hearing Research 373* (2019), 23–31.

[44] Wong, D. D. E., Fuglsang, S. A., Hjortkjaer, J., Ceolini, E., Slaney, M., and de Cheveigné, A. A comparison of regularization methods in forward and backward models for auditory attention decoding. *Frontiers in Neuroscience 12* (2018).

[45] Yurgil, K. A., Velasquez, M. A., Winston, J. L., Reichman, N. B., and Colombo, P. J. Music training, working memory, and neural oscillations: A review. *Frontiers in psychology 11* (2020), 266–266.