EEG-based decoding and recognition of imagined music

Marthe Tibo[1], Simon Geirnaert[1,2], and Alexander Bertrand[1,*]

[1]KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
[2]KU Leuven, Department of Neurosciences, Research Group ExpORL
[*]alexander.bertrand@kuleuven.be

ABSTRACT

When listening to music, the brain generates a neural response that follows the amplitude envelope of the musical sound. Previous studies have shown that it is possible to decode this envelope-following response from electroencephalography (EEG) data during music perception. However, a successful decoding and recognition of imagined music, without the physical presentation of a music stimulus, has not been established to date. During music imagination, the human brain internally replays a musical sound, which naturally leads to the hypothesis that a similar envelope-following response might be generated. In this study, we demonstrate that this response is indeed present during music imagination and that it can be decoded from EEG data. Furthermore, we show that the decoded envelope allows for classification of imagined music in a song recognition task, containing tracks with lyrics as well as purely instrumental tasks. A two-song classifier achieves a median accuracy of 95%, while a 12-song classifier achieves a median accuracy of 66.7%. The results of this study demonstrate the feasibility of decoding imagined music, thereby setting the stage for new neuroscientific experiments in this area as well as for new types of brain-computer interfaces based on music imagination.

Introduction

Every day, our brains are exposed to millions of different sounds, ranging from simple tone beeps to highly complex stimuli, like music. When musical sounds are picked up by the human ear, signals travel via the auditory nerve to a variety of brain areas, where a complex chain of processing events occurs. These processes can for example make us emotionally react to music, can bring back memories, or let us play along to the song on an instrument.

Apart from listening to music - termed *music perception* - the human brain is capable of an even more remarkable task. Our minds can imagine complex sounds without the presence of a physical source of sound. This phenomenon, so-called *music imagination*, can for example lead to an 'earworm', where a song is stuck inside of our heads for a long duration of time[1].

However, a lot about the specific cortical processing and spatial activation during the perception and imagination of music remain unknown. To shed more light on these functions, the perception and imagination processes can be summarized into models, which define the relation between an auditory stimulus and its neural response. By designing these models, one can reconstruct the amplitude envelope of a sound stimulus from the neural responses to this stimulus or vice versa.

Multiple previous studies have created such models for music *perception*. Di Liberto et al.[2] reconstructed stimulus envelopes out of EEG responses during music perception experiments with musicians and non-musicians to explore possible neural differences between both groups. Sturm et al.[3] decoded audio power slopes out of EEG responses and investigated their relation to behavioral tension data. Classification approaches, such as in the study of Schaefer et al.[4], have shown that it is possible to distinguish between seven distinct songs during music perception by using a quadratically regularized linear logistic regression classifier on EEG data. Stober et al.[1,5] conducted classification experiments for 12 songs during music perception, using a similarity-constraint encoding approach. Apart from working on perception data, the latter also concentrated on the EEG responses of imagination experiments. However, no consistent patterns were found in the data and an insignificant classification accuracy was obtained.

In this paper, we provide evidence that this EEG-based decoding and song classification is also possible for *imagined* music, which up to our knowledge has not been established to date. We adopt a linear stimulus reconstruction approach as a prior step to perform song classification, which is borrowed from the neighboring domain of neural decoding of speech. In the field of speech processing, linear decoding techniques achieve good results with regards to the auditory attention decoding problem, such as in the studies of O'Sullivan et al.[6], Fuglsang et al.[7], Wong et al.[8], Das et al.[9], and Geirnaert et al.[10]. While such decoding strategies have been successfully applied for music perception[2–4,11–14], they have not been successful in decoding imagined music so far[5]. In this work, the decoder is trained on EEG data of a music perception experiment and then applied to EEG data from a music imagination experiment. For the latter, the decoder is complemented with a delay operator that searches

for the (unknown) start of the imagined music, leading to an improved quality of the obtained reconstructions. Finally, these reconstructed envelopes will be labeled by classifiers w.r.t. various properties, obtaining a significant classification accuracy in both the perception and imagination experiments. In essence, this study shows a way to successfully classify imagined music (or possibly other stimuli), paving the way for new software tools for analyzing neural activity during music imagination and new brain-computer interface paradigms. In the long term, this could even lead to new futuristic applications such as hands-free and voice-free song selection, song recognition, a 'silent' version of the popular Shazam app, etc.

In the next sections, we will first give an overview of the data used in this study, followed by the applied preprocessing steps. Afterwards, the adapted linear decoding model is introduced in detail, together with its modified decoding strategy. Subsequently, we will introduce various classification tasks, which are discussed and analyzed for both the perception and imagination of music, where the former acts as a baseline.

Methodology

Data

Two publicly available data sets are used, containing the EEG responses on musical stimuli for various subjects. Both data sets were recorded using a 64-channel BioSemi Active Two system at a sampling frequency of 512 Hz.

OpenMIIR data set

The OpenMIIR data set, published by Stober et al.[1], contains 12 stimuli with polyphonic music, consisting of 4 nursery rhymes with a singing voice and musical accompaniment, the same 4 nursery rhymes without singing voice, and 4 well-known orchestral works (8 out of 12 stimuli are purely instrumental without singing voice). The EEG responses to each of these stimuli were recorded in 10 subjects during four different experiments:

- music perception, preceded by cue clicks (with the frequency equal to the beat of the corresponding music stimulus);
- music imagination, preceded by cue clicks;
- two music imagination experiments without cue clicks.

For the last two imagination experiments, one set-up was recorded in the same block as the previous ones, while another was recorded as a standalone trial. These standalone trials were registered for all music stimuli immediately after all recording blocks of the previous three conditions were performed (approximately 30 minutes after the start of the first recording block). Five repetitions of each music presentation/imagination are available per subject, summing to 60 trials per subject and experiment. The music stimuli in this data set have an average length of 10.4 s.

Bach data set

This data set, published by Di Liberto et al.[2], contains 10 excerpts from the partitas and sonatas of J.S. Bach and their EEG responses for 20 subjects. All stimuli are purely instrumental and played from a MIDI file with piano samples. This data set only contains a perception EEG experiment, meaning that these trials will solely be used to investigate the effect of the modified decoding strategy on perception data. Each of these stimuli is presented three times to a subject, leading to a total of 30 trials. The average length of the excerpts is 120 s.

Preprocessing

First, the EEG trials are re-referenced to the average of the two recorded mastoid channels. For the OpenMIIR data set, however, this step is omitted, since the mastoid channels were only recorded for half of the subjects. Instead, we retained the original reference (common mode sense active electrode) on all EEG data (incl. mastoid channels). Afterwards, a multi-channel Wiener filter[15] removes eye blinks and other disturbances from the data. After this step, the clean EEG channels are filtered using a first-order Butterworth filter with a frequency range between 1 and 10 Hz. This choice is inspired by research in the domain of speech processing, where it was shown that the envelope and its EEG response are linearly relatable over the aforementioned frequency range[6,8,9]. As a final step, the EEG is downsampled to 64 Hz.

For each of the music stimuli, an envelope is extracted based on the procedure described in Biesmans et al.[16]. In short, the signal is split into 15 subbands using a gammatone filter bank, with center frequencies ranging from 1 Hz to 5 kHz. The amplitude envelope of each subband signal is compressed using a power law with the exponent equal to 0.6. Afterwards, the envelopes are summed across the subbands, resulting in a single aggregated envelope. Finally, the aggregated envelope is filtered between 1–10 Hz using the same bandpass filter as in the EEG preprocessing and subsequently downsampled to 64 Hz.

Decoding and classification

Stimulus reconstruction

To reconstruct an envelope from the corresponding EEG response, a linear decoding model is used. Consider a music stimulus envelope $s(t)$ and its N-channel EEG response $\mathbf{r}(t) = \begin{bmatrix} r_1(t) & \cdots & r_N(t) \end{bmatrix}^T$, here represented as an N-dimensional vector, where

$r_n(t)$ represents the t-th sample in the n-th EEG channel, and where $(\cdot)^T$ denotes the transpose operator. The decoder then linearly combines time-lagged versions of the EEG channels to estimate the envelope $\hat{s}(t)$[6]:

$$\hat{s}(t) = \sum_{n=1}^{N} \sum_{\tau=0}^{L-1} g_n(\tau) r_n(t+\tau). \tag{1}$$

where $g_n(\tau)$ denotes the decoder weight for the n-th channel on the τ-th time lag. In this model, we choose a range of time lags up to 250 ms, which is inspired by the speech decoding literature[6]. At a sampling rate of 64 Hz, this corresponds to up to $L = 17$ sample shifts (incl. the zero shift). However, this model omits an important parameter for music imagination. During these experiments, the exact onset of the imagined music is unknown, resulting in a potential mismatch between the decoded music envelope and the baseline envelope. To counteract this effect, one could choose to broaden the range of time lags. However, this introduces a large amount of extra parameters into (1), resulting in overfitting effects and ill-conditioned matrix inversions in the calculation of the decoder (see below). Therefore, we include an extra global time shift ν into the model, independent of the time lags τ, which globally shifts the onset of the EEG signal:

$$\hat{s}_\nu(t) = \sum_{n=1}^{N} \sum_{\tau=0}^{L-1} g_n(\tau) r_n(t+\tau+\nu). \tag{2}$$

Since the time delay ν between the actual music envelope and the response generated by the brain is unknown during (uncued) music imagination, an optimal value for ν has to be estimated in order to maximize the neural decoding performance. We will return to this issue later on (see 'Testing the decoder').

Finally, we note that the decoder is modeled as a linear time-invariant system, which implies that the reconstructed envelope $\hat{s}(t)$ in (2) is equal to a delayed version of the one in (1), i.e., $\hat{s}_\nu(t) = \hat{s}(t+\nu)$. Therefore, in the remaining of this paper, we will omit the subscript ν in $\hat{s}_\nu(t)$ and will simply write $\hat{s}(t+\nu)$ when referring to the delayed version.

Training the decoder

The decoder is trained on the music *perception* trials. In this case the global delay parameter ν can be set to $\nu = 0$, since the neural activity then entrains to the stimulus with only a short neural path delay, which is mostly covered by the time shifts represented by τ. The decoder weights are optimized such that they minimize the squared error between the original music envelope $s(t)$ and the reconstructed envelope $\hat{s}(t)$ in (1) on a training set of T samples:

$$\min_{\mathbf{g}} \sum_{t=1}^{T} \left(s(t) - \hat{s}(t)\right)^2 + \lambda \|\mathbf{g}\|_2^2 \tag{3}$$

where $\mathbf{g}$ is an NL-dimensional vector in which all decoder weights $g_n(\tau)$ are stacked (for all channels n and time shifts τ). The second term is a regularization term to counteract overfitting effects[6,8], where λ is a tunable regularization weight. (3) can be rewritten in matrix-vector format as:

$$\min_{\mathbf{g}} \|\mathbf{s} - \mathbf{R}\mathbf{g}\|_2^2 + \lambda \|\mathbf{g}\|_2^2, \tag{4}$$

where $\mathbf{s} = [s(1) \ldots s(T)]^T$ and $\mathbf{R}$ is a $T \times NL$ matrix containing all N EEG channels plus their $L-1$ time-lagged copies in its columns:

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}(1)^T & \mathbf{r}(2)^T & \cdots & \mathbf{r}(L)^T \\ \mathbf{r}(2)^T & \mathbf{r}(3)^T & \cdots & \mathbf{r}(L+1)^T \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{r}(T-1)^T & \mathbf{r}(T)^T & \cdots & 0 \\ \mathbf{r}(T)^T & 0 & \cdots & 0 \end{bmatrix}. \tag{5}$$

The least squares decoder is then found as

$$\mathbf{g} = \left(\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I}\right)^{-1} \left(\mathbf{R}^T\mathbf{s}\right) = \left(\mathbf{C}_{RR} + \lambda\mathbf{I}\right)^{-1} \mathbf{C}_{Rs}, \tag{6}$$

where $\mathbf{I}$ denotes the identity matrix, and $\mathbf{C}_{RR}$ and $\mathbf{C}_{Rs}$ are respectively the estimated autocorrelation matrix of the EEG data and the estimated cross-correlation vector between the EEG data and the stimulus envelope. The decoder is always trained on a per-subject basis. The regularization hyperparameter λ is tuned based on a leave-one-song-out cross-validation procedure over a grid defined by $\lambda = 1.848^k \times 10^{-6}$ with $k = [0, 50]$[8]. The value of λ that leads to the highest Pearson correlation coefficient $\rho(s(t), \hat{s}(t))$ between the reconstructed and the original envelope (averaged across all folds) is selected as the optimal regularization parameter, after which the decoder is retrained once more on all training data using this parameter.

Testing the decoder

When decoding imagined music, the global shift parameter ν can not be set equal to 0 due to an unknown delay between the imagined/decoded music track and the original envelope of the corresponding audio track. This delay parameter ν is different in each trial and is therefore estimated on a per-trial basis. To this end, we test the pre-trained decoder over a range of values for ν and select the value that yields a reconstructed envelope $\hat{s}(t+\nu)$ that has the highest Pearson correlation coefficient with the original stimulus $s(t)$, i.e.,

$$\nu_{\text{opt}} = \underset{\nu}{\text{argmax}}\ \rho(s(t), \hat{s}(t+\nu)). \tag{7}$$

The shift $\nu = \nu_{\text{opt}}$ then denotes the global shift at which the reconstruction is maximally similar to the real envelope. During music imagination, this delay parameter ν is varied between 0–2 s.

For the decoding experiments in case of music imagination, the training data consists of all repetitions of all music tracks from the perception experiment, while the test set consists of the trials from the imagination experiment. For the tuning of the regularization parameter λ, a cross-validation procedure is used on the training set. We test all music imagination conditions, i.e., with and without cue clicks. When testing the decoder performance for the perception case, a nested cross-validation scheme is used, where in each fold a single music track (incl. all repeated perception trials) is used as the test set, while all repeated perception trials of all other music stimuli of the same subject is used as training set. The cross-validation for the tuning of the regularization parameter λ is now nested inside this outer-loop cross-validation.

The decoding performance is quantified by the Pearson correlation coefficient $\rho(s(t), \hat{s}(t+\nu_{\text{opt}}))$ between the reconstructed and original envelope on the test trials. While the introduction of the global shift parameter ν was introduced to cope with the unknown time offset when decoding music imagination, we also include it here for the testing of decoding music perception to allow for a fair comparison between both cases. This is because the omittance of the extra optimization in (7) might introduce an unfair disadvantage for the music perception case as the latter would then be restricted to neural delays up to $L-1$ samples. The delay parameter ν is again varied between 0–2 s for the perception experiment in the OpenMIIR data set, but a range of 0–1 s was chosen for the Bach data set, due to a limited amount of additional EEG data with respect to the stimulus length.

Classification of songs

The main goal of this work is to identify with which music track a decoded music envelope corresponds. To this end, the decoded envelope $\hat{s}(t)$ is compared to each music envelope $s_i(t)$ in the database of music tracks or songs ($i = 1, \ldots, 12$ for the OpenMIIR data set, and $i = 1, \ldots, 10$ for the Bach data set). The maximal correlation coefficient ρ_i is identified for each music track i as

$$\rho_i = \max_{\nu} \rho(s_i(t), \hat{s}(t+\nu)). \tag{8}$$

The perceived/imagined song can then be detected by selecting the music track i for which ρ_i is maximal. It is noted that the quality of the estimated Pearson correlation coefficient depends on the number of available time samples over which it is estimated. If more time samples are available, the estimation noise on ρ_i will be lower, thereby becoming a more discriminative feature for classification. This introduces an important tradeoff between the decision window length or required recording time on the one hand, and the accuracy of the classification on the other hand. A similar tradeoff exists in the context of auditory attention decoding on speech[10,17]. In this paper, Pearson correlation coefficients are always computed over the full trial length, unless specified otherwise.

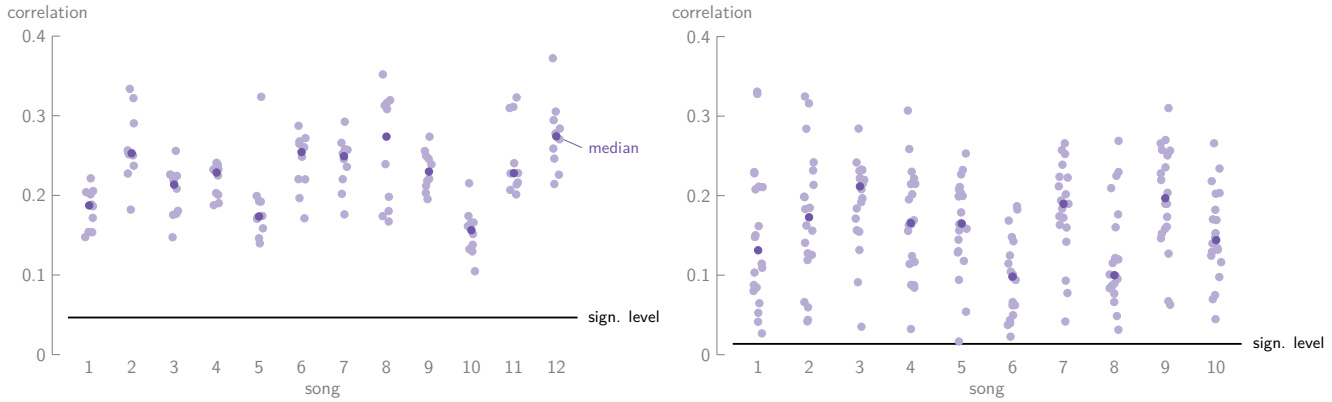
We will consider three different classification experiments; a two-song classifier, a K-song classifier (with K the number of available songs in the respective data set), and a music category classifier, where the latter distinguishes between the three song categories in the OpenMIIR data set.

Results and discussion

Music perception

First, we assess the performance of our linear decoding model on the music perception experiments. For both data sets, Figure 1 shows the median (across repeated trials) Pearson correlation between the optimal reconstructed envelopes and their corresponding real envelopes for each subject. The significance level is calculated by correlating 10,000 randomly time-permuted reconstructed envelopes with their corresponding real envelopes[18], thereby destroying the correlation structure. The median correlation (across subjects and repetitions) is then computed, and the 97.5 percentile of these median correlations is used as the significance level.

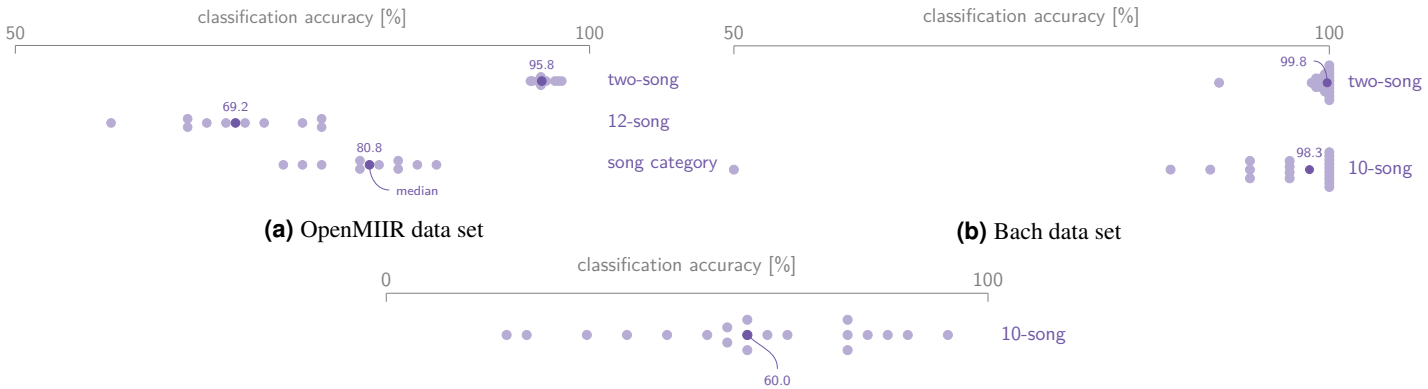
Using this permutation test, we see significant median Pearson correlations for the perception experiment over all stimuli (songs) in both data sets. The median over all subjects and songs is equal to 0.21 for the OpenMIIR data set and 0.16 for the Bach data set.



(a) OpenMIIR data set

(b) Bach data set

Figure 1. Music perception: Median correlations for each subject (light purple dots) and song in both data sets. Darker colored dots display the median over all subjects.



(a) OpenMIIR data set

(b) Bach data set

(c) Bach data set: 10 s decision window

Figure 2. Music perception: Classification accuracy for each subject (light purple dots) during the two-song, K-song, and song category classification tasks. The darker colored dots represent the median accuracy over all subjects.

Note that de Cheveigné et al.[19] investigated the influence of a delay parameter during the decoding of perceived speech signals. They found, for these signals, similar higher correlations between the real and reconstructed speech envelopes when including this delay parameter into the model, compared to the omission of this time shift.

Two-song classifier

In this task, we investigate whether it is possible to distinguish the reconstructed envelopes of two stimuli from each other, which corresponds to a match-mismatch task. We therefore correlate each reconstructed envelope with the real envelopes of two stimuli, of which one of both is matched to the corresponding reconstructed envelope. We select the highest correlation out of this duo, which is ideally linked to the matching real envelope. Figure 2 shows the mean accuracy of this classification task over all possible stimulus pairs for each subject. The overall median accuracy over all subjects is equal to 95.8% for the OpenMIIR data set and 99.8% for the Bach data set. The seemingly high results for this classifier prompts us to explore more complex strategies in the next subsections.

K-song classifier

Secondly, we investigate whether the previous strategy can be extended to discern K stimuli from each other, where K is the total number of songs in each data set ($K = 12$ for the OpenMIIR data set, and $K = 10$ for the Bach data set.). The accuracy per subject for this K-song classifier is shown in Figure 2. Again a high and significant accuracy per subject is found (McNemar test, $p \ll 10^{-4}$), with a median over all subjects of 69.2% for the OpenMIIR data set and 98.3% for the Bach data set. Significance is determined by means of a McNemar test per subject, comparing these results to those of a classifier that always predicts one of the perfectly balanced stimulus classes. We observe an overall higher performance for the Bach data set, which can be attributed due to the smaller number of classes (12 vs 10) and the availability of longer trials leading to more accurate Pearson correlation estimates. To investigate the influence of trial length on the K-song classification task, this set-up was repeated using only the first 10 s of the reconstructed and real stimulus as decision window for classification, comparable to the OpenMIIR

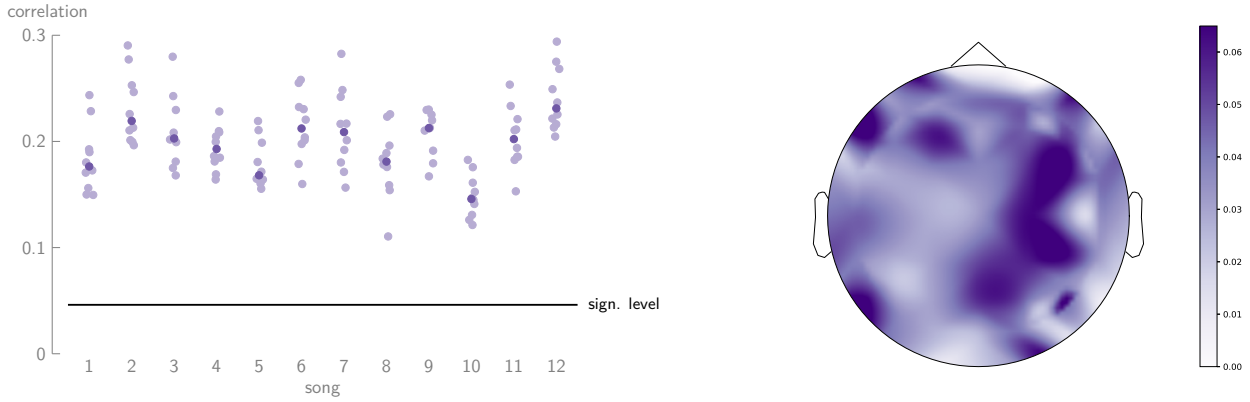
data set, leading to a median accuracy of 60.0% (Figure 2c). This implies that the high performance on the Bach data set is attributed to the longer trials over which the correlation is calculated.

Song category classifier

In a third classification task, the stimuli of the OpenMIIR data set will be labeled as one of three music categories: nursery rhymes with lyrics (category 1), the same nursery rhymes without lyrics (category 2), and orchestral music (category 3). The per-subject accuracies are shown in Figure 2a. All of them are significant (McNemar test per subject, $p \ll 10^{-4}$), with a median over all subjects of 80.8%. As expected, the accuracy for each subject is higher than for the K-song classification, but lower than for the two-song classification.

Music imagination

Secondly, we assess the decoding performance during music imagination. We first test the decoder without applying a global time shift, i.e., $\nu = 0$. It was found that this leads to non-significant and even negative median correlations across subjects and songs for all imagination experiments, based on the previously defined permutation test. Similarly, performing the aforementioned classification tasks on the imagined music leads to an overall low accuracy. The results for the song category classifier are all non-significant (McNemar test, $0.075 < p < 0.89$ across subjects), with a median accuracy across subjects equal to 29.1% for the cued imagination, 33.0% for the uncued imagination, and 37.0% for the standalone experiment. For the 12-song classification task, the median accuracies are equal to 5.8% for the cued imagination, 6.7% for the uncued imagination, and 11.7% for the standalone experiment, which again are all non-significant. We conclude that decoding imagined music does not work if there is no compensation mechanism for the (unknown) onset mismatch of the imagined music, even during cued experiments.



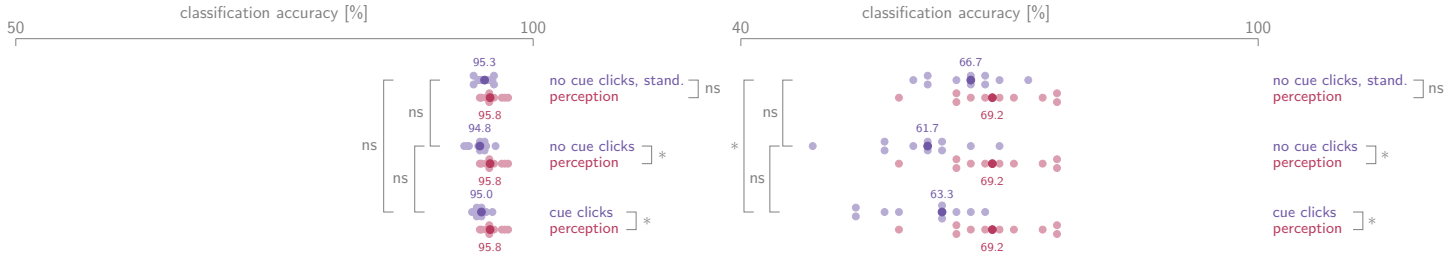
(a) Median correlations

(b) Decoder weight topology

Figure 3. Music imagination: (a) Median correlations (across repeated trials) for each subject (lighter dots) and song during cued music imagination for the model with optimal time shift. Darker colored dots display the median over all subjects. (b) Decoder weight topology used for envelope reconstruction, for all 64 EEG channels of the OpenMIIR data set. The displayed weights are computed by averaging the absolute values of the subject-specific decoders across subjects and time lags τ.

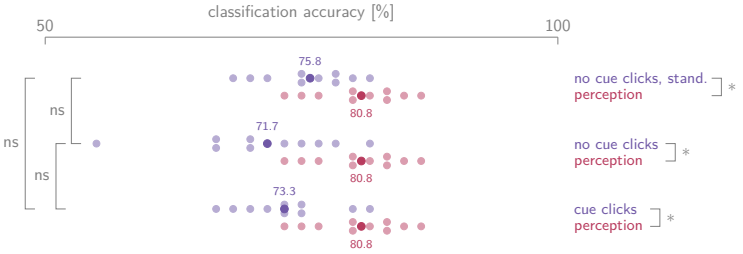
Therefore, we repeat the same process using an optimal time shift $\nu = \nu_{\text{opt}}$ for each of the three imagination experiments. Figure 3a shows the median (across repeated trials per subject) Pearson correlation between the optimally reconstructed envelopes and their corresponding real envelopes during the cued imagination experiment. The median correlation over all subjects and songs is equal to 0.20 for the cued music imagination, 0.19 for the uncued music imagination, and 0.21 for the standalone experiment. Interestingly, the obtained correlations are all above the significance level, meaning that the envelope reconstruction benefits from the inclusion of the time shift ν, confirming the hypothesis of a possible mismatch between the trial and imagination onsets. However, as expected, these correlation values are lower than for the perception experiment in the OpenMIIR data set (Wilcoxon signed rank test, $p < 10^{-3}$), indicating possible distortions in the imagined stimuli (e.g., tempo inconsistencies), which lead to a reduced similarity between the reconstructed and real envelopes during testing.

Figure 3b shows the topology of the decoder weights used for envelope reconstruction, which are trained on perception data. These weights are obtained by averaging the absolute values of the subject-specific decoders over all time lags and subjects. A complex activity pattern arises, involving many different brain areas. However, by the nature of the used decoding technique, it should be noted that these weights do not necessarily display the activation pattern in the brain related solely to music processing. Decoder weights extract the data of interest out of the EEG signal, improving the signal-to-noise ratio, but this process is dependent on the existing noise pattern. This filtering process could therefore lead to lower weights at musically



(a) Two-song classifier

(b) 12-song classifier



(c) Song category classifier

Figure 4. Music imagination: Classification accuracy for each subject (light purple dots) during the two-song, 12-song, and song category classification tasks, compared to the perception case (red dots). The darker colored dots represent the median accuracy over all subjects. * refers to a significant difference on the 0.05-level (Wilcoxon signed rank test), ns to a non-significant difference.

relevant channels, while less informative channels are included into the decoding model with substantial weights[20]. A forward modeling technique, which directly relates a music stimulus to its neural activation, could overcome this issue[12].

Two-song classifier

When applying the two-song classifier to our imagination data, we see a high mean accuracy over all possible song categories for every subject in Figure 4a. The overall median over all subjects is 95.0% for the cued, 94.9% for the uncued, and 95.3% for the standalone imagination experiment. Comparing these accuracies to the results during music perception, the latter significantly outperforms the music imagination experiments (Wilcoxon signed rank test, $p < 0.0039$), except for the standalone imagination experiment ($p = 0.055$). The results for this standalone condition, which is recorded last for all subjects, might be influenced by a better memorization of the music stimulus across the experiments. No significant differences were found between the three imagination experiments.

12-song classifier

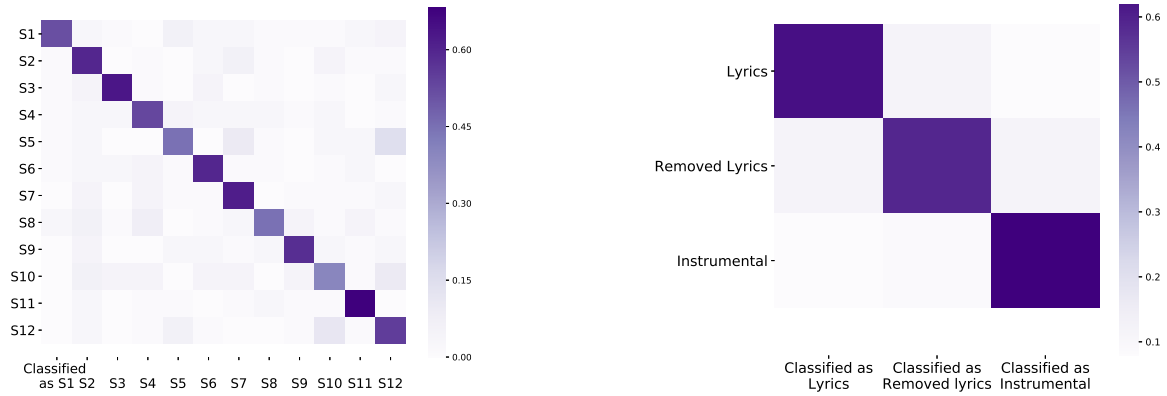
In Figure 4b, the 12-song classification accuracy is shown for each subject. For each of the three imagination experiments, the results are significant (McNemar test per subject, $p \ll 10^{-4}$), although significantly lower than the perception case. The median accuracy over all subjects is 63.3% for the cued imagination, 61.7% for the uncued imagination, and 66.7% for the standalone experiment. Comparing these accuracies to the results during music perception, the latter significantly outperforms the accuracies obtained for the music imagination experiments (Wilcoxon signed rank test, $p < 0.016$), except for the standalone imagination experiment ($p = 0.23$), similar to what was observed in the two-song classification task. Furthermore, a significant difference between the cued and standalone imagination trial is found.

Figure 5a shows the confusion matrix of the 12-song classification task during the standalone imagination experiment. Looking at Table 1, even highly similar stimuli, with (almost) equal tempos are significantly distinguished. This indicates that the classification is based on more than only the beat tempo. Furthermore, it seems that there are no more errors between the same songs with and without lyrics (and thus with the same tempo) compared to other songs.

Song category classifier

In Figure 4c, the accuracy values of the song category classifier per subject are displayed, which are significant for all imagination conditions (McNemar test per subject, cue clicks: $p \ll 10^{-3}$, no cue clicks: $p < 0.01$, standalone: $p \ll 10^{-4}$). The median accuracy over all subjects is 73.3% for the cued imagination, 71.7% for the uncued imagination, and 75.8% for the standalone experiment. Comparing these accuracies to the results during music perception, the latter significantly outperforms all music imagination experiments (Wilcoxon signed rank test, $p < 0.04$). Between the imagination experiments, no significant differences were found.

This classification task shows that a division according to song category can be made using brain responses. Figure 5b shows the confusion matrix for the song category classifier during the standalone imagination experiment. Similar to the



(a) 12-song classifier

(b) Song category classifier

Figure 5. Music imagination: Confusion matrices for the 12-song classifier and the song category classifier during the standalone imagination experiment. Darker colored dots display a dense amount of classified stimuli (perfect classification = 1), while lighter colors represent a sparse amount of classified stimuli. We see a good overall classification, even for the same songs with/without lyrics or with similar tempos (Table 1).

Category	Stimulus ID (Tempo [BPM])			
Lyrics	S1 (212)	S2 (189)	S3 (200)	S4 (160)
Removed lyrics	S5 (212)	S6 (189)	S7 (200)	S8 (160)
Instrumental	S9 (178)	S10 (166)	S11 (104)	S12 (140)

Table 1. Music stimulus properties (OpenMIIR data set).

12-song classification task, these results show that even for highly similar stimuli (nursery rhymes with and without lyrics) that have exactly the same tempo (Table 1), the song classifier can determine the right category based on brain responses.

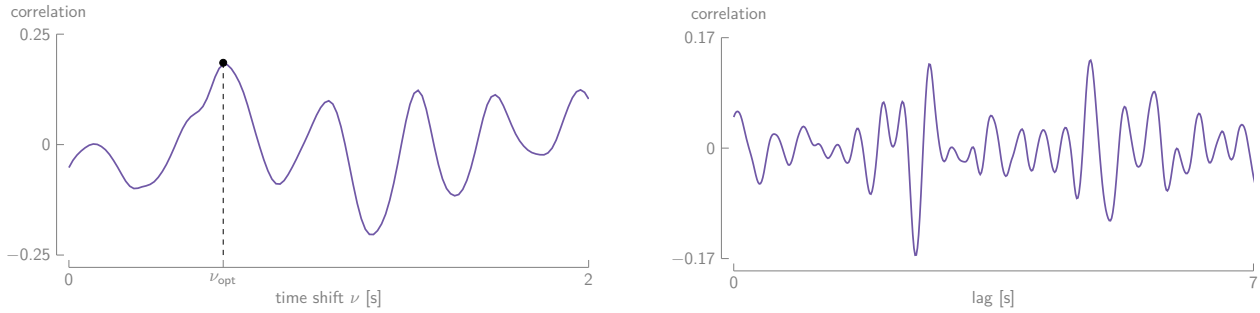
Periodicity

In this section, we will investigate the Pearson correlation coefficient $\rho(s(t), \hat{s}(t+\nu))$ in function of the time shift ν. When plotting these Pearson correlations as a function of ν per trial, a seemingly periodical or oscillatory behavior arises, as shown in Figure 6a. This suggests that the neural response has a periodical behavior, possibly following the tempo of the music (typically measured in beats per minute or BPM). However, it is possible that this periodical behavior of the Pearson correlation coefficient is an artifact induced by the periodical nature of the original stimulus, as the latter is also included in the computation of $\rho(s(t), \hat{s}(t+\nu))$.

To investigate this phenomenon of beat or tempo tracking, the leakage effect from the periodicity of the real envelopes must first be ruled out. We do this by testing whether this periodical behavior also appears in the reconstructed envelope, which is exclusively constructed from the EEG data. If a similar oscillatory behavior is present, this would mean that the human brain also registers the tempo (BPM) in the neural processing pathway. To this end, shorter excerpts of the real audio envelopes, equal to the length of two song periods, are cut out and cross-correlated with the corresponding decoded envelopes. If a periodical pattern is observed for more than two song periods in this cross-correlation, this behavior is not only caused by the real envelope, but by a combined oscillatory behavior of both the neural response and the original music envelope.

Figure 6b shows the cross-correlation between the decoded envelope and the two-period excerpt for the same two imagined stimuli, in which we can see a similar pattern of minima and maxima extending over more than two stimulus periods. This observation confirms that the exhibited periodical pattern is also present in the neural activity generated by the imagined music. To further investigate this beat tracking phenomenon, the median distance between two peaks in these cross-correlations is extracted, leading to an estimated beat frequency per trial. The median beat frequency over all trials and subjects is subsequently correlated with the real tempi of the stimuli. Figure 7 shows that these tempo estimates are highly similar to the real tempi of the corresponding stimuli. Using a t-test, significant correlations with the real tempo are found for the three imagination experiments of the OpenMIIR data set (cue clicks: correlation = 0.62, $p = 0.032$, no cue clicks: correlation = 0.67, $p = 0.018$, standalone: correlation = 0.64, $p = 0.024$).

Previous research of Tal et al.[21] used an MEG experiment and drum patterns to uncover that this neural entrainment, in the case of music perception, is at least partially internally produced instead of stimulus-driven. A phase-locking to the beat is created in the brain even if some beats are missing. Finding this beat tracking pattern in the case of music imagination further broadens this observation to an entirely internally generated song.



(a) Correlation i.f.o. time shift.

(b) Cross-correlation reconstructions - two periods envelope.

Figure 6. (a) The correlation i.f.o. the time shift ν for a representative trial and subject of the OpenMIIR data set, for the standalone imagination experiment (no cue clicks). (b) The cross-correlation of the reconstruction with two periods of the envelopes, for the same representative trial, shows a similar periodical pattern.

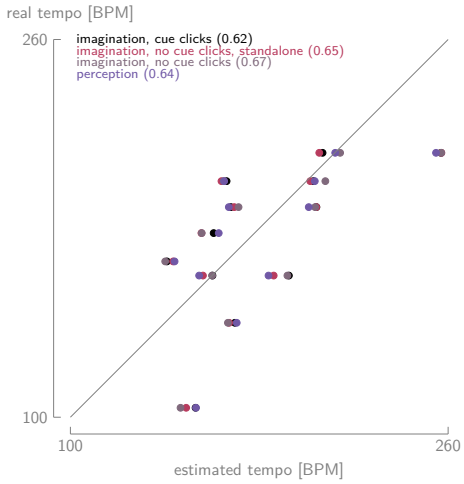


Figure 7. Plot of the real vs. estimated tempo for each of the twelve songs per perception/imagination experiment. One dot represents the tempo of one song in one of the four experiments. The black line depicts the bisecting diagonal (corresponding to $\rho = 1$) as a reference. When comparing the real and estimated tempos of all songs, we find that they are significantly correlated during both music perception and imagination.

Conclusions and outlook

In this study, we have demonstrated the possibility to decode and recognize imagined music. Our methodology employs a linear neural decoding model, as traditionally used for speech and perceived music[2,6,8,13,16,19]. We have shown that a similar linear neural decoding paradigm can also be used to decode imagined music if a time shift parameter ν is introduced, which controls for the unknown onset of the imagined song, and which can be automatically extracted from the data.

The three classification tasks - a two-song, K-song, and song category classifier - are based on the correlation between the real music envelopes and the EEG-based reconstructed versions obtained via the linear decoding model. While the standard decoding model (with $\nu = 0$) results in non-significant classification accuracies, the results improve drastically when using the time-shifted version of the model, resulting in significant classification accuracies. Surprisingly, we have demonstrated that all imagined music tracks in a data set can be distinguished from each other with a relatively high accuracy based on single-trial EEG data. This is also true for highly similar stimuli with (almost) equal tempos. Furthermore, even the same songs, once with and once without lyrics, could correctly be identified with a high accuracy.

Classification accuracies in the case of imagined music were observed to be slightly lower than for music perception on the same music stimuli (63.3%, 61.7%, and 66.7% vs. 69.2% respectively on the 12-song classification task). We hypothesize that this is caused by mismatches between the original music envelope and the imagined music, for example, due to tempo variation. This may be overcome by using other similarity metrics that are robust to such tempo variations such as dynamic time warping or more advanced methods. Classification accuracies could be further improved by using more advanced features. For example, since the estimated beat or tempo tracking in the neural responses are correlated to the tempo of the original song (see Figure 7), one could include the estimated tempo as an additional feature.

Moreover, investigating this beat tracking phenomenon teaches us more about the processing of auditory stimuli by the brain. In the case of music perception, previous literature shows that this neural entrainment is at least partially internally produced instead of stimulus-driven, thereby creating regular beat patterns in the brain even if some beats are missing in the perceived music[21]. Finding a beat pattern in the case of music imagination, as shown in this study, further broadens this observation to an entirely internally generated song.

However, a few limitations exist in this current approach. First, for each subject, the experiments were recorded in one continuous experiment. If this would not be the case, a retest would usually require calibration and retraining of the decoders due to a different fitting of the EEG cap. Moreover, 64 channels are used for the recordings in both data sets, which is too bulky for practical brain-computer interface (BCI) applications. Further research could investigate if similar results can be obtained using a smaller amount of EEG electrodes.

A better understanding of the process of music imagination and its decoding possibilities could increase our knowledge of the (musical) brain and its highly complex workings. Furthermore, with the upcoming interest in BCI technology, the results of this study open the door for several futuristic applications. For example, one can imagine a hands-free and silent BCI which allows one to select and play a song from a pre-defined playlist by simply thinking of it. Although the classification results obtained in this study are still far away from a practically viable accuracy, it provides an important initial step, which demonstrates that imagined music recognition is possible, and which paves the way for new research endeavors to further explore and improve this new paradigm.

References

1. Stober, S., Sternin, A., Owen, A. M. & 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*, 763–769.
2. Di Liberto, G., Pelofi, C., Shamma, S. & de Cheveigné, A. Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoust. Sci. Technol.* **41**, 361–364 (2020).
3. Sturm, I., Dahne, S., Blankertz, B. & Curio, G. Multi-Variate EEG Analysis as a Novel Tool to Examine Brain Responses to Naturalistic Music Stimuli. *Plos One* **10**, DOI: 10.1371/journal.pone.0141281 (2015).
4. Schaefer, R. S., Farquhar, J., Blokland, Y., Sadakata, M. & Desain, P. Name that tune: Decoding music from the listening brain. *Neuroimage* **56**, 843–849, DOI: 10.1016/j.neuroimage.2010.05.084 (2011).
5. Stober, S. Toward Studying Music Cognition with Information Retrieval Techniques: Lessons Learned from the OpenMIIR Initiative. *Front. Psychol.* **8**, DOI: 10.3389/fpsyg.2017.01255 (2017).
6. O'Sullivan, J. A. *et al.* Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cereb. Cortex* **25**, 1697–1706, DOI: 10.1093/cercor/bht355 (2015).
7. Fuglsang, S. A., Dau, T. & Hjortkjaer, J. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage* **156**, 435–444, DOI: 10.1016/j.neuroimage.2017.04.026 (2017).
8. Wong, D. D. E. *et al.* A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding. *Front. Neurosci.* **12:531**, DOI: 10.3389/fnins.2018.00531 (2018).
9. Das, N., Bertrand, A. & Francart, T. EEG-based auditory attention detection: boundary conditions for background noise and speaker positions. *J. Neural Eng.* **15**, 066017 (2018).
10. Geirnaert, S. *et al.* Neuro-Steered Hearing Devices: Decoding Auditory Attention From the Brain (2020). 2008.04569.
11. Gang, N., Kaneshiro, B., Berger, J. & 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*, 131–138.
12. Di Liberto, G. M. *et al.* Cortical encoding of melodic expectations in human temporal cortex. *eLife* **9** (2020).
13. Cantisani, G., Essid, S. & 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)*, vol. 2019-, 80–84 (IEEE, 2019).
14. Treder, M. S., Purwins, H., Miklody, D., Sturm, I. & Blankertz, B. Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *J. Neural Eng.* **11**, 10, DOI: 10.1088/1741-2560/11/2/026009 (2014).
15. Somers, B., Francart, T. & Bertrand, A. A generic EEG artifact removal algorithm based on the multi-channel Wiener filter. *J. Neural Eng.* **15**, DOI: 10.1088/1741-2552/aaac92 (2018).

16. Biesmans, W., Das, N., Francart, T. & Bertrand, A. Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 402–412, DOI: 10.1109/TNSRE.2016.2571900 (2017).

17. Geirnaert, S., Francart, T. & Bertrand, A. An Interpretable Performance Metric for Auditory Attention Decoding Algorithms in a Context of Neuro-Steered Gain Control. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28**, 307–317 (2020).

18. Verschueren, E., Somers, B. & Francart, T. Neural envelope tracking as a measure of speech understanding in cochlear implant users. *Hear. Res.* **373**, 23–31 (2019).

19. de Cheveigné, A. *et al.* Decoding the auditory brain with canonical component analysis. *Neuroimage* **172**, 206–216, DOI: 10.1016/j.neuroimage.2018.01.033 (2018).

20. Haufe, S. *et al.* On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage (Orlando, Fla.)* **87**, 96–110 (2014).

21. Tal, I. *et al.* Neural Entrainment to the Beat: The "Missing-Pulse" Phenomenon. *J. Neurosci.* **37**, 6331–6341 (2017).

Author contributions statement

M.T., S.G., and A.B. conceived the experiment(s), M.T. conducted the experiment(s), analyzed the results, and wrote the first manuscript draft. M.T. and S.G. produced the images of this manuscript. All authors reviewed and edited the manuscript.

Additional information

Funding This research is funded by an Aspirant Grant from the Research Foundation - Flanders (FWO) (for S. Geirnaert), the KU Leuven Special Research Fund C14/16/057, FWO project nr. G0A4918N, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802895 and grant agreement No 637424), and the Flemish Government (AI Research Program). The scientific responsibility is assumed by its authors. A. Bertrand and S. Geirnaert are also affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

Accession codes

Competing interests The authors declare no competing interests.

Figure and table legends

Figure 1. Music perception: Median correlations for each subject (light purple dots) and song in both data sets. Darker colored dots display the median over all subjects.

Figure 2. Music perception: Classification accuracy for each subject (light purple dots) during the two-song, K-song, and song category classification tasks. The darker colored dots represent the median accuracy over all subjects.

Figure 3. Music imagination:(a) Median correlations (across repeated trials) for each subject (light purple dots) and song during cued music imagination for the model with optimal time shift. Darker colored dots display the median over all subjects. (b) Decoder weight topology used for envelope reconstruction, for all 64 EEG channels of the OpenMIIR data set. The displayed weights are computed by averaging the absolute values of the subject-specific decoders across subjects and time lags τ.

Figure 4. Music imagination: Classification accuracy for each subject (light purple dots) during the two-song, 12-song, and song category classification tasks, compared to the perception case (red dots). The darker colored dots represent the median accuracy over all subjects. * refers to a significant difference on the 0.05-level (Wilcoxon signed rank test), ns to a non-significant difference.

Figure 5. Music imagination: Confusion matrices for the 12-song classifier and the song category classifier during the standalone imagination experiment. Darker colored dots display a dense amount of classified stimuli (perfect classification $= 1$), while lighter colors represent a sparse amount of classified stimuli. We see a good overall classification, even for the same songs with/without lyrics or with similar tempos (Table 1).

Table 1. Music stimulus properties (OpenMIIR data set).

Figure 6. (a) The correlation i.f.o. the time shift ν for a representative trial and subject of the OpenMIIR data set, for the standalone imagination experiment (no cue clicks). (b) The cross-correlation of the reconstruction with two periods of the envelopes, for the same representative trial, shows a similar periodical pattern.

Figure 7. Plot of the real vs. estimated tempo for each of the twelve songs per perception/imagination experiment. One dot represents the tempo of one song in one of the four experiments. The black line depicts the bisecting diagonal (corresponding to $\rho = 1$) as a reference. When comparing the real and estimated tempos of all songs, we find that they are significantly correlated during both music perception and imagination.