Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification

Treder, Matthias S.; Purwins, Hendrik; Miklody, Daniel; Sturm, Irene; Blankertz, Benjamin

Published in:
Journal of Neural Engineering

DOI (link to publication from Publisher):
10.1088/1741-2560/11/2/026009

Publication date:
2014

Document Version
Early version, also known as pre-print

Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification

MS Treder[1,2], H Purwins[1, 4], D Miklody[1,2], I Sturm[1,3], and B Blankertz[1,2]

[1] Neurotechnology Group, Technische Universität Berlin, Berlin, Germany
[2] Bernstein Focus: Neurotechnology, Berlin, Germany
[3] Berlin School of Mind and Brain, Berlin, Germany
[4] Sound and Music Computing Group, Department of Architecture, Design & Media Technology, Aalborg University Copenhagen

E-mail: matthias.treder@tu-berlin.de

Abstract. Objective: Polyphonic music (music consisting of several instruments playing in parallel) is an intuitive way of embedding multiple information streams. The different instruments in a musical piece form concurrent information streams that seamlessly integrate into a coherent and hedonistically appealing entity. Here, we explore polyphonic music as a novel stimulation approach for use in a brain-computer interface.
Approach: In a multi-streamed oddball experiment, we had participants shift selective attention to one out of three different instruments in music audio clips. Each instrument formed an oddball stream with its own specific standard stimuli (a repetitive musical pattern) and oddballs (deviating musical pattern).
Main results: Contrasting attended versus unattended instruments, ERP analysis shows subject- and instrument-specific responses including P300 and early auditory components. The attended instrument can be classified offline with a mean accuracy of 91% across 11 participants.
Significance: This is a proof of concept that attention paid to a particular instrument in polyphonic music can be inferred from ongoing EEG, a finding that is potentially relevant for both brain-computer interface and music research.

Keywords: brain-computer interface, EEG, auditory, music, attention, oddball paradigm, P300

Background

A popular approach to brain-computer interfacing (BCI) is the use of sensory stimulation. Typically, in order to make a selection the participant is required to attend to a particular sensory event within a stream of events. With reference to the oddball paradigm [1], this event is a rare *deviant* occuring among more frequent *standard* events. Attention to this event can modulate components of the event-related potential (ERP). The most prominent of these ERP responses is the P3 component [2], but a number of earlier components that are modulated by attention and that contribute to classification performance have been identified [3–6].

While the utility of many visual paradigms is to some extent restricted by their gaze dependence (cf. [7–9]), auditory and tactile BCIs are not only gaze-independent but even vision-independent by design, at least when it comes to stimulation. Particularly in the auditory domain, there have been successful approaches to develop spellers, that enable users to spell words by deploying attention to acoustic stimuli [10–17]. Since a substantial part of BCI research effort goes into the simultaneous increase of classification accuracy and speed, researchers typically resort to streams of isolated sensory stimuli, that have simple physical characteristics and sharp onset and offset. There has been research using spoken or sung syllables or even natural sounds, and it was shown that the stimuli are perceived as more pleasant and in some cases even lead to better classification perormance [13,14,18].

So far, music has been addressed in BCI-related research in two different scenarios. Several works performed a sonification, that is, rendering audible, of ongoing EEG by transforming it into acoustic signals. Following seminal work by Lucier (*Music for Solo Performer*), composed in 1965, other used real-time EEG analysis for production of music scores based on the EEG frequency spectrum [19,20] and real-time

mechanical control of musical instruments by means of affective mental states [21,22]. Furthermore, it has been shown, that individual pieces of music induce individual signatures in the EEG, and some of these characteristics are even preserved when music is only imagined [23–25].

Music has not been harnessed as a stimulation paradigm before, although it has several intriguing properties. First, in contrast to virtually all other non-natural auditory stimuli, it appears to have a special cognitive and emotional status. It has profound effects on the neural chemistry and its psychological effects include regulating mood, increasing joy and happiness, and enhancing attention and vigilance [26]. Second, in Western societies, the skills involved in music listening and, partly, music understanding are typically overlearnt. In other words, perceiving music is a natural and intuitive task. Third, music integrates several instruments into an aesthetic entity; listeners are able to follow individual instruments while being immersed in a holistic listening experience. In Western major-minor tonal music, repetition and variation of patterns are an essential part of the structure that plays with the listeners' expectations.

Taking advantage of that, the aim of this paper is to explore a multi-streamed musical oddball paradigm as a novel approach to brain-computer interfaces. Hill and Schölkopf [27] demonstrated that when participants are presented two concurrent streams of auditory beeps, each having its own standard and deviant stimuli, the deviant in the attended auditory stream can be detected on a single-trial basis. In a similar fashion, we embed three concurrent streams in form of musical instruments. Each instrument repeats a characteristic pattern (standard stimulus) that is varied infrequently (deviant stimulus) without violating the characteristics of the musical idiom. Note that, in contrast to a standard oddball experiment, the goal is not to have a classifier differentiate between a standard stimulus and a deviant stimulus; rather, the goal is to differentiate between deviants in the attended auditory stream (attended deviants) and deviants in the unattended auditory stream (unattended deviants). The paradigm is illustrated in Figure 1.

As a first attempt to shed some light on the relevant stimulus parameters, we tested two different kinds of musical pieces. One was designed to resemble 1980s synthesizer pop music. Bass, drums, and keyboard take a stereotypical musical role and depend on each other especially with respect to their metrical structure. This strong interdependence of the voices gives rise to the conjecture that these voices might fuse to a holistic percept, or Gestalt, making it more difficult to disentangle the individual instruments mentally. In contrast, the second musical piece using samples of acoustic instruments (double-bass, piano, flute) is designed to maximize the independence of the voices, by employing distinct timbre, register, spatial direction and different metrical structure. We hypothesize that *attended* musical deviants induce specific modulations of ERP components that are different from *unattended* deviants and that can be classified on a single-trial basis. We further hypothesize that the latter musical scenario, due to its independence of instruments, eases the deployment of attention on a particular instrument and hence aids classification performance.

Methods

Participants

Eleven participants (7 male, 4 female), aged 21–50 years (mean age 28), all but one right-handed, took part in the experiment. All were naive with respect to BCI research and they received money for their participation. Participants gave written consent and the study was performed in accordance with the Declaration of Helsinki.

Apparatus

EEG was recorded at 1000 Hz, using a Brain Products (Munich, Germany) actiCAP active electrode system with 64 electrodes. We used electrodes Fp1–2, AF3,4,7,8, Fz, F1–10, FCz, FC1–6, FT7,8, T7,8, Cz, C1–6, TP7,8, CPz, CP1–6, TP7,8, Pz, P1–10, POz, PO3,4,7,8, and Oz,1,2, placed according to the international 10-10 system. Active electrodes were referenced to left mastoid, using a forehead ground. All skin-electrode impedances were kept below 20 kΩ. The bandpass of the hardware filter was 0.016–250 Hz. Visual stimuli

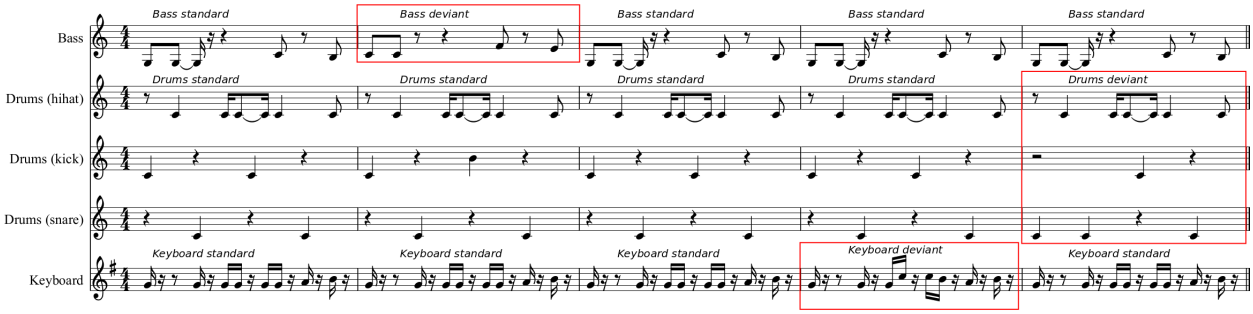

Figure 1. Score sheet illustrating the multi-streamed musical oddball paradigm for the Synth-Pop stimulus. There are 3 concurrent streams of stimuli, being represented by bass, drums (split into hi-hat, kick drum and snare), and keyboard. Each instrument has its own standard and deviant patterns. In the experiment, one of the three instruments would be attended while the other two are unattended, giving rise to attended deviants and unattended deviants.

were shown on a standard 22" TFT screen. Music stimuli were presented using Sennheiser PMX 200 headphones.

Stimuli

Stimuli consist of 40-seconds music clips. Each clip comprises 6 tracks, two of which are meant for stereo audio playback. The other four contain a mono mix of the stereo clip and three trigger channels (one for each instrument) that code the occurrence of standard and deviant stimuli. These four tracks were recorded as additional EEG channels. The stereo part of the clip is composed of three overlaid instruments, each playing frequent repetitions of a standard bar-long pattern, once in a while interrupted by a deviant bar-long pattern. Deviants of different instruments are non-overlapping, i.e. a deviant in one instrument is always accompanied by standard patterns in the other two instruments. Deviants are defined by a single tone or a whole sequence of tones deviating from the standard pattern. Each clip contains 3–7 deviants for each instrument. We tested two different kinds of music:

Synth-Pop A minimalistic adaptation of *Just can't get enough* by the Synth-Pop band *Depeche Mode*. A corresponding sample score is depicted in Figure 1. It features three instruments: drums consisting of kick drum, snare and hi-hat; a synthetic bass; and a keyboard equipped with a synthetic piano sound. The instruments play an adaptation of the chorus of the original song with the keyboard featuring the main melody of the song. Deviants are defined as follows: For the drums, the kick drum on the first quarter note is replaced by eigth notes featuring snare and then kick drum; for the bass, the whole 4-tone standard sequence is transposed up by 5 semitones; for the keyboard, tones 4–6 of the 8-tone standard sequence are transposed. The relative loudness of the instruments has been set by one of the authors such that all instruments are roughly equally audible.
Panning: None (all instruments panned to center).
Beats-per-minute: 130.

Jazz Stylistically, the *Jazz* clips are located half-way between a minimalistic piece by Philip Glass and a jazz trio comprising double-bass, piano, and flute. Each of the three voices is generated through frequent repetition of a standard pattern composed of 3-5 tones, once in a while replaced by a deviant pattern that differs from the standard pattern in one note. One clip consists of three overlaid voices. The *Jazz* music clips differ from the *Synth-Pop* clips in various ways. The *Jazz* clips sound more natural. This is achieved

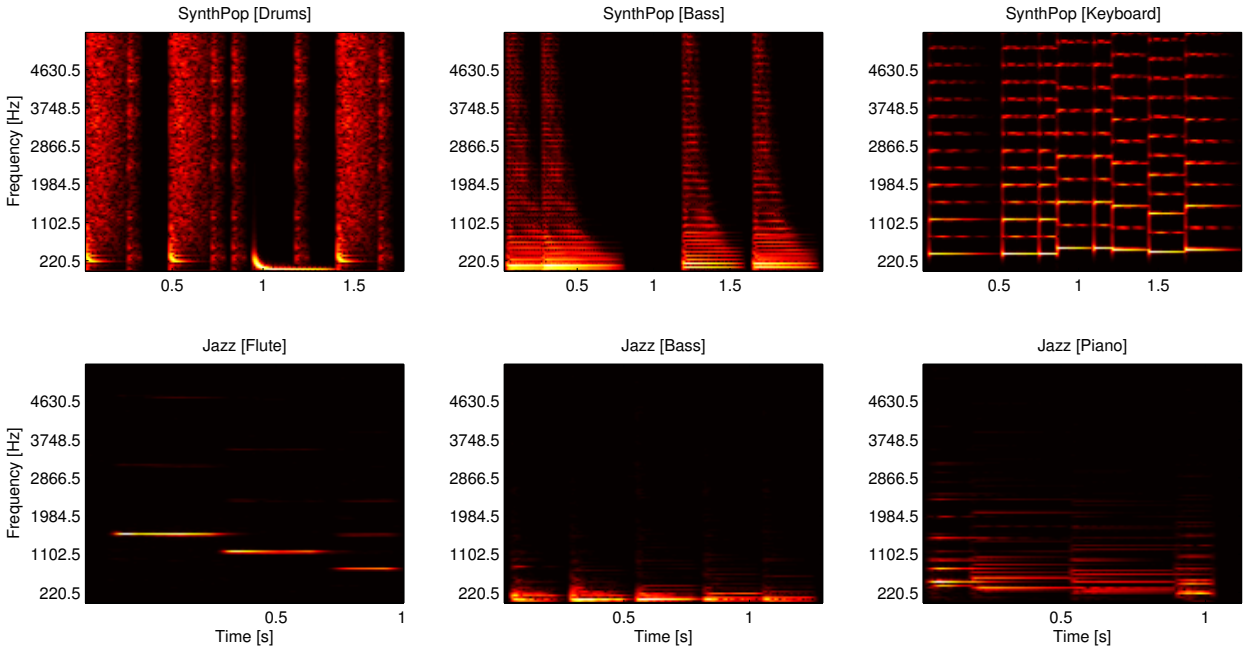

Figure 2. Log-amplitude spectrograms of the deviant stimuli for each instrument and each condition.

by selecting samples of acoustic instruments. In addition, loudness and micro-timing are manually adjusted for each tone of the basic pattern in order to make the entire phrase sound more musical. Apart from timbre (double-bass, piano, flute) and pitch range (low, medium, high), for the *Jazz* clips, another parameter is use to make the voices independent from each other. Each voice consists of patterns of different length, namely of 3, 4, and 5 beats per pattern. Through rhythmical interference a polymetrical rhythmical texture is generated. For better separation of the musical instruments, also panning is chosen to locate musical instruments in different directions from the user. This independence of the different voices is aimed at helping the user to focus on one particular instrument (stream segregation, [28]). The relative loudness of the instruments has been set by one of the authors such that deviants in all instruments are roughly equally audible. In particular, the double-bass had to be amplified, while the flute was turned down.
Panning: Flute left, Bass central, piano right.
Beats-per-minute: 120

For each music condition, 10 different music clips were created with variable amounts and different positions of the deviants in each instrument. Additionally, we exported solo versions with each of the instruments playing in isolation. Sample stimuli are provided as supplemental material. Figure 2 depicts spectrograms of the deviant stimuli for each of the instruments.

Procedure

Participants were seated in a comfortable chair at a distance of about 60 cm from the screen. Instruction was given both in written and verbal form. They were instructed to sit still, relax their muscles and try to minimize eye movements during the course of a trial. After EEG preparation, they first completed a short standard oddball experiment wherein they had to listen to a repeating standard tone that was replaced by a deviant tone of a different pitch with a probability of 15%.

Prior to the main experiment, participants were presented the different music stimuli and it was verified

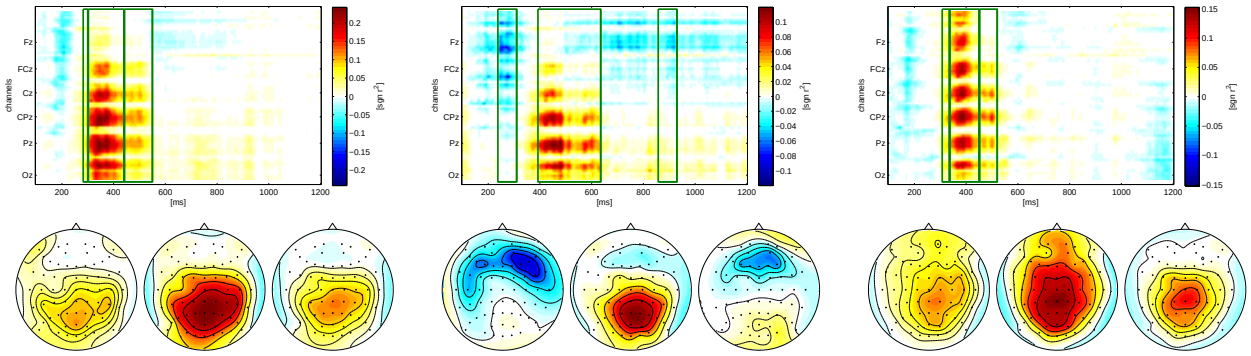

Figure 3. Feature selection for participant `aan` in the Synth-Pop condition, for each instrument separately. In each plot, the $sgnr^2$ between attended and nonattended stimuli (color coded) is plotted across time within epoch (x-axis) for each electrode (y-axis). The features (green boxes) were found heuristically by searching for peaks in the $sgn\, r^2$. The scalp plots below display the mean voltage distribution for the specific time windows. Due to the temporally extended nature of the oddballs discriminative infomation is sometimes at later time points than in typical oddball experiments.

that they can recognize the deviants. The main experiment was split into 10 blocks and each block consisted of 21 music clips. All clips in a block belonged to a single music condition: *Synth-Pop* (SP), *Jazz* (J), *Synth-Pop solo* (SPS), or *Jazz solo* (JS). The solo clips were identical to the mixed clips except for featuring only the cued instrument. The 21 music clips were played in random order. Each of the three instruments served as the cued instrument for 7 clips within a block. The music conditions were presented in an interleaved order as: SP, J, SPS, JS, SP, J, SPS, JS, SP, J. In other words, there were 3 mixed blocks (= 63 clips) and 2 solo blocks (= 42 clips) for each music condition.

Each trial started with a visual cue indicating the to-be-attended instrument. Then, the standard stimulus and the deviant stimulus of that particular instrument were played. Subsequently, a fixation cross was overlayed on the cue and after 2s, the music clip started. The cue and the fixation cross remained on the screen throughout the playback and participants were instructed to fixate the cross. To assure that participants deployed attention to the cued instrument, their task was to count the number of deviants in the cued instrument, ignoring the other two instruments. After the clip, a cue on the screen indicated that they should enter the count using the computer keyboard. After each block, they took a break of a few minutes.

Data analysis

For offline analysis, the data was downsampled to 250 Hz and lowpass filtered using a Chebyshev filter (with passbands and stopbands of 42 Hz and 49 Hz, respectively). The data was sectioned into epochs ranging from -200 ms prestimulus to 1200 ms poststimulus for each deviant. The prestimulus interval was used for baseline correction. A min-max criterion was used to reject artifacts (epochs in which the difference of maximum and minimum value exceeds 100 μV in one of the channels Fp1 or Fp2 are discarded). For classification, artifacts were only rejected for the training set and preserved in the test-set. Only deviants were subjected to analysis. They were assigned to one of two classes, namely *attended deviants* (i.e., deviants in the attended instrument) and *unattended deviants* (i.e., deviants occuring in any of the two unattended instruments).

Classification was based on two-class linear discriminant analysis (LDA) with shrinkage of the covariance matrix [29]. Linear classifiers are characterized by a projection vector $\boldsymbol{w}$ and a bias b, with the distance to the separating hyperplane given by $\boldsymbol{w}^\top \boldsymbol{x} - b$. In LDA, parameters are given as

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}\ (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1),$$
$$b = \boldsymbol{w}^\top (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2,$$

where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^N$ are the class means and $\boldsymbol{\Sigma}$ is the feature covariance matrix (averaged across the two classes, or pooled covariance). A given input $\boldsymbol{x} \in \mathbb{R}^N$ is assigned to one of the classes according to the sign of the distance to the hyperplane. Since the distributions parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$ are not known, they have to be estimated from collected calibration data. For high-dimensional features, the empirical covariance matrix tends to have a systematic bias. To counteract this detrimental effect on classification, we employed the shrinkage technique for the estimation of the covariance matrix [30], with the analytical solution to determine the shrinkage parameter as suggested in [31]; see [29,32] for an application of LDA-shrinkage in BCI context.

In our case, the classifier is used to discriminate *attended deviants* from *unattended deviants*. In other words, standard stimuli were not considered during classifier training or evaluation. We employed spatio-temporal features [29] for which all electrodes and three time intervals have been considered. The selection of these time intervals followed a heuristic searching for peaks in the point-biserial correlation coefficient $sgn\, r^2$ between attended deviants and nonattended deviants in the poststimulus interval (cf. [29]). Voltages were averaged within these three selected time intervals such that features of $3 \times 63 = 189$ dimensions were obtained. Averaging voltage corresponds to lowpass filtering of the signal, which owes to the fact that frequency peak of ERP components is typically in the theta or lower apha range; furthermore, averaging gives some robustness to the trial-to-trial variability of peak latencies. An example for feature selection is given in Figure 3. Classification performance was estimated using leave-one-clip-out cross-validation: the test-set comprised the data from a single music clip and the rest of the data was used for training; this procedure was repeated for each of the music clips.

For the investigation of the temporal and the spatial distribution of the discriminative information, also purely spatial (voltages at all channels averaged within a given time interval) and purely temporal features (voltages at all time points within an epoch for a given channel) have been used.

In order to take into account the fact that the physical differences between the different instruments could affect the shape of the ERPs, we compared two classification procedures:

(i) *General classifier.* A single binary classifier was trained using all attended deviants and all unattended deviants, discarding the instrument of origin. During testing, the instrument yielding the lowest mean classifier output was selected as attended instrument.

(ii) *Instrument-specific classifier based on posterior probabilities.* Attended and unattended deviants were split into 3 groups, one for each instrument. For each instrument, a separate binary classifier was trained. The training data was then split into two sets $\mathcal{T}_1$ and $\mathcal{T}_2$ according to class membership, obtaining two sets of projected data points $\{\boldsymbol{w}^\top \boldsymbol{x} + b \mid \boldsymbol{x} \in \mathcal{T}_k\}$. The two class-conditional distributions associated with these sets were modelled as Gaussians using maximum likelihood estimates of mean and variance. In the testing phase, Bayes formula was used to obtain posterior probabilities

$$P(\mathcal{C}_1 \,|\, \boldsymbol{x}) = \frac{P(\boldsymbol{x} \,|\, \mathcal{C}_1)\; P(\mathcal{C}_1)}{\sum_{k=1,2} P(\boldsymbol{x} \,|\, \mathcal{C}_k)\; P(\mathcal{C}_k)},$$

where $\mathcal{C}_1$ refers to the attended class and $\mathcal{C}_2$ refers to the unattended class, $P(\mathcal{C}_1 \,|\, \boldsymbol{x})$ is the posterior probability that the data $\boldsymbol{x}$ belongs to the attended class, $P(\boldsymbol{x} \,|\, \mathcal{C}_k)$ is the likelihood of the data, and $P(\mathcal{C}_1) = 1/3$ and $P(\mathcal{C}_2) = 2/3$ are the prior probabilities for an instrument being attended or unattended. For each clip, the instrument yielding the highest mean posterior probability on the deviants was selected as attended instrument.

Results

Event-related potentials (ERPs)

Grand average ERPs for each each music condition and each instrument are depicted in Figure 4. The grand average was calculated by weighting each participant's dataset according to the inverse of its variance. By this, noisy datasets were penalized and contributed less to the grand average waveform. In all cases, there is a difference between attended and unattended deviants. The peak difference is at about 500 ms

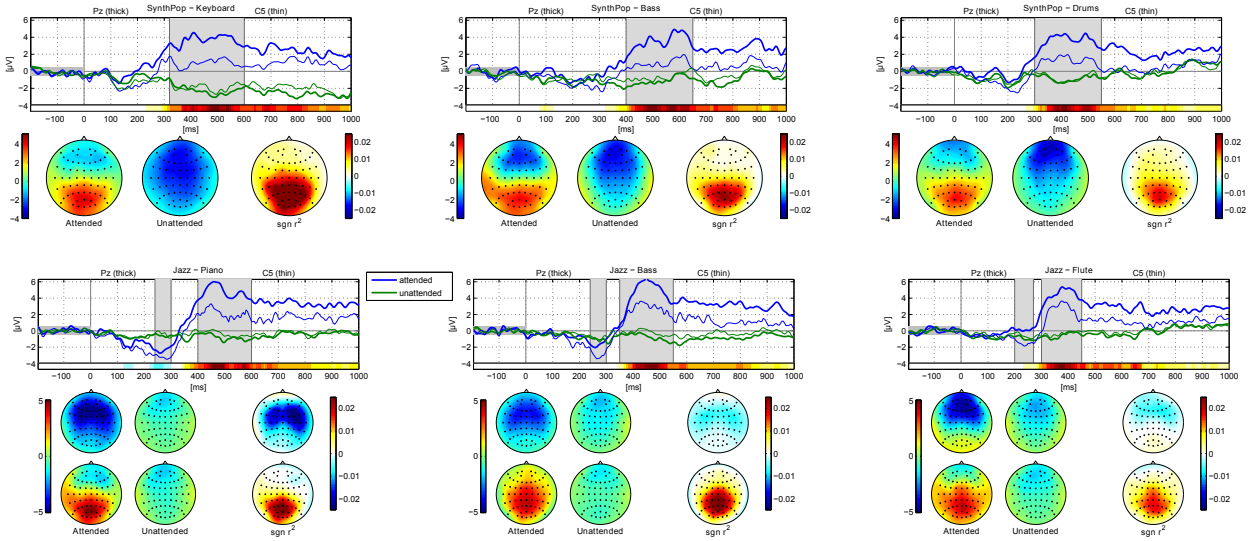

Figure 4. Grand average ERPs for each of the *Synth-Pop* condition (upper row) and the *Jazz* condition (lower row), separately for each instrument. Each channel plot shows attended deviants (blue lines) against unattended deviants (green line) for Pz (thick) and C5 (thin) electrodes. The horizontal colorbar at the bottom of the channel plot indicates *sgn* r^2 values for channel Pz. Below each channel plot, topographies are given for the grey shaded intervals. For the the *Synth-Pop* condition, there is a temporally extended component with a topography similar to P3. For the *Jazz* condition, there is additionally an earlier negativity that could be associated with auditory processing (bottom plots; upper row of topographies), followed by a P3 (bottom plots; lower row of topographies).

with a broad spatial topography typical for the P3 component. However, the difference persists throughout the whole epoch. In the *Jazz* condition, the P3 component is temporally more localized and there is also evidence of an earlier negativity at around 300 ms, particularly for the piano and the bass. This is most probably related to auditory processing of the deviating stimulus sequence.

Classification

R1.4

Classification results are depicted in Figure 5. Selection accuracy for selecting the correct instrument (chance level 33%) using the general classifier was 69.25±2.36% (where the error refers to standard error of the mean or SEM) for *Synth-Pop* and 71.47±3.33% for *Jazz*. With instrument-specific classifiers based on posterior probabilities, accuracy rose to 91±3.1% for *Synth-Pop* and 91.5±2.79% SEM for *Jazz*. A two-way repeated measures ANOVA with factors *Classification* [Binary, Posterior] and *Music* [*Synth-Pop*, *Jazz*] showed a significant effect of *Classification* ($F = 51.17, p < 0.001$). Classification was significantly better using instrument-specific classifiers than using a general classifier. There was no effect of Music ($p = 0.6447$) and no significant interaction ($p = 0.77$).

Spatial and temporal classification

Figure 6 shows the temporal and spatial distribution of information used during classification. For each kind of music and each instrument separately, a binary classifier was trained to discriminate between attended and nonattended deviants. Leave-one-clip-out cross-validation was used to obtain estimates of classification accuracy. To obtain the temporal distribution of information, the mean voltage in 100 ms windows centered in the interval [-500, 1000] relative to stimulus onset was used. Clearly, information is broadly distributed across the epoch with a peak at around 500 ms. In the *Jazz* condition, for instrument flute and keyboard, small early peaks suggest the involvement of an early auditory component. To obtain the spatial distribution of information, we trained classifiers on single electrodes, using all samples in the epoch as features. In

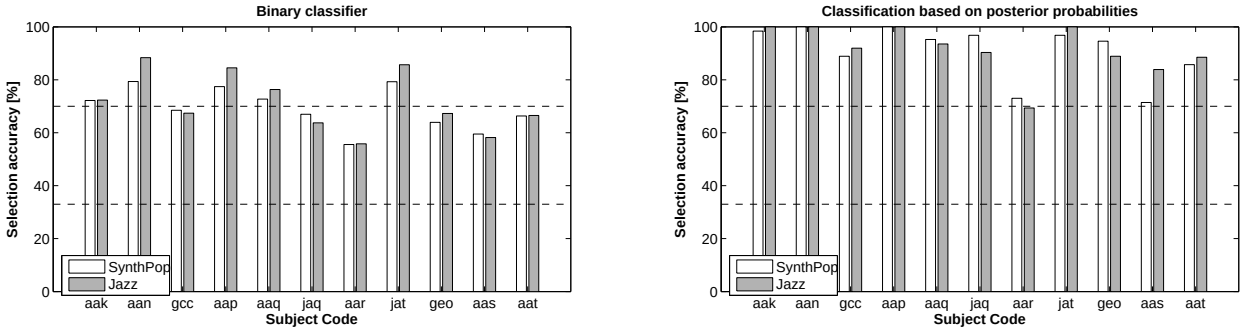

Figure 5. Classification performance for each participant. Dashed lines indicate chance level (33%) and the 70% benchmark for good BCI performance. Left: Selection accuracy using the general classifier. Right: Selection accuracy using the instrument-specific classifier based on posterior probabilities.

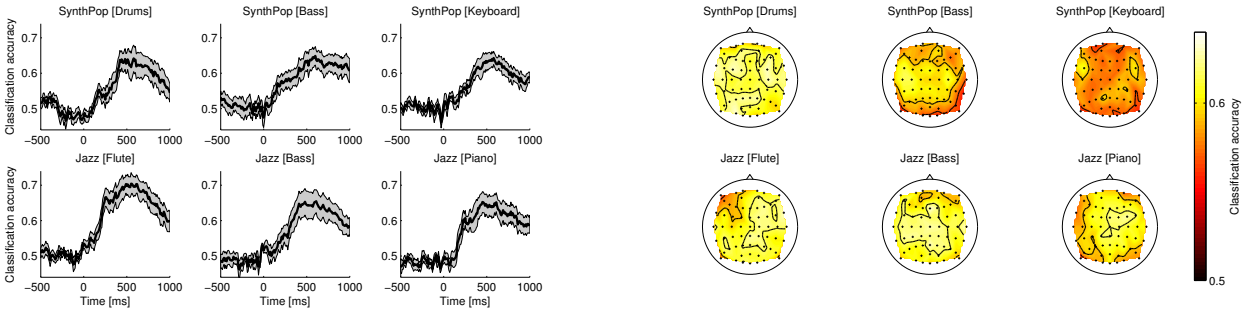

Figure 6. Temporal (left) and spatial (right) distribution of information for each kind of music and each instrument for binary classification (chance level 50%). The shaded areas in the left plots indicate 1 SEM across participants.

Figure 6 right, classification accuracy for across electrodes is depicted as scalp topographies. In most cases, classification performance is worst for pre-frontal and occipital electrode sites and best for central and/or temporal electrode sites.

Behavioral performance

For each condition and each instrument, we investigated participants' counting accuracy. The results are shown in Figure 7 left. A two-way repeated measures ANOVA with factors *Music* [Synth-Pop, Jazz] and *Instrument* [Drums/Flute, Bass, Keyboard/Piano] showed a significant effect of *Music* ($F = 3.78, p < 0.05$), with a higher accuracy for Synth-Pop. The *Music* $\times$ *Instrument* interaction was not significant ($p = 0.95$). The main effect of *Instrument* was not significant ($p = 0.12$), although t-tests with a Bonferroni-corrected criterion $\alpha = 0.05/3$ showed that counting accuracy was significantly lower for Bass than for Keyboard/Piano ($t = 4.87; p < 0.001$). No significant differences were found for Drums/Flute vs Bass ($p = 0.02$) and Drums/Flute vs Keyboard/Piano ($p = 0.10$).

We also investigated the relationship between counting performance and classification accuracy using instrument-specific classifiers. To this end, within music conditions, we averaged counting accuracies across instruments, and correlated the resulting statistic with classification performance using posterior probabilities. Results are shown in Figure 7 right. We found significant high correlations, both for *Synth-Pop* ($r = 0.87, p < 0.001$) and for *Jazz* ($r = 0.91, p < 0.001$).

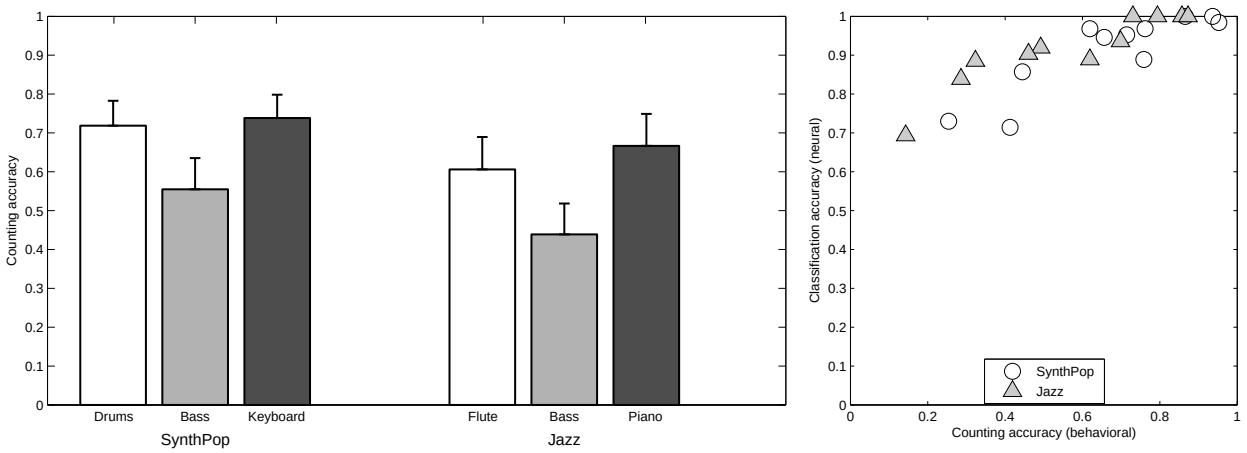

Figure 7. Behavioral data. Left: Behavioral performance of participants. Counting accuracy is shown for each kind of music and each instrument separately. Errorbars depict 1 SEM. Right: Correlation between behavioral performance (x-axis) and instrument selection accuracy (y-axis; chance level 33%) using posterior probabilities. Single data points represent single participants.

Discussion

Using a multi-streamed oddball paradigm with three concurrently playing instruments, we found that only deviants in the attended instrument produce a P300 while deviants in the unattended instrument do not. Furthermore, particularly in the *Jazz* condition, we found that auditory potentials following attended deviants are more pronounced than auditory potentials following unattended deviants in the same instrument.

Using a single binary classifier, the attended instrument can be predicted correctly with an accuracy of 69.3±2.4% in the *Synth-Pop* condition and 71.5±3.3% in the *Jazz* condition. Classification accuracy rises to 91±3.1% and 91.5±2.8%, respectively, using three classifiers and posterior probabilities. This suggests that there is substantial variability across instruments in terms of the temporal and spatial shape of the ERP. Furthermore, all but one participant (69.3% in the Jazz condition) exceed the 70%-benchmark that is generally considered as a threshold for acceptable BCI performance. Classification on spatial or temporal features alone showed a broad distribution of class-discriminative information, both spatially and temporally. The differences in spatial distribution of information could possibly stem from the different physical characteristics of the stimuli. Some instruments have short sounds with rather sharp on- and offsets, while others have more soft onsets, and are temporally more extended or consist of multiple deviant tones.

The behavioral analysis shows a far from perfect counting performance of the participants. Performance is worse in the *Jazz* condition, which can be explained by the fact that deviants consist of a single note whereas in the *Synth-Pop* condition deviants formed a sequence of several notes. It is hard to pinpoint the origin of these lapses during counting, since it can occur at two levels of cognitive processing. First, the lapse can occur at the perceptual level, with the participant simply not perceiving the deviant. Second, it can occur at a cognitive level, with the participant simply making an error during the counting task. In fact, some participants reported forgetting the exact count in a number of trials. The counting task can be said to implicitly induce a dual-task situation where the participant is required to not only attend to a particular instrument but also mentally add up the number of deviants, which involves simply arithmetics and memory. However, the significant correlation between counting performance and classification performance suggests that there is a relationship also at the level of perception and attention.

In the introduction, we conjectured that the relative independence of the instruments in the *Jazz* piece would ease the deployment of selective attention to a particular instrument. However, we found no effect in terms of classification accuracy and behaviorally, counting performance was even worse in the *Jazz* condition.

This might be due to several reasons. First, participants reported finding the *Synth-Pop* stimulus more pleasant, and this motivational effect might have counteracted beneficial perceptual effects. Second, the deviant was a sequence of notes for the *Synth-Pop* stimulus but only a single note for the *Jazz* stimulus, so that a *Synth-Pop* deviant could still be recognized even when a single note was missed. However, a more thorough analysis of the underlying factors requires separate parametric studies that contrast various parameters such as timbre, pitch, music genre and deviant length systematically.

Towards a musical BCI

The key point about a new BCI application is robust detection of the mental states of the user. Although we do not present a full-fledged BCI application, we demonstrate that the mental states (here: the attended instruments) can be detected with an accuracy of over 90% across 11 participants. To give an example, in a BCI equipped with with the musical oddball paradigm, each of the three instruments could be associated with a particular message to be conveyed, such as "YES", "NO", and "NOT SURE", or any other expedient set of messages. The user would then select one of these messages by attending to the corresponding instrument. In auditory BCI research, a transition from simple artificial tones (that were regarded as unpleasant or annoying) to sung syllables has been shown to increase users' ergonomics ratings [18].

Since our stimulus material is not only acoustically naturalistic, but even structurally close to original music, a similar effect can be expected. A direct comparison of our system with a standard auditory stimuli in terms of information throughput and ergonomics is a future task. Finally, since our approach successfully implemented two different musical styles, in the future it might be an option to design stimuli according to the user's preferred genre.

Despite the musical oddball paradigm not being competitive compared to state-of-the-art auditory BCIs in terms of information thoughput, our results show that it is possible to design a BCI that is linked to an important source of joy for individuals, namely music listening. Even though the number of possible concurrent streams is limited, within these limits the richer structure of the stimulus material might be beneficial for classification performance as suggested in [18]. Furthermore, the observation that complex naturalistic stimuli can reduce systematic class confusions might also apply to our stimulus material. The arbitrariness of the deviant patterns within the flow of music to some extent violates the musical structure, which, especially in Pop music, is simple, repetitive and therefore highly predictable. However, recent evidence demonstrates that switching from random sequences to fixed sequences does not hinder performance and that it can, on the contrary, even improve it [33]. Thus, it appears to be possible to make the musical oddball paradigm more musical by abandoning the classical oddball paradigm.

Single-trial classification for music research

Music perception has been investigated in several EEG studies. For instance, Besson et al. [34] demonstrated that a P300 component is evoked by sung melodies that are out of key. Granot et al. [35] reported a P300 component associated with expectation violation, using monophonic singing. However, previous research typically used monophonic musical pieces and based conclusions on group averages. For the first time, we study the oddball paradigm in a polyphonic musical context. Moreover, we are the first to use machine learning in decoding ERPs evoked by polyphonic music. In particular, we demonstrate that auditory attention to instruments in polyphonic music can be decoded on a single-trial basis. Although our present stimuli are far from, say, a concert hall-like music experience, the music clips that we use approach original music in instrumentation, style and structure. Within the domain of music perception research, brain responses to specific aspects of music typically are examined using well-controlled, artificial stimuli. Approaches using naturalistic stimuli have become more popular only recently, and they are impeded by the difficulty of having a small number of complex, unbalanced long stimuli [36,37]. Along this continuum of stimulus complexity, the present setting is at a half-way position between strict experimental control and ecological validity that, in principle, allows to investigate a range of aspects of music perception. More specifically, it opens an avenue for investigating the role of selective auditory attention in music, as it

potentially allows to further characterize attention-related features of the EEG, and eventually transfer this knowledge to a more complex musical context. For instance, one could use classification to monitor the moment-to-moment fluctuation of attention while listening to music and relate it to specific musical signatures. Insights into what captures a listeners attention may be relevant for direct creators of music, such as composers, but also in the domain of auditory interface design [38,39] and advertisement [40].

Limitations

A few limitations of the present study warrant consideration. First, although we ultimately pursue the implementation of a realistic music BCI, the musical pieces used do not yet correspond to real songs. To implement the oddball paradigm, the deviant is played at random points in time; to control the complexity of the stimulus, we restricted the number of instruments and the number of different musical patterns per instrument. These restrictions do not apply to many kinds of real music. However, as stated before, abandoning the random sequences of an oddball paradigm and turning to more musical, and hence more predictable, structures is viable possibility as suggested by recent evidence on BCI classification using fixed sequences [33]. Furthermore, the complexity of the musical piece (by increasing the number of instruments and/or number of patterns per instrument) can probably be further increased. To what extent this is possible has to be identified in future work.

Second, participants had an explicit counting task instead of simply attending. Since there was no online BCI feedback, the counting task was deemed necessary to ensure sustained attention thoughout the experiment. Future work should consider whether classification is possible without participants performing such an explicit task.

Third, the musical oddball paradigm was partly motivated by the ergonomic argument that music constitutes an aesthetically more pleasing stimulus than the sharp and abstract stimuli typically used. However, the truth of this statement is not verified yet. To this end, a comparative study needs to be conducted wherein the musical oddball paradigm is compared to an auditory BCI and users' ergonomy rating need to be registered.

Conclusions

The multi-streamed musical oddball paradigm exploits the fact that during listening to polyphonic music one is able to follow individual instruments while still being immersed in a holistic listening experience. Thus, our approach capitalizes on an overlearnt ability and simultaneously increases the usability of an auditory BCI providing the user with a more enjoyable and intuitive situation.

Additionally, music is an intuitive way of embedding parallel, though not fully independent, streams of information within a holistic percept or Gestalt. Our results show that it is possible to design a BCI that is linked to an important source of joy for individuals, namely music listening. Furthermore, this approach opens an avenue for investigating selective auditory attention to music and how it relates to stream segregation [28], supported by differentiating the streams with respect to timbre, pitch range, and rhythmical structure. Finally, this approach could be used to investigate which signatures in a complex musical score involuntarily call the attention of the listener.

Acknowledgements

We acknowledge financial support by the German Bundesministerium für Bildung und Forschung (Grant Nos. 16SV5839 and 16SV5839).

[1] N. K. Squires, K. C. Squires, and S. A. Hillyard, "Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man," *Electroencephalogr Clin Neurophysiol*, vol. 38, no. 4, pp. 387–401, Apr 1975. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/46819

[2] M. Fabiani, G. Gratton, D. Karis, and E. Donchin, "Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential," *Adv Psychophysiol*, vol. 2, pp. 1–78, 1987.

[3] L. Bianchi, S. Sami, A. Hillebrand, I. P. Fawcett, L. R. Quitadamo, and S. Seri, "Which physiological components are more suitable for visual ERP based brain-computer interface? A preliminary MEG/EEG study," *Brain Topogr*, vol. 23, pp. 180–185, Jun 2010.

[4] M. S. Treder and B. Blankertz, "(C)overt attention and visual speller design in an ERP-based brain-computer interface," *Behav Brain Funct*, vol. 6, p. 28, May 2010. [Online]. Available: http://www.behavioralandbrainfunctions.com/content/6/1/28

[5] P. Brunner, S. Joshi, S. Briskin, J. R. Wolpaw, H. Bischof, and G. Schalk, "Does the "P300" speller depend on eye gaze?" *J Neural Eng*, vol. 7, p. 056013, 2010.

[6] S. L. Shishkin, I. P. Ganin, I. A. Basyul, A. Y. Zhigalov, and A. Y. Kaplan, "N1 wave in the P300 BCI is not sensitive to the physical characteristics of stimuli," *J Integ Neuroscience*, vol. 8, no. 4, pp. 471–485, 2009.

[7] M. S. Treder, "Special section on gaze-independent brain-computer interfaces (editorial)," *J Neural Eng*, vol. 9, no. 4, p. 040201, 2012.

[8] M. S. Treder, N. M. Schmidt, and B. Blankertz, "Gaze-independent brain-computer interfaces based on covert attention and feature attention," *J Neural Eng*, vol. 8, no. 6, p. 066003, 2011, open Access. [Online]. Available: http://dx.doi.org/10.1088/1741-2560/8/6/066003

[9] A. Riccio, D. Mattia, L. Simione, M. Olivetti, and F. Cincotti, "Eye gaze independent brain computer interfaces for communication," *J Neural Eng*, vol. 9, p. 045001, 2012.

[10] M. Schreuder, B. Blankertz, and M. Tangermann, "A new auditory multi-class brain-computer interface paradigm: Spatial hearing as an informative cue," *PLoS ONE*, vol. 5, no. 4, p. e9813, 2010. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0009813

[11] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann, "A novel 9-class auditory ERP paradigm driving a predictive text entry system," *Front Neuroscience*, vol. 5, p. 99, 2011. [Online]. Available: http://www.frontiersin.org/Journal/Abstract.aspx?s=763&name=neuroprosthetics&ART_DOI=10.3389/fnins.2011.00099

[12] A. Furdea, S. Halder, D. J. Krusienski, D. Bross, F. Nijboer, N. Birbaumer, and A. Kübler, "An auditory oddball (P300) spelling system for brain-computer interfaces," *Psychophysiology*, vol. 46, pp. 617–625, 2009.

[13] J. Guo, S. Gao, and B. Hong, "An auditory brain-computer interface using active mental response," *IEEE Trans Neural Syst Rehabil Eng*, vol. 18, no. 3, pp. 230 –235, june 2010.

[14] A. Kübler, A. Furdea, S. Halder, E. M. Hammer, F. Nijboer, and B. Kotchoubey, "A brain-computer interface controlled auditory event-related potential (p300) spelling system for locked-in patients," *Annals of the New York Academy of Sciences*, vol. 1157, pp. 90–100, Mar 2009.

[15] M. Schreuder, T. Rost, and M. Tangermann, "Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI," *Front Neuroscience*, vol. 5, no. 112, 2011. [Online]. Available: https://www.frontiersin.org/Journal/Abstract.aspx?s=763&name=neuroprosthetics&ART_DOI=10.3389/fnins.2011.00112

[16] M. Schreuder, A. Riccio, F. Cincotti, M. Risetti, B. Blankertz, M. Tangermann, and D. Mattia, "Putting AMUSE to work: an end-user study," *Int J Bioelectromagnetism*, vol. 13, no. 3, pp. 139–140, 2011. [Online]. Available: http://ijbem.k.hosei.ac.jp/2006-/volume13/number3/2011_v13_no3_139-140.pdf

[17] N. Hill, T. Lal, M. Schröder, T. Hinterberger, N. Birbaumer, and B. Schölkopf, "Selective attention to auditory stimuli: A brain-computer interface paradigm," in *Proceedings of the 7th Tübingen Perception Conference*, H. Bülthoff, H. Mallot, R. Ulrich, and F. Wichmann, Eds. Kirchentellinsfurt, Germany: Knirsch Verlag, 2004, p. 102.

[18] J. Höhne, K. Krenzlin, S. Dähne, and M. Tangermann, "Natural stimuli improve auditory BCIs with respect to ergonomics and performance," *J Neural Eng*, vol. 9, no. 4, p. 045003, 2012. [Online]. Available: http://stacks.iop.org/1741-2552/9/i=4/a=045003

[19] E. R. Miranda, K. Sharman, K. Kilborn, and A. Duncan, "On harnessing the electroencephalogram for the musical braincap," *Comp Music J*, vol. 27, no. 2, pp. 80–102, 2003.

[20] E. R. Miranda and A. Brouse, "Interfacing the brain directly with musical systems: on developing systems for making music with brain signals," *Leonardo*, vol. 38, no. 4, pp. 331–336, 2005.

[21] S. Makeig, G. Leslie, T. Mullen, D. Sarma, N. Bigdely-Shamlo, and C. Kothe, "First demonstration of a musical emotion bci," in *Lecture Notes in Computer Science*, ser. Affective Computing and Intelligent Interaction, S. D. et al, Ed. Springer Berlin Heidelberg, 2011, vol. 6975, pp. 487–496.

[22] T. Mullen, R. Warp, and A. Jansch, "Minding the (transatlantic) gap: An internet-enabled acoustic brain-computer music interface," in *Proceedings of the International Conference on New Interfaces for Musical Expression, 30 May - 1 June 2011, Oslo, Norway*, vol. 2011, 2011, pp. 469–472.

[23] R. S. Schaefer, J. Farquhar, Y. Blokland, M. Sadakata, and P. Desain, "Name that tune: Decoding music from the listening brain," *Neuroimage*, Jun 2010.

[24] R. S. Schaefer, R. J. Vlek, and P. Desain, "Music perception and imagery in eeg: alpha band effects of task and stimulus. international," *Int J Psychophysiol*, vol. 82, no. 3, pp. 254–259, 2011.

[25] R. S. Schaefer, P. Desain, and J. Farquhar, "Shared processing of perception and imagery of music in decomposed eeg," *Neuroimage*, vol. 70, pp. 317–326, 2013.

[26] M. L. Chanda and D. J. Levitin, "The neurochemistry of music," *Trends Cogn Sci*, vol. 17, no. 4, pp. 179–193, 2013.

[27] N. Hill and B. Schölkopf, "An online brain–computer interface based on shifting attention to concurrent streams of auditory stimuli," *J Neural Eng*, vol. 9, no. 2, p. 026011, 2012. [Online]. Available: http://stacks.iop.org/1741-2552/9/i=2/a=026011

[28] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[29] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – a tutorial," *Neuroimage*, vol. 56, pp. 814–825, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2010.06.048
[30] J. H. Friedman, "Regularized discriminant analysis," *J Amer Statist Assoc*, vol. 84, no. 405, pp. 165–175, 1989.
[31] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J Multivar Anal*, vol. 88, pp. 365–411, 2004.
[32] C. Vidaurre, N. Krämer, B. Blankertz, and A. Schlögl, "Time domain parameters as a feature for eeg-based brain computer interfaces," *Neural Networks*, vol. 22, pp. 1313–1319, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2009.07.020
[33] M. Tangermann, J. Höhne, H. Stecher, and M. Schreuder, "No surprise — fixed sequence event-related potentials for brain-computer interfaces," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 2501–2504.
[34] M. Besson, F. Faïta, I. Peretz, A.-M. Bonnel, and J. Requin, "Singing in the brain: Independence of lyrics and tunes," *Psychological Science*, vol. 9, no. 6, pp. 494–498, 1998.
[35] R. Granot and E. Donchin, "Do re mi fa sol la ti-constraints, congruity, and musical training: An event-related brain potentials study of musical expectancies," *Music Perception*, vol. 19, no. 4, pp. 487–528, 2002.
[36] R. S. Schaefer, J. Farquhar, Y. Blokland, M. Sadakata, and P. Desain, "Name that tune: decoding music from the listening brain," *neuroimage*, vol. 56, no. 2, pp. 843–849, 2011.
[37] F. Cong, V. Alluri, A. K. Nandi, P. Toiviainen, R. Fa, B. Abu-Jamous, L. Gong, B. G. Craenen, H. Poikonen, M. Huotilainen, *et al.*, "Linking brain responses to naturalistic and continuous music through analysis of ongoing eeg and stimulus features," *IEEE Transactions on Multimedia*, 2013.
[38] B. D. Simpson, R. S. Bolia, and M. H. Draper, "Spatial audio display concepts supporting situation awareness for operators of unmanned aerial vehicles," *Human Performance, Situation Awareness, and Automation: Current Research and Trends HPSAA II, Volumes I and II*, vol. 2, p. 61, 2013.
[39] H. Gamper, C. Dicke, M. Billinghurst, and K. Puolamäki, "Sound sample detection and numerosity estimation using auditory display," *ACM Transactions on Applied Perception (TAP)*, vol. 10, no. 1, p. 4, 2013.
[40] C. Fraser and J. A. Bradford, "Music to your brain: Background music changes are processed first, reducing ad message recall," *Psychology & Marketing*, vol. 30, no. 1, pp. 62–75, 2013.