Extracting the neural representation of tone onsets for separate voices of ensemble music using multivariate EEG analysis.

Irene Sturm[1,2,4*], Matthias Treder[7,8], Daniel Miklody[2], Hendrik Purwins[2,6], Sven Dähne[3], Benjamin Blankertz[2,5,9], Gabriel Curio [1,4,5,9]

[1]Berlin School of Mind and Brain

[2]Technische Universität Berlin, Neurotechnology Group

[3]Technische Universität Berlin, Machine Learning Group

[4]Charité University Medicine Berlin, Department of Neurology, Neurophysics Group

[5]Bernstein Focus Neurotechnology, Berlin

[6]Audio Analysis Lab/Sound and Music Computing Group, Aalborg University Copenhagen

[7] Behavioural & Clinical Neuroscience Institute, Department of Psychiatry, University of Cambridge, UK

[8] Cambridge Centre for Ageing and Neuroscience (Cam-CAN), University of Cambridge and MRC Cognition and Brain Sciences Unit, Cambridge, UK

[9] Bernstein Center for Computational Neuroscience, Berlin

*corresponding author: irene.sturm@charite.de

Keywords: auditory stream segregation, music perception, naturalistic music, EEG, event-related potentials

Irene Sturm’s work was funded by the Berlin School of Mind and Brain and by the Christiane-Nüsslein-Volhard Foundation. This work was presented in parts as a poster at the Neurosciences and Music conference in Dijon 2014.

The work of MT, DM, HP and BB was supported by a grant of the BMBF under contract #01GQ0850 (BFNT).

Correspondence concerning this article should be addressed to Irene Sturm, Neurotechnology Group, Technische Universität Berlin, Sekr. MAR 3-4, Marchstr. 23, 10587 Berlin, Germany.
E-mail: irene.sturm@charite.de

Author biographies:

Irene Sturm holds degrees in music performance, as a music pedagogue and in computer science. Currently, she is a PhD student and scholarship-holder at the Excellence Graduate Berlin School of Mind and Brain. She is affiliated with the Neurotechnology Group at Technische Universität Berlin and with the Neurophysics Group at the Department of Neurology, Charité — University Medicine, Berlin. Her research interests comprise investigating music processing in naturalistic listening scenarios with EEG/ECoG.

Matthias Treder did his PhD on vision research at the Donder institute in Nijmegen, the Netherlands. From 2009-2014, he worked as a PostDoc in the Machine Learning Lab at TU Berlin, Germany, where he did research into visual brain-computer interfaces and single-trial EEG analysis. He is currently a PostDoc with the Department of Psychiatry, University of Cambridge, where he does connectivity and network analysis on ageing using MEG.

Daniel Miklody (MSc) holds a bachelor degree in Electrical Engineering and Information Technology from the Technical University of Vienna and a master’s degree in Computational Neuroscience from the Bernstein Center for Computational Neuroscience, Berlin. He is currently working as a research assistant in the Neurotechnology group of Professor Benjamin Blankertz at Technical University Berlin. His research focus is on individualized head models through electrical impedance measurements.

Hendrik Purwins is Assistant Professor at Audio Analysis Lab, Aalborg University Copenhagen, Denmark. His interests comprise statistical, unsupervised models for machine listening, music generation, sound resynthesis.

Sven Dähne obtained a B.Sc. in Cognitive Science in 2007 from the University of Osnabrück and a M.Sc. in Computational Neuroscience in 2010 from TU Berlin and the Bernstein Center for Computational Neuroscience (BCCN) in Berlin. He is currently working towards his Ph.D. at TU Berlin and the BCCN. He has been working on machine learning methods for online-adaptation of EEG based BCIs. His current research focuses on the multivariate analysis of bandpower modulations in the context of uni- and multimodal neuroimaging data.

Benjamin Blankertz holds the chair for Neurotechnology at Technische Universität Berlin and is one of the heads of the Berlin Brain Computer Interface project. His scientific interests comprise the application of multivariate decoding methods for neurocognitive and psychophysical experiments.

Gabriel Curio holds board specializations in neurology and psychiatry. He is leading the Neurophysics Group at the Department of Neurology, Charité — University Medicine, Berlin, Germany. Presently, he is a Professor of Neurology and Deputy Director of the Department of Neurology at the Charité Campus Benjamin Franklin. He is founding Co-Director of the Bernstein Center of Computational Neuroscience Berlin, lead PI in the Bernstein Focus Neurotechnology Berlin, and Faculty Member of the Excellence Graduate School of Mind and Brain.

Abstract

When listening to ensemble music even non-musicians can follow single instruments effortlessly. Electrophysiological indices for neural sensory encoding of separate streams have been described using oddball paradigms which utilize brain reactions to sound events that deviate from a repeating standard pattern. Obviously, these paradigms put constraints on the compositional complexity of the musical stimulus. Here, we apply a regression-based method of multivariate EEG analysis in order to reveal the neural encoding of separate voices of naturalistic ensemble music that is based on cortical responses to tone onsets, such as N1/P2 ERP components. Music clips (resembling minimalistic electro-pop) were presented to 11 subjects, either in an ensemble version (drums, bass, keyboard) or in the corresponding three solo versions. For each instrument we train a spatio-temporal regression filter that optimizes the correlation between EEG and a target function which represents the sequence of note onsets in the audio signal of the respective *solo* voice. This filter extracts an EEG projection that reflects the brain's reaction to note onsets with enhanced sensitivity. We apply these instrument-specific filters to 61-channel EEG recorded during the presentations of the *ensemble* version and assess by means of correlation measures how strongly the voice of each solo instrument is reflected in the EEG. Our results show that the reflection of the melody instrument keyboard in the EEG exceeds that of the other instruments by far, suggesting a high-voice superiority effect in the neural representation of note onsets. Moreover, the results indicated that focusing attention on a particular instrument can enhance this reflection. We conclude that the voice-

discriminating neural representation of tone onsets at the level of early auditory ERPs parallels the perceptual segregation of multi-voiced music.

Introduction

Natural 'soundscapes' of everyday life, e.g., communication in a crowded get-together or noisy environment, challenge our proficiency in organizing sounds into perceptually meaningful sequences. All the more music might spark our processing capabilities as it provides acoustic scenes with a large number of concurring sound sources. Yet, when listening to music we are able to organize the complex soundscape into streams, segregate foreground and background, recognize voices, melodies, patterns, motifs, and switch our attention between different aspects of a piece of music. Auditory stream segregation (ASS), the perceptional process which underlies this capability, has fascinated researchers for many years, resulting in numerous studies exploring its mechanisms and determinants. In a nutshell (for a detailed review see Moore and Gockel, 2002), the segregation of a complex audio signal into streams can occur on the basis of many different acoustic cues (Van Noorden, 1975); it is assumed to rely on processes at multiple levels of the auditory system; and it reflects a number of different processes, some of which are stimulus-driven while others are of more general cognitive nature, i.e., involving attention and/or knowledge (Bregman, 1994).

Electrophysiological indices of auditory stream segregation have been detected in several approaches (Sussman, 2005; Sussman, Horváth, Winkler, & Orr, 2007; Winkler, Takegata, &

Sussman, 2005; Yabe, et al., 2001; for an overview see Snyder and Alain, 2007). One line of research focused on the Mismatch Negativity (MMN) as neural index for a distinct perceptional state of stream segregation by constructing tone sequences such that only a perceptual segregation into two streams would allow a MMN-generating sound pattern to emerge. Following a similar principle, neural steady-state responses were found to reflect the formation of separate streams (Chakalov, Draganova, Wollbrink, Preissl, & Pantev, 2013) in MEG. Using EEG an influence of frequency separation of consecutive tones on the N1-P2 complex amplitudes was reported (Gutschalk, et al., 2005; Snyder, Alain, & Picton, 2006). Critically, this trend correlated with the perception of streaming in individual participants; a similar effect was reported for the N1 component.

This suggests that the amplitude of early auditory ERP components like the N1-P2 complex can inform about the perceptional state with respect to segregation/coherence of complex auditory stimuli. Since the N1-P2 complex as a sensory-obligatory auditory-evoked potential can be utilized without imposing a complex structure, e.g., an oddball paradigm, on the stimulus material, it may be promising for investigating ASS in more naturalistic listening scenarios.

In the domain of speech processing cortical onset responses that reflect changes in the waveform envelope (termed Envelope Following Responses, EFRs), have been a target of interest for a long time (Kuwada & Maher, 1986; Purcell, John, Schneider, & Picton, 2004; Aiken & Picton, 2005). Several approaches and methods aiming at extracting EFRs in naturalistic listening scenarios from continuous EEG or MEG have been proposed (Aiken & Picton, 2008; Kerlin & Miller, 2010; Lalor, Power, Reilly, & Foxe, 2009; Lalor & Foxe, 2010 and O'Sullivan, 2014). These methods have provided a distinct picture of the brain signals ‘following’ the speech waveform envelope and, in particular, been utilized to study the human ‘cocktail party problem’ of understanding speech in noisy settings. In the domain of music processing a marked reflection of the sound envelope has been detected in the EEG signal of short segments of naturalistic music (Schaefer, Farquhar, Blokland, Sadakata, & Desain, 2011). Unsupervised approaches (Cong, et al., 2012; Thompson, 2013) have confirmed that note onsets leave a reflection in the listener's EEG consistently across subjects and stimuli. However, these reflections have not been investigated in detail for longer musical contexts and, in particular, an analogue to the 'cocktail party' problem in speech processing has not been investigated specifically, even though composing music from several 'voices' is a common musical practice.

Considering the general characteristics of the N1-P2 response as a stimulus-driven sensory component that varies as a function of the physical properties of the sound like its frequency (Dimitrijevic, Michalewski, Zeng, Pratt, & Starr, 2008; Pratt, et al., 2009) or spectral complexity (Maiste & Picton, 1989; Shahin , Roberts, Pantev, Trainor, & Ross, 2005), it is an interesting question whether in a music-related scenario where perception of separate streams is highly likely, this typical onset-related ERP can be utilized to extract a neural representation related to these streams from the brain signal. In principle, this task taps into two so-called inverse problems that do not have a unique solution: (1) We have a number of sound sources that produce a mixed audio signal, and from the mixed signal it is not possible (without further assumptions) to infer the original configuration of sources. This audio signal is assumed to result in stimulus-related neural activity in the listener. (2) What we record in the listener's EEG is a mixture of stimulus-related neural activity, unrelated neural activity, and non-cerebral noise. Inferring these sources from the EEG signal, the so-called inverse problem of EEG generation, is likewise a problem without unique solution. In the present analysis we aim in a first step to learn a solution for the second of these inverse problems, to extract stimulus-related activity from the EEG in the case of a solo stream. Subsequently, we apply the derived solution in scenario with mixed sound sources. We explore in how far the stimulus-related activity related to the solo stream can be extracted from the EEG of the mixed (multi-voiced) ensemble presentation.

We re-analyze a data set from a study proposing a ‘musical’ brain computer interface application (Treder, Purwins, Miklody, Sturm, & Blankertz, 2014) where participants listened to short clips of a complex semi-naturalistic, multi-voiced music stimulus. In the music clips of 40 s duration three musical instruments (drums, keyboard, and bass) were presented, each playing a (different) sequence of a repetitive standard pattern, interspersed by an infrequent deviant pattern. Playing as an ensemble, the instruments produced a sequence resembling a minimalistic version of Depeche Mode’s ‘Just can’t get enough’ (1980s Electro Pop). The experiment consisted of 63 presentations of the ensemble version in which the instruments played together and 14 solo clip presentations for each instrument (42 solo clips in total). During the ensemble presentations participants were instructed to attend to a target instrument and to silently count the number of deviant patterns in this instrument. The original analysis showed that P3 ERP components to deviant patterns in the target instrument sufficiently differ from those in the non-target instruments and, thus, allow to decode from the EEG signal which of the instruments a subject is attending to. These results can be considered as a proof-of-concept that our capability of shifting attention to one voice in an ensemble may be exploited in order to create a novel music-affine stimulation approach for use in a brain-computer interface.

In contrast to the previous analysis that focused solely on P3 responses to deviations in the patterns, here, we propose to exploit the fact that *all* note onsets in a music clip should evoke ERP responses. Therefore, the sequence of onset events that constitutes each instrument’s part should elicit a corresponding series of ERP events in the listener’s EEG. Since onset characteristics critically contribute to an instrument’s specific timbre (McAdams, 1995) and onset-triggered ERPs are known to be responsive to subtle spectral and temporal changes (Meyer, Baumann, & Jancke, 2006) it can be assumed that the properties of this ERP response might differ for musical instruments with different tone onset characteristics. We introduce a novel multivariate method to extract this sequence of ERPs from the single-trial EEG by training a spatio-temporal filter that optimizes the relation between the sequence of onsets in the solo audio signal and the concomitant EEG. We (1) explore whether such a spatio-temporal filter obtains EEG projections from the solo-instrument trials that are significantly correlated with the sequence of onsets of the respective solo music clip; and we (2) probe (by correlation measures) whether these filters trained on the solo trials can be used to reconstruct a representation of this solo voice from the EEG of participants listening to the ensemble version clips. Finally, we test whether the reconstruction quality increases if participants focus their attention on the respective instrument.

Methods

Participants

Eleven participants (7 male, 4 female), aged 21-50 years (mean age 28), all but one right-handed, were paid to take part in the experiment. 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, Pz, P1-10, POz, PO3,4,7,8, and Oz,1,2, placed according to the international 10-20 system. In addition to these 63 EEG channels one electrode was used to measure the electrooculgram (EOG). 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 at 0.016-250 Hz. Visual stimuli providing the cues related to the participant’s task (details see below) were shown on a standard 22" TFT screen. Music stimuli were presented using Sennheiser PMX 200 headphones. The audio signal was recorded as an additional EEG channel.

Stimuli

Stimuli consisted of 40-seconds music clips in 44.1 kHz mono WAV format, delivered binaurally, i.e., listeners were presented with the identical audio stream at each ear. The ensemble version clip is composed of three overlaid instruments, each repeating 21 times the respective bar-long standard sound pattern depicted in Figure 1. In the following, the term ‘single trial’ denotes a single presentation of one of these 40s-long music clips. Once in a while, instead of the bar-long standard pattern a deviant pattern occurs in one of the instruments. Each clip contains 3-7 deviant bar-long patterns (out of 21 bars) for each instrument. Deviants of different instruments are non-overlapping and there is only one deviant pattern per instrument. Deviant patterns are defined by 1 (drums), 4 (bass) or 3 (keyboard) tone(s) deviating from the standard pattern in pitch or timbre (drums), but not changing the onset pattern in time (for a detailed description see Treder, Purwins, Miklody, Sturm and Blankertz (2014)). The stimulus represents a minimalistic adaptation of the chorus of ‘Just can't get enough’ by the Synth-Pop band Depeche Mode. 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 playing the main melody of the song. The relative loudness of the instruments has been set by one of the authors such that all instruments are roughly equally audible. The tempo is 130 beats-per-minute.

These stimuli are multi-voiced in the sense that they represent a musical texture consisting of more than one voice, not in the sense of independent melody lines. This interdependence is also reflected in the correlation between the audio power slopes that is given in Table 4. The bar-long patterns consist of nine onsets for drums, four onsets for bass and eight onsets for keyboard. Drums and keyboard have one onset each that is not shared by one of the other instruments; all other onsets coincide for at least two instruments.

In the original experiment two different kinds of musical pieces were tested: in addition to the ‘Just can't get enough’ adaptation (music condition SP) a stimulus resembling a jazz-like minimalistic piece of music (music condition J) was presented. This jazz-like piece of music was in stereo format, i.e., left ear and right ear were stimulated with different streams. The present analysis focused on utilizing continuous onset-related brain responses for the investigation of stream segregation. Therefore, the jazz-like stereo stimulus which introduced additional spatial cues for stream segregation was not appropriate for the present analysis.

According to the pattern of standard and deviant, 10 different music clips were created with variable amounts and different positions of the deviants in each instrument. Additionally, solo versions with each of the instruments playing in isolation were generated. Sample stimuli are provided as supplemental material.

Procedure

Participants were seated in a comfortable chair at a distance of about 60 cm from the screen. Instruction was given in both, 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. Prior to the main experiment, participants were presented with the different music stimuli and it was verified that they can recognize the deviants. The main experiment was split into 10 blocks and each block consisted of 21 40s-long music clips (containing 21 bars each). All clips in a block featured one 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 one instrument. Within one block the 21 music clips were played according to a randomized playlist containing the ten clips that differed with respect to the position of deviant patterns. 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 blocks with ensemble presentations (= 63 clips, 21 for each target instrument) and 2 solo blocks (= 42 clips, 14 for each instrument) for each music condition; only conditions SP and SPS are part of the present analysis.

Each trial started with a visual cue indicating the to-be-attended instrument. Then, the standard bar-long pattern and the deviant bar-long pattern of that particular instrument were played. Subsequently, a fixation cross was overlaid 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 prompted participants to enter the count using the computer keyboard. After each block, they took a break of a few minutes.

Data Analysis

Pre-processing of EEG data

The EEG data was lowpass-filtered using a Chebyshev filter (with passbands and stopbands of 42 Hz and 49 Hz, respectively) and then downsampled to 100 Hz. Since electrodes F9 and F10 were not contained in the head model used in the later analysis (see below ‘Training of regression filters on solo clips’) they were not considered in the analysis. This left 61 EEG channels for analysis. In order to remove signal components of non-neural origin, such as eye artifacts, muscle artifacts or movement artifacts while preserving the overall temporal structure of clips we separated the 61-channel EEG data into independent components using the TDSEP algorithm (Temporal Decorrelation source SEParation, (Ziehe, Laskov, Nolte, & Müller, 2004)) . ICA components that were considered as purely or predominantly driven by artifacts based on visual inspection of power spectrum, time course and topography (see also McMenamin et al. (2010) and McMenamin, Shackman, Greischar and Davidson (2011)) were discarded and the remaining components were projected back into the original sensor space.

Pre-processing of audio wave files

For each music clip (solo and ensemble stimuli) we determined the slope of the audio power envelope. To this end, we first segmented the audio signal into 50% overlapping time windows of 50 ms width and then calculated the average power of each window. Subsequently, the resulting time course was smoothed using a Gaussian filter of three samples width and the first derivative was taken, yielding the power slope. Then, the extracted power slope was interpolated to match the sampling frequency of the EEG.

Linear Ridge Regression with temporal embedding

In order to extract a component from the ongoing EEG that reflects a brain response to the sequence of onsets of a music stimulus we apply Linear Ridge Regression (Hoerl, 1970). Regression-based techniques have been applied in the context of cortical speech envelope tracking before (O'Sullivan, 2014). The related Canonical Component Analysis has been applied in studies related to the perception of complex natural stimuli, e.g. for identifying common networks of activation in a group of participants who were presented with movie clips (Dmochowski, Sajda, Dias, & Parra, 2012; Gaebler, et al., 2014) or in subjects listening to narrations (Kuhlen, Allefeld, & Haynes, 2012). Here, we utilize Linear Ridge Regression in order to optimally extract ERP responses that are phase-locked to rapid intensity changes indicating tone onsets in the music stimulus from the listener’s EEG. We train regression models to optimize the correlation between a surrogate channel extracted from the 61-channel EEG of single subjects and the power slope of the audio signal, a feature that, according to our experience, represents best the intensity changes that are expected to trigger ERP responses. Since it is not clear by how much the EEG response lags behind the presented stimulus, we apply regression to temporally embedded EEG data, a technique that was proposed in (Bießmann, et al., 2010) in order to deal with couplings between signals with unknown delay: To the EEG data set X1,…,Xn additional dimensions that are copies of X, time-shifted by 1, . . . , 25 data points are added as 'artificial' channels. This allows to capture brain responses within a latency of 0 to 250 ms.

Figure 2 summarizes the workflow of the generic regression analysis that was performed on the solo stimuli.

Training of regression filters to EEG during presentation of solo clips

In the first stage of the analysis regression filters that maximize the correlation between EEG and audio power slope were determined for the solo clips of the three instruments for each subject separately. In a leave-one-clip-out cross-validation approach clips for each instrument were divided into training and test sets, so that each clip acted as the test set once while the remaining clips formed the training set. Regression filters were calculated on the training set and applied to the test clip resulting in one uni-dimensional EEG projection for each of the 14 music clips. The correlation coefficients of the 14 derived EEG projections for one instrument and the respective power were calculated in order to determine how well the extracted brain response reflects the onset sequence of the stimulus at the level of single subjects and single trials. In the following, we use the term 'reconstruction quality' if we refer to the correlation coefficient between EEG projections and audio power slope. Additionally, the correlation coefficient for the mean EEG projection and the audio power slope was determined for each subject and instrument, and the grand average across all subjects was calculated.

The resulting regression filters, matrices of the dimensionality 61 channels x 26 time lags can be translated into spatio-temporal patterns that indicate to which extent each sensor contributes to the optimal EEG projection at which time lag (Haufe, et al., 2014). This allows to examine how the information that is used to reconstruct the audio power slope is distributed in space and time (relative to the stimulus). An example of such a spatio-temporal pattern is given in Figure 5. In order to get a better neurophysiological understanding of our results, we decomposed these 61 x 26 dimensional matrices into spatial components using a least-squares source reconstruction approach, the MUltiple SIgnal Classification (‘MUSIC’) algorithm (Mosher & Leahy, 1998) and determined the corresponding time evolution for each component. This gives a set of scalp topographies (called spatial MUSIC components in the following) that contain a certain proportion of the spatial variance of a regression pattern and a corresponding set of time courses (called temporal MUSIC components in the following) that informs about their temporal distribution.

Application of regression filter to EEG during presentation of ensemble version

Then, we applied the regression filters derived in step 1 to the EEG responses of the ensemble version stimuli. This was done for each subject and each instrument separately, resulting in three uni-dimensional EEG projections for each ensemble version clip per subject. As before, these projections were averaged across the 63 ensemble version clips for each subject (separately for the instruments) as well as across all subjects.

Statistical analysis

It is important to recognize that both, the EEG signal and the audio power slopes, contain serial correlation, i.e., subsequent samples are not independent of each other. Thus, the assumptions that underlie the standard tests for significance of correlation are not satisfied. To obtain a significance measure that takes this into account we followed the approach proposed by Pyper and Peterman (1998) and determined for each correlation coefficient the effective degrees of freedom based on the cross-correlation between the two respective time courses. This value, which is an estimate of the number of independent samples in both signals, is then used to determine the significance of the correlation coefficient. In order to account for the repetitiveness of the music clips, we considered the cross-correlation for all possible time lags within a music clip, drastically reducing the effective degrees of freedom. The original and estimated effective degrees of freedom for the Grand Average correlation coefficients are given in Table 2 in the bottom line.

The correlation coefficients of the subject-individual mean EEG projections were corrected for multiple testing for N=11 subjects with a Bonferroni correction. Significance of correlation was determined to the level of alpha=0.05.

Results

Solo stimulus presentations

Figure 3 shows examples of the EEG projections that reconstruct the audio power slope; for illustration purposes these were collapsed across 11 subjects, 14 clips for each instrument and 21 bars in each clip. A comparison of the EEG-reconstructed power slope (grey line) with the audio power slope (black line) shows that onset events in the audio signal are accompanied by peaks in the brain signal. Furthermore, the brain signal contains additional peaks that occur in absence of a corresponding onset event in the audio power slope.

Table 1 gives the percentage of solo clips (14 for each instrument) in which the EEG-reconstructed power slope is significantly correlated with the audio power slope at the level of each individual clip. Note that this measure relates to the significance of single trial clips of 40 s duration and was derived without averaging of EEG data. Table 2 shows the magnitude of correlation of the *averaged* EEG-reconstructed power slopes (for the 14 solo presentations of each instrument) with the audio power slope for single subjects, revealing significant correlation in 7/11 subjects for *drums*, in 9/11 subjects for *bass*, and in 8/11 subjects for *keyboard*. The bottom line of Table 2 shows that taking the mean of all subject's EEG projections (Table 2, bottom line 'GA') produces time courses that are significantly correlated with the original audio power slopes for all three instruments with magnitude of correlation r=0.60 for *drums* (p=0.00014, effective degrees of freedom: 34), r=0.52 for *bass*(r=0.52, p=0.00011, effective degrees of freedom: 48) and r=0.54 for *keyboard* (p=0.0000004, effective degrees of freedom: 72). Note that the original number of degrees of freedom of 3968 was drastically reduced by Pyper et al.'s method (Pyper & Peterman, 1998) that was applied to account for serial correlation in both time courses. All power slopes in Figure 3 are scaled for illustrational purposes. The absolute values of the audio power slopes for the three instruments are depicted in Figure 4, indicating differences in amplitudes and rise times.

Decomposition of regression patterns

Figure 5 shows an example of the spatio-temporal patterns that were derived from regression filters of a representative subject. The spatio-temporal patterns matrices that are directly derived from the regression filters are shown in the top panel. They show the distribution of information that is used to optimally reconstruct the stimulus' power slope in time and sensor space with time lags from 0 to 250 ms in the abscissa and the EEG channels on the ordinate. Note that the x-axis in milliseconds carries a different meaning than in standard ERP analysis, since it denotes the time lag between stimulus and EEG signal. Decomposing the spatio-temporal patterns with the MUSIC algorithm (see section Methods) results in a fronto-central scalp topography, resembling the topography of the N1/P2 complex. This scalp pattern is consistent for the three instruments. Its evolution over time differs, showing a change from positive to negative weights with extrema at 40 ms and 210 ms time lag for *drums*, broadly spread negative weights between 0 ms and 220 ms for *bass*, and a time evolution with two distinct positive peaks at 50 ms and 150 ms for *keyboard*.

Ensemble version stimulus presentations

Applying the three regression filters (trained on the solo stimulus presentations for the three instruments) to the EEG of the ensemble version stimulus presentation extracts an EEG projection that is significantly correlated with the solo audio power slope of each instrument in 3/11 subjects for *drums*, in 2/11 subjects for *bass*, and in 9/11 subjects for *keyboard* (Table 3). In one of the subjects EEG projections significantly correlated with all three solo power slopes could be derived in parallel from the (same) EEG of the ensemble presentation, in 3/11 subjects the audio power slopes of two instruments in parallel, in 5/11 subjects for one instrument, and for 2/11 subjects for none of them. The EEG Grand Average (11 subjects, 63 EEG projections for each ensemble version clip each) is significantly correlated with the audio power slope of a solo instrument only for *keyboard* (r=0.45, p=0.001, effective degrees of freedom 88).

Specificity of reconstruction

Since the solo power slopes are correlated with each other to different degrees as well as with the audio power slope of the ensemble version stimulus (Table 4), there is no straightforward way to estimate whether the EEG projections extracted by the instrument-specific filters are indeed specific for the instrument. To learn about the specificity, we put forward the null hypothesis that the instrument-specific filter extracts a representation of *all* onsets of the ensemble version stimulus. We compare Fisher-z-transformed correlation coefficients between EEG projections derived by the instrument-specific filter and solo audio power slopes to those between the same EEG projections and ensemble version audio power slopes in a paired

Wilcoxon signed rank test. Figure 6 shows that for *keyboard* in all but one subject the EEG projection is more highly correlated with the *keyboard* audio power slope than with the ensemble version audio power slope, resulting in a significant difference between the distributions of correlation coefficients at group level (p=0.002). For *drums* and *bass* there were no significant differences.

Effect of attention

When listening to the 63 ensemble version clips subjects were instructed to focus on a specific instrument before each clip, resulting in 21 trials of an ‘attended condition’ and 42 trials with an ‘unattended condition’ for each instrument. We tested whether the correlation between the EEG-reconstructed instrument-specific audio power slope and the respective audio power slope significantly differed between these two conditions by performing a random partition test with 1000 iterations. For single subjects a significant increase in correlation was present for *drums* in one subject (S1), for *bass* in two subjects (S5, S11), and for *keyboard* in five subjects (S6, S7, S8, S9, and S10). Within the group of subjects a significant effect of attention was present for *keyboard* (p = 0.001).

Behavioral performance

The behavioral performance differs for the three instruments with highest counting accuracy for *keyboard* (Grand Average: 74% correctly counted deviant stimuli), second highest accuracy for drums (71%) and lowest for *bass* (54%). The previous analysis of this data set (Treder, Purwins, Miklody, Sturm, & Blankertz, 2014) reported the absence of a significant main effect of the category instrument on the counting accuracy (ANOVA, p=0.12), but found a significantly lower counting accuracy for *bass* than for *Keyboard* (Bonferroni-corrected t-test, t = 4.87; p = 0.001).

Discussion

The present study demonstrates that multichannel EEG recordings can reveal neural responses to acoustic onset patterns of a single voice embedded in an ensemble of musical instruments: To this end 11 subjects listened to a set of music clips where three instruments played short repetitive patterns, either in a solo version (three solo conditions) or together, forming a minimalistic electro pop-like sound pattern (multi-voiced `ensemble' condition). Methodologically, we found that Linear Ridge Regression with temporal embedding enables to extract neural responses to the tone onset structure of a continuous music stimulus. In a first step using the solo stimulus presentations, such an onset sequence was reconstructed from the group average of EEG projections of each of the three instruments; for each single subject it was recovered at least for one of the instruments, in 4/11 subjects for all three instruments. Topographically, the maps derived from the spatio-temporal regression filters resembled a N1-P2 complex, as, e.g., described in Shahin, Roberts, Pantev, Traino and Ross (2005), while their time evolution seem to be influenced by the stimulus properties of each instrument's part. In a second step, applying these instrument-specific regression filters to the EEG recorded during the ensemble version presentation successfully extracted onset representations of at least one instrument's solo voice in 9/11 single subjects, and in the Grand Average for the melody instrument *keyboard*. Third, in the melody instrument the reconstruction quality was found significantly enhanced when this instrument was the target of attention.

Note onsets in music are acoustic landmarks providing auditory cues that underlie the perception of more complex phenomena such as beat, rhythm, and meter (Cameron & Grahn, 2014). Event-related brain responses to these low-level constituents of rhythm have been studied in numerous contexts in the music domain (Meyer, Baumann, & Jancke, 2006; Schaefer, Desain, & Suppes, 2009; Shahin A. , Roberts, Pantev, Trainor, & Ross, 2005) and in the speech domain (Hertrich, Dietrich, Trouvain, Moos, & Ackermann, 2012). In order to detect differences between conditions in the ERP, applications typically rely on averaging techniques. Thus, they require a large number of presentations of the same stimulus, therefore constraining the stimulus material in duration and complexity.

In the first part of the present analysis we have demonstrated that the proposed regression method allows to robustly track the onset sequence of three monophonic complex music-like stimuli in the listener's EEG. This corresponds to results from the domain of speech processing where Envelope Following Responses (EFRs) have been extracted from continuous EEG and MEG by combining source reconstruction techniques based on explicit modeling of the N1-P2 complex with convolution models (Aiken & Picton, 2008), with spatial filtering methods (Kerlin & Miller, 2010) or by estimating the impulse response of the auditory system (Lalor, Power, Reilly, & Foxe, 2009; Lalor & Foxe, 2010).

In particular, the proposed method is related to the reverse correlation approach of O'Sullivan et al. (2014) since we regress EEG onto a sound envelope-related target function and operate on single trials. Our results demonstrate that such an approach can be successfully applied in a music-related context and, moreover, we extend O'Sullivan's technique by providing a way to transform the regression filters into a format that is neurophysiologically interpretable.

Our approach was successful in single subjects in a considerable proportion of presentations (music clips of 40 s duration (see Table 1)) without any averaging of EEG data. By following a cross-validation approach we demonstrated that this relationship between EEG and stimulus reflects genuine stimulus-related activity in the listener's EEG that generalizes across presentations of the same stimulus.

Compared with averaging techniques the proposed EEG decomposition approach allows to examine also non-repetitive stimuli that would lead to 'blurred' ERPs for single tones in the average. It extracts an EEG projection that represents the cortical onset responses with enhanced signal-to-noise-ratio at the original time-resolution and, thus, enhances the sensitivity for small-scale differences between conditions such as, e.g., those related to the target status of an auditory stream. Furthermore, it allows for subsequent investigations at several time scales.

Extending the results by Schaefer et al. (2011) and Cong et al. (2012) the present results add to the growing body of knowledge about how a naturalistic complex music signal is represented in the brain.

Patterns

The extracted MUSIC components (see Methods) revealed a scalp pattern that was consistent between subjects and instruments while time courses strongly varied between instruments. This common scalp pattern is reminiscent of a N1-P2 complex. The P1-N1-P2 complex is a sequence of 'obligatory' auditory event-related potentials that index detection of the onset of auditory stimuli (Näätänen & Picton, 1987). Latency and amplitude of the P1, N1 and P2 (which are assumed to reflect different neural generators and functional processes, but typically occur together) are influenced by a variety of factors related to stimulus properties and context, but also to subject-individual variables, such as age, arousal or attention (for a review see Crowley and Colrain (2004)). Taken together, given the N1-P2-like scalp topography in the present results, the latency range of up to 250 ms, and the fact that the target function for defining the spatio-temporal regression filter emphasized rapid changes in sound intensity, the regression-derived EEG-projections appear to reflect a sequence of onset-triggered early auditory ERPs, similar to those reported for single musical tones (Shahin, Roberts, Pantev, Trainor, & Ross, 2005).

The temporal dimension of the extracted components of the three instruments is much more variable. When interpreting these time courses, one has to recognize that they differ from averaged ERPs (even though they are on the same time scale), as they represent the weighting of the corresponding spatial component over time and, thus, rather resemble a convolution model or FIR filter than an ERP time course. Nonetheless, time lags with large weights in principle can be compared to latencies of canonical ERP components. As such, the range where the extracted time courses peak is in line with the optimal time lag of cross-correlation between brain signal and sound envelope of 180 ms reported in (Aiken & Picton, 2005) and with results of O'Sullivan (2014). In the present stimuli, however, note onsets occur in quick succession, such that the window of 0 to 250 ms time lag of the regression model potentially covers more than a single onset/ERP component. This means that the regression model not only might 'learn' latency and spatial distribution of onset-related brain responses, but could be sensitive also to the rhythmic structure of the stimulus sequence. Most likely, the two peaks that are 115 ms apart (corresponding to the inter-onset-interval between two semi-quavers) in the temporal MUSIC component of *keyboard* can be attributed to this effect. Along this line, the flat shape of the temporal MUSIC component for *bass* may be related to the fact that its rhythmic pattern is the most inhomogeneous with respect to inter-onset-intervals and, the (relatively) better pronounced peaks of *drums* to quavers being the most frequent inter-onset-interval in this voice. In summary, while the spatial patterns are consistent across instruments, the extracted time courses seem to be influenced by stimulus properties. However, a future systematic parametric investigation is needed to clarify factors determining such instrument-specific time courses.

Ensemble version stimuli

In the second part of the analysis the regression filters that were fine-tuned to each subject's individual brain response and each stimulus' properties were applied to the subject's EEG recorded during the ensemble presentation. We assessed how well the solo parts of the three instruments were recovered by comparing the instrument-specific EEG projections to the respective audio power slopes. Our results show that at the level of single subjects this approach was successful for *keyboard* in all but two subjects, while a reconstruction for *drums* and *bass* failed in most subjects. In one subject (S1) all three instruments were reconstructed in parallel (from the same EEG signal) with significant correlation and in three subjects in two instruments.

The study goal was to approach the two-fold inverse problem of reconstructing (known) sound sources that create a mixed sound signal from the EEG signal of an individual who listened to this mixed signal. This intricate enterprise capitalized on the assumption that the brain performs auditory scene analysis and creates a representation of these single sources. In the present scenario the listener was presented with a sound scene that is stylistically relatively close to real music and, therefore, may invoke our natural abilities to stream music. The present stimulus provides a whole range of spectral, timbral and rhythmic cues on several time scales and these occur both, sequentially and simultaneously, promoting the segregation into streams. In the present scenario, thus, users were expected to perceive separate streams, and this assumption was confirmed by the behavioral results.

The present results are a proof-of-concept that a neural representation of such a stream can be extracted from the EEG, at least for one of the sound sources, here for the melody instrument *keyboard*. The scalp topographies derived from the regression models and the latency range of the EEG features suggest that the same ‘mid-latency’ auditory ERP components play a role in this process that have been found indicative of the percept of streaming, as reported previously in (Gutschalk, et al., 2005; Gutschalk, Oxenham, Micheyl, Wilson, & Melcher, 2007; Snyder, Alain, & Picton, 2006; Snyder & Alain, 2007; Weise, Bendixen, Müller, & Schröger, 2012). Furthermore, the corresponding instrument-specific time courses suggest that the temporal characteristics of ERP responses (latency, rise time) are critical for detecting the neural representation of distinct sound streams. Since we do not know whether a neural representation of distinct sound streams would be detectable in the case where subjects do *not* perceive separate streams, we cannot infer a causal relationship between the detectability of the neural representation and the percept of a stream. However, our approach prepares the ground for expanding the existing literature on EEG-correlates of auditory streaming with respect to more complex stimulus material.

Our results represent a link to the great number of studies that investigate the human 'cocktail party' problem (Power, 2012) by examining cortical activity that tracks the sound envelope of speech (for an overview see Ding (2014)) in multi-speaker environments.

These have demonstrated that Envelope-Following-Responses (EFRs) can be utilized to decompose the brain signal into representations of auditory streams. Moreover, selective attention leads to an enhanced representation in the attended stream while the to-be-ignored stream is suppressed (Kerlin & Miller, 2010). Several studies identified acoustic and higher-level influences on stream representation and associated time windows of processing (Ding & Simon, 2012; Ding & Simon, 2012b; Power, 2012; O'Sullivan, 2014; Horton, 2013). Our results contribute to this field in so far as they (at least partially) show a similar cortical representation of the single voices of a music-like stimulus. At group level the reconstruction quality of *keyboard*, the voice that is represented best, was significantly higher if *keyboard* was the target of attention. No such effect was present for *drums* and *bass* where reconstruction quality was poor. This means that we have found an analogue effect to an enhanced representation of an attended auditory stream in speech processing in the processing of a multi-voiced music-like stimulus. In particular, our results suggest that this effect is due to a synchronization of cortical activity to the rhythmic structure of the stimulus.

Critically, however, our stimulation scenario differs in some important points. In contrast to typical ‘cocktail party’ situations, the voices that constitute the present ensemble version stimulus are more strongly correlated and do not compete, but are integrated into an aesthetic entity. Furthermore, subjects were presented the same multi-voiced stream at both ears, while multi-speaker paradigms typically make use of a spatial separation of streams. Our results show that in absence of spatial cues and with a high coincidence of onsets between streams still at least two neural representations of streams could be extracted in parallel for some subjects. The time signatures that we derived from the regression filters suggest that such neural representations depend on differences in the shape of the time course of related ERPs.

Our results may contribute to the domain of auditory ERP-based BCI where early ERPs like the N1 and P2 have been exploited alone (Choi, 2013) or in combination with the P3 in order to decode the user’s target of attention from the EEG (Hill, Bishop, & Miller, 2012; Treder & Blankertz, 2010; Treder, Purwins, Miklody, Sturm, & Blankertz, 2014). In this context our results may give a first hint that such applications may in principle be designed without an oddball paradigm and based on more naturalistic stimuli.

The number of subjects with successfully recovered EEG-reconstructed solo power slopes differed for the three instruments, with *keyboard* outperforming *bass* and *drums* by far. In contrast, in the solo condition all instruments could be reconstructed similarly well, even though their audio power slopes differed in amplitude, rise times, and number of onsets. Therefore, it is not likely that the differences observed in the ensemble version condition reflect differences solely in the stimulus characteristics. It rather points to a strong influence of the context on the neural representation of the instruments’ parts, i.e., whether an instrument plays alone or is part of an ensemble. Our findings are in line with the high-voice superiority effect for pitch encoding that has been demonstrated by means of the Mismatch Negativity (MMN) in (Fujioka

T. T., 2005; Marie & Trainor, 2012; Marie & Trainor, 2014). In contrast, our results do not reveal a low-voice superiority effect that has been shown for timing in (Hove, 2014). This can be explained considering the two-tone masking effect (for a summary see Trainor L. J. (2015)): when a low-pitched and a high-pitched tone are presented together, the harmonics of the higher pitched tone tend to mask the harmonics of the lower pitched tone. In the present stimulus instruments play their notes mostly simultaneously. Consequently, the high-pitched keyboard masks the other instruments, while an opportunity for a low-voice superiority effect for timing to arise is not given, due to the absence of ‘unmasked’ bass tones.

The high-voice superiority effect is consistent with the musical practice of putting the melody line in the highest voice and has been supported by concomitant behavioral observations of superior pitch salience in the high voice (Crawley, 2002; Palmer, 1994). Our findings complement these results in so far as they indicate the N1-P2 as a further ERP component that reflects the high-voice superiority effect. Moreover, the present results demonstrate the presence of this effect in a more naturalistic listening scenario and, with *keyboard* being the instrument with the highest accuracy in the counting task, also find consistent behavioral evidence that agrees with previous results.

When evaluating correlation-related results in this scenario one has to keep in mind that the audio power slopes of all instruments and the ensemble version audio power slope are not independent of each other, but correlated to different degrees. This makes a comparison of correlation coefficients difficult; the periodic nature of the stimuli adds further limitations. Consequently, differences in absolute correlation coefficients are hard to interpret. Therefore, the present analysis was based on significance measures taking into account differences in the periodicity of the signals (see Methods). One possible concern is that the differences in reconstruction quality between *keyboard* and the other two solo instruments in the ensemble condition might just reflect the relations between the respective audio power slopes, more specifically, that the higher fidelity of the EEG-reconstructed *keyboard* slope is due to its relation to the ensemble version audio power slope. While such effects are inherent in this context and cannot be ruled out completely, two points argue in favor of a genuine instrument-specific EEG-based representation of the *keyboard*'s part in the ensemble condition: First, the correlation of the (original) slope of *drums* with the ensemble version slope is much higher than that of the (original) *keyboard* slope (see Table 3), but its reconstruction quality is poor in most subjects. Second, the EEG-reconstructed *keyboard* slope in all but one subjects is more similar to the original *keyboard* slope than to the ensemble version audio power slope (Figure 6), suggesting that this reconstruction indeed is specific for the *keyboard* part.

Limitations

The results presented here show that multivariate methods of EEG analysis can achieve considerable advances, on the one hand transferring previous results on the processing of tone onsets to more complex stimulation scenarios, on the other hand, dealing with complex challenges like the reconstruction of streams. Notwithstanding, several issues call for further exploration. First, the stimulus sequence contains infrequently occurring deviant sound patterns in each instrument's part. These trigger a P300 component which is the key EEG feature on in the operation of the original 'musical' BCI application. Yet, the present analysis uses only time lags between 0 and 250 ms and, consequently, should not make direct use of the 'strong' P300 component. Even though P3 to deviants may be picked up by our spatio-temporal filter, its reflection in the EEG projection will not be in 'sync' with the audio power slope and will rather lead to lower correlation with the power slope. However it cannot be completely ruled out that the processing of deviants influences also the earlier components. Since deviants occurred only infrequently, a possible influence would be ‘diluted’ strongly. Still, at this point, no strong claim can be made whether this approach can be transferred to a truly oddball-free, even more naturalistic paradigm and whether, in particular, the effect of attention is detectable in this case.

Even though the proposed method produces EEG-projections for single trials (given that training data of the same stimulus are available), a considerable part of the present effects was detected in averaged EEG projections. This means that, in a more general sense, the present approach can be regarded as an effective preprocessing step that exploits the wealth of the multivariate EEG in order to enhance the signal-to-noise-ratio and, thus, enables to extract stimulus-related activity from brain signals in far more complex stimulation scenarios. Moreover, the regression-derived patterns represent a kind of group average across the set of training data and, thus, cannot be regarded as single-trial results. In the present analysis the stimuli used for training the regression models were repetitions of one rhythmic pattern. This is not a prerequisite for applying Linear Ridge Regression, but most probably was beneficial for the ‘learning processes’ of the regression model. In principle, however, if an onset sequence has fairly stationary characteristics, e.g., timbre and attack, the brain response to these onsets should be extractable even in the absence of a strongly repetitive structure as in the present stimuli. This hypothesis could be addressed in future experiments.

Conclusion

The present results demonstrate that the sequence of note onsets forming a semi-natural rhythmically complex music stimulus can be reconstructed from the listener’s EEG using spatio-temporal regression filters. Furthermore, if the characteristics of a naturalistic complex sound pattern can be encoded by such a model, in principle this can be applied to extract an EEG representation of the respective sound pattern even if it is embedded into an ensemble of several voices. Thus, the EEG can provide a neural representation of separate streams a listener might perceive. Specifically, in congruence with behavioral results we found that the melody instrument of an ensemble music stimulus was represented most distinct and that focused attention enhanced this effect.

References

Aiken, S. J., & Picton, T. W. (2005). Envelope following responses to natural vowels. *Audiology & Neurootology, 11*(4), 213-232. DOI:10.1159/000092589

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing, 29*(2), 139-157. doi: 10.1097/AUD.0b013e31816453dc

Baumann, S., Meyer, M., & Jäncke, L. (2008). Enhancement of auditory-evoked potentials in musicians reflects an influence of expertise but not selective attention. *Journal of Cognitive Neuroscience, 20*(12), 2238-2249. doi:10.1162/jocn.2008.20157

Bießmann, F., Meinecke, F. C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N. K., & Müller, K.-R. (2010). Temporal kernel CCA and its application in multimodal neuronal data analysis. *Machine Learning, 79*(1-2), 5-27. doi:10.1007/s10994-009-5153-3

Billings, C. J., Tremblay, K. L., & Miller, C. W. (2011). Aided cortical auditory evoked potentials in response to changes in hearing aid gain. *International Journal of Audiology, 50*(7), 459-467. doi:http://dx.doi.org/10.3109%2F14992027.2011.568011

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound.* MIT press.

Cameron, D. J., & Grahn, J. A. (2014). Neuroscientific Investigations of Musical Rhythm. *Acoustics Australia, 42*(2), 111.

Campbell, J. D., Cardon, G., & Sharma, A. (2011). Clinical application of the P1 cortical auditory evoked potential biomarker in children with sensorineural hearing loss and auditory neuropathy spectrum disorder. *NIH Public Access*, *32*, S. 147. doi:http://dx.doi.org/10.1055%2Fs-0031-1277236

Chakalov, I., Draganova, R., Wollbrink, A., Preissl, H., & Pantev, C. (2013). Perceptual organization of auditory streaming-task relies on neural entrainment of the stimulus-presentation rate: MEG evidence. *BMC Neuroscience, 14*(1), 120. doi:10.1186/1471-2202-14-120

Choi, I. R.-C. (2013). Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in Human Neuroscience*, 7. doi:http://dx.doi.org/10.3389%2Ffnhum.2013.00115

Cirelli, L. K., Bosnyak, D., Manning, F. C., Spinelli, C., Marie, C., Fujioka, T., . . . Trainor, L. J. (2014). Beat-induced fluctuations in auditory cortical beta-band activity: using EEG to measure age-related changes. *Frontiers in Psychology, 5*, 1-9. doi:http://dx.doi.org/10.3389%2Ffpsyg.2014.00742

Coch, D., Sanders, L. D., & Neville, H. J. (2005). An event-related potential study of selective auditory attention in children and adults. *Journal of Cognitive Neuroscience, 17*(4), 605-622. doi:10.1162/0898929053467631

Cong, F., Phan, A. H., Zhao, Q., Nandi, A. K., Alluri, V., Toiviainen, P., . . . Ristaniemi, T. (2012). Analysis of ongoing EEG elicited by natural music stimuli using nonnegative tensor factorization. *Signal Processing Conference (EUSIPCO), 2012*, 494-498.

Crawley, E. J.-M. (2002). Change detection in multi-voice music: the role of musical structure, musical training, and task demands. *Journal of Experimental Psychology: Human Perception and Performance, 28*(2), 3. doi:http://psycnet.apa.org/doi/10.1037/0096-1523.28.2.367

Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: age, sleep and modality. *Clinical Neurophysiology, 115*(4), 732-744. doi:http://dx.doi.org/10.1016/j.clinph.2003.11.021

Dimitrijevic, A., Michalewski, H. J., Zeng, F.-G., Pratt, H., & Starr, A. (2008). Frequency changes in a continuous tone: auditory cortical potentials. *Clinical Neurophysiology, 119*(9), 2111-2124. doi:10.1016/j.clinph.2008.06.002

Ding, N. & Simon, J. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers . *Proceedings of the National Academy of Sciences, 109*(29), 11854-11859. doi:10.1073/pnas.1205381109

Ding, N. &. Simon, J. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology, 107*(1), 78-89. doi:10.1152/jn.00297.2011

Ding, N. &. Simon, J. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human Neuroscience*, 8. doi:http://dx.doi.org/10.3389%2Ffnhum.2014.00311

Dmochowski, J. P., Sajda, P., Dias, J., & Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention--a possible marker of engagement? *Frontiers in Human Neuroscience, 6*. doi:http://dx.doi.org/10.3389%2Ffnhum.2012.00112

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta--theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage, 85*, 761-768. doi:http://dx.doi.org/10.1016/j.neuroimage.2013.06.035

Fujioka, T. T. (2005). Automatic encoding of polyphonic melodies in musicians and nonmusicians. *Journal of Cognitive Neuroscience, 17*(10), 1578-1592. doi:10.1162/089892905774597263

Fujioka, T., Ross, B., Kakigi, R., Pantev, C., & Trainor, L. J. (2006). One year of musical training affects development of auditory cortical-evoked fields in young children. *Brain, 129*(10), 2593-2608. doi:http://dx.doi.org/10.1093/brain/awl247

Gaebler, M., Biessmann, F., Lamke, J.-P., Müller, K.-R., Walter, H., & Hetzer, S. (2014). Stereoscopic depth increases intersubject correlations of brain networks. *NeuroImage, 100*, 427-434. doi:10.1016/j.neuroimage.2014.06.008

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., & Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *The Journal of Neuroscience, 25*(22), 5382-5388. doi:10.1523/JNEUROSCI.0347-05.2005

Gutschalk, A., Oxenham, A. J., Micheyl, C., Wilson, E. C., & Melcher, J. R. (2007). Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation. *The Journal of Neuroscience, 27*(48), 13074-13081. doi:10.1523/JNEUROSCI.2299-07.2007

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Biessmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage, 87*, 96-110. doi:http://dx.doi.org/10.1016/j.neuroimage.2013.10.067

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., & Ackermann, H. (2012). Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology, 49*(3), 322-334. doi:10.1111/j.1469-8986.2011.01314.x

Hill, K. T., Bishop, C. W., & Miller, L. M. (2012). Auditory grouping mechanisms reflect a sound's relative position in a sequence. *Frontiers in Human Neuroscience, 6*. doi:http://dx.doi.org/10.3389%2Ffnhum.2012.00158

Hoerl, A. E. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67. DOI:10.1080/00401706.1970.10488634

Horton, C. D. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology, 109*(12), 3082-3093. doi:10.1152/jn.01026.2012

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*(28), 321-377.

Hove, M. J. (2014). Superior time perception for lower musical pitch explains why bass-ranged instruments lay down musical rhythms. *Proceedings of the National Academy of Sciences, 111*(28), 10383-10388. 10.1073/pnas.1402039111

Kerlin, J. R., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *The Journal of Neuroscience, 30*(2), 620-628. doi: 10.1523/JNEUROSCI.3631-09.2010

Kuhlen, A. K., Allefeld, C., & Haynes, J.-D. (2012). Content-specific coordination of listeners' to speakers' EEG during communication. *Frontiers in Human Neuroscience, 6*. doi:http://dx.doi.org/10.3389%2Ffnhum.2012.00266

Kuwada, S. B., & Maher, V. L. (1986). Scalp potentials of normal and hearing-impaired subjects in response to sinusoidally amplitude-modulated tones. *Hearing research, 21*(2), 179-192. http://dx.doi.org/10.1016/0378-5955(86)90038-9

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience, 31*(1), 189-193. doi:10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A., Reilly, R. B., & Foxe, J. J. (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *JN Physiology, 102*(1), 349-359. doi:10.1152/jn.90896.2008

Maiste, A., & Picton, T. (1989). Human auditory evoked potentials to frequency-modulated tones. *Ear and Hearing, 10*(3), 153-160.

Marie, C., & Trainor, L. J. (2012). Development of simultaneous pitch encoding: infants show a high voice superiority effect. *Cerebral Cortex*, S. bhs050. doi:10.1093/cercor/bhs050

Marie, C., & Trainor, L. J. (2014). Early development of polyphonic sound encoding and the high voice superiority effect. *Neuropsychologia, 57*, 50-58. doi:http://dx.doi.org/10.1016/j.neuropsychologia.2014.02.023

Martin, B. A., Tremblay, K. L., & Korczak, P. (2008). Speech evoked potentials: from the laboratory to the clinic. *Ear and Hearing, 29*(3), 285-313. doi:10.1097/AUD.0b013e3181662c0e

Martin, B., Tremblay, K., & Stapells, D. (2007). Principles and applications of cortical auditory evoked potentials. *Auditory evoked potentials: basic principles and clinical application*. 482-507. *ISBN*-13: 978-0781757560

McAdams, S. W. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes.,. *Psychological Research, 58*(3), 177-192. doi:10.1007/BF00419633

McMenamin, B. W., Shackman, A. J., Greischar, L. L., & Davidson, R. J. (2011). Electromyogenic artifacts and electroencephalographic inferences revisited. *NeuroImage, 54*(1), 4-9. doi:http://dx.doi.org/10.1016%2Fj.neuroimage.2010.07.057

McMenamin, B. W., Shackman, A. J., Maxwell, J. S., Bachhuber, D. R., Koppenhaver, A. M., Greischar, L. L., & Davidson, R. J. (2010). Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG. *NeuroImage, 49*(3), 2416-2432. doi:http://dx.doi.org/10.1016%2Fj.neuroimage.2009.10.010

Meyer, M., Baumann, S., & Jancke, L. (2006). Electrical brain imaging reveals spatio-temporal dynamics of timbre perception in humans. *NeuroImage, 32*(4), 1510-1523. doi:http://dx.doi.org/10.1016/j.neuroimage.2006.04.193

Moore, B. C., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica, 88*(3), S. 320-333.

Mosher, J. C., & Leahy, R. M. (1998). Recursive MUSIC: a framework for EEG and MEG source localization. *Biomedical Engineering, IEEE Transactions on, 45*(11), 1342-1354. doi:http://dx.doi.org/10.1109/10.725331

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology, 24*(4), 375-425. doi:10.1111/j.1469-8986.1987.tb00311.x

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *The Journal of Neuroscience : the official journal of the Society for Neuroscience, 32*(49), 17572-17581. doi:10.1523/JNEUROSCI.3203-12.2012

Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *The Journal of Neuroscience : the official journal of the Society for Neuroscience, 31*(28), 10234-10240. doi:10.1523/JNEUROSCI.0411-11.2011

O'Sullivan, J. A.-C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, bht355. doi:10.1093/cercor/bht355

Palmer, C. &. (1994). Harmonic, melodic, and frequency height influences in the perception of multivoiced music. . *Perception & Psychophysics, 56*(3), 301-312. doi:10.3758/BF03209764

Power, A. J. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience, 35*(9), 1497-1503. DOI: 10.1111/j.1460-9568.2012.08060.x

Pratt, H., Starr, A., Michalewski, H. J., Dimitrijevic, A., Bleich, N., & Mittelman, N. (2009). Cortical evoked potentials to an auditory illusion: Binaural beats. *Clinical Neurophysiology, 120*(8), 1514-1524. doi:10.1016/j.clinph.2009.06.014

Purcell, D. W., John, S. M., Schneider, B. A., & Picton, T. W. (2004). Human temporal auditory acuity as assessed by envelope following responses. *The Journal of the Acoustical Society of America, 116*(6), 3581-3593. http://dx.doi.org/10.1121/1.1798354

Pyper, B. J., & Peterman, R. M. (1998). Comparison of methods to account for autocorrelation in correlation analyses of fish data. *Canadian Journal of Fisheries and Aquatic Sciences, 55*(9), 2127-2140. doi:10.1139/f98-104

Schaefer, R. S., Desain, P., & Suppes, P. (2009). Structural decomposition of EEG signatures of melodic processing. *Biological Psychology, 82*(3), 253-259. doi:http://dx.doi.org/10.1016/j.biopsycho.2009.08.004

Schaefer, R. S., Farquhar, J., Blokland, Y., Sadakata, M., & Desain, P. (2011). Name that tune: decoding music from the listening brain. *NeuroImage, 56*(2), 843-849. doi:http://dx.doi.org/10.1016/j.neuroimage.2010.05.084

Schaefer, R., Vlek, R., & Desain, P. (2011). Decomposing rhythm processing: electroencephalography of perceived and self-imposed rhythmic patterns. *Psychological Research, 75*(2), 95-106. doi:10.1007/s00426-010-0293-4

Shahin, A., Bosnyak, D. J., Trainor, L. J., & Roberts, L. E. (2003). Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. *The Journal of Neuroscience, 23*(13), S. 5545--5552. ISSN: 1529-2401

Shahin, A., Roberts, L. E., Chau, W., Trainor, L. J., & Miller, L. M. (2008). Music training leads to the development of timbre-specific gamma band activity. *NeuroImage, 41*(1), 113-122. doi:http://dx.doi.org/10.1016/j.neuroimage.2008.01.067

Shahin, A., Roberts, L. E., Pantev, C., Trainor, L. J., & Ross, B. (2005). Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. *Neuroreport, 16*(16), 1781-1785. doi:10.1097/01.wnr.0000185017.29316.63

Shahin, A., Trainor, L., Roberts, L., Backer, K., & Miller, L. (2010). Development of auditory phase-locked activity for music sounds. *Journal of Neurophysiology, 103*(1), 218. doi:10.1152/jn.00402.2009

Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin, 133*(5), 780. doi: http://dx.doi.org/10.1037/0033-2909.133.5.780

Snyder, J. S., Alain, C., & Picton, T. W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *Journal of Cognitive Neuroscience, 18*(1), 1-13. doi:10.1162/089892906775250021

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *The Journal of the Acoustical Society of America, 117*(3), 1285-1298. doi: http://dx.doi.org/10.1121/1.1854312

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics, 69*(1), 136-152. doi:10.3758/BF03194460

Thompson, J. (2013). *Neural decoding of subjective music listening experiences - unpublished Master's Thesis.* Dartmouth College Hanover, New Hampshire. Hanover: Dartmouth College Hanover, New Hampshire.

Trainor, L. J. (2015). The origins of music in auditory scene analysis and the roles of evolution and culture in musical creation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 370*(1664), 20140089. DOI: 10.1098/rstb.2014.0089

Trainor, L., McDonald, K., & Alain, C. (2002). Automatic and controlled processing of melodic contour and interval information measured by electrical brain activity. *Journal of Cognitive Neuroscience, 14*(3), 430-442. doi:10.1162/089892902317361949

Treder, M. S., & Blankertz, B. (2010). Research (C) overt attention and visual speller design in an ERP-based brain-computer interface. *Behav. Brain Funct, 6*, 1-13. doi: 10.1186/1744-9081-6-28

Treder, M. S., Purwins, H., Miklody, D., Sturm, I., & Blankertz, B. (2014). Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *Journal of Neural Engineering, 11*(2), 026009. doi:dx.doi.org/10.1088/1741-2560/11/2/026009

Tremblay, K. L., & Kraus, N. (2002). Auditory training induces asymmetrical changes in cortical neural activity. *Journal of Speech, Language, and Hearing Research, 45*(3), 564-572. doi:10.1044/1092-4388(2002/045)

Tremblay, K. L., Ross, B., Inoue, K., McClannahan, K., & Collet, G. (2014). Is the auditory evoked P2 response a biomarker of learning? *Frontiers in Systems Neuroscience, 8*. doi:http://dx.doi.org/10.3389%2Ffnsys.2014.00028

Tremblay, K., Kraus, N., McGee, T., Ponton, C., Otis, & Brian. (2001). Central Auditory Plasticity: Changes in the N1-P2 Complex after Speech-Sound Training. *Ear & Hearing, 22*(2), 79-90. doi:10.1097/00003446-200104000-00001

Van Noorden, L. H. (1975). *Temporal coherence in the perception of tone sequences, unpublished PhD thesis.* Unpublished doctoral dissertation, Technische Hogeschool Eindhoven, the Netherlands, Netherlands.

Weise, A., Bendixen, A., Müller, D., & Schröger, E. (2012). Which kind of transition is important for sound representation? An event-related potential study. *Brain Research, 1464*, 30-42. doi:http://dx.doi.org/10.1016/j.brainres.2012.04.046

Will, U., & Berg, E. (2007). Brain wave synchronization and entrainment to periodic acoustic stimuli. *Neuroscience Letters, 424*(1), 55-60. doi:http://dx.doi.org/10.1016/j.neulet.2007.07.036

Winkler, I., Takegata, R., & Sussman, E. (2005). Event-related brain potentials reveal multiple stages in the perceptual organization of sound. *Cognitive Brain Research, 25*(1), 291-299. doi:10.1016/j.cogbrainres.2005.06.005

Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi, R., Sutoh, T., . . . Kaneko, S. (2001). Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration. *Brain Research, 897*(1), 222-227. doi:10.1016/S0006-8993(01)02224-7

Ziehe, A., Laskov, P., Nolte, G., & Müller, K.-R. (2004). A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *The Journal of Machine Learning Research, 5*, 777-800.

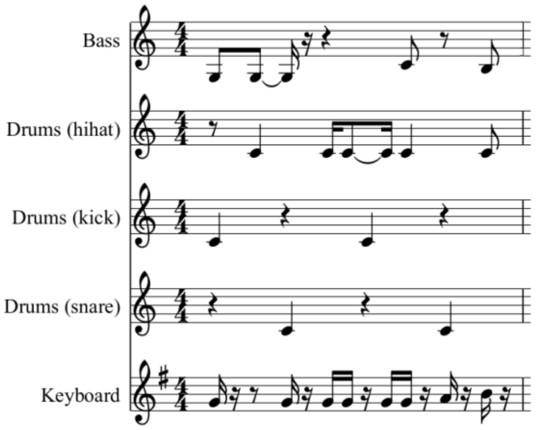

Figure 1: Score of ensemble version stimulus. Drums, although consisting of three instruments, are treated as one voice in the analysis. One (out of 63) music clips of 40 s duration consists of 21 repetitions of the depicted one-bar pattern. In addition, 14 solo clips were presented for each instrument.

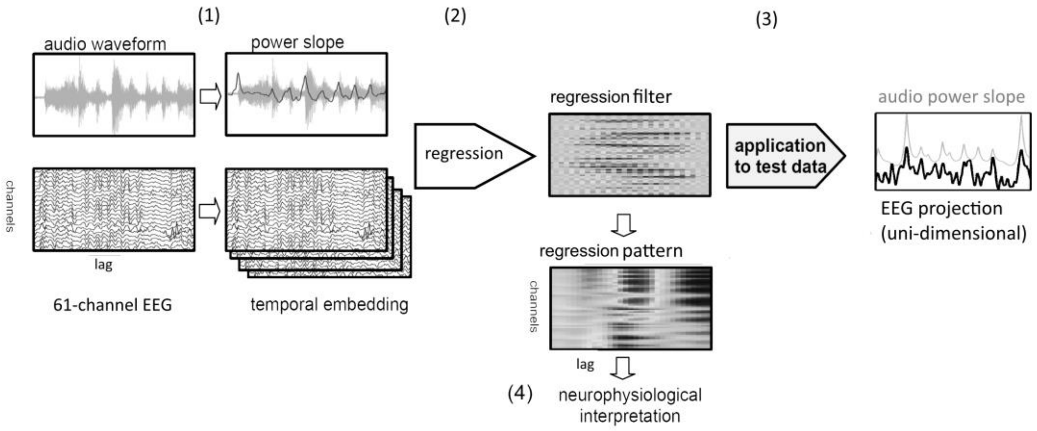

Figure 2: Analysis pipeline: (1) In a preprocessing step the power slope of the audio waveform is extracted by taking the first derivative of the signal's envelope. The 61-channel EEG signal is expanded with time shifted versions of the data to provide a range of time lags from 0 to 250 ms. (2) Training: A regression filter is trained to maximize the correlation between EEG recorded during the audio presentation and audio power slope. (3) Application to new data: The regression filter is applied to test data (EEG recordings of another presentation of the same stimulus, preprocessed as in step (1)) resulting in a uni-dimensional EEG projection. The goodness-of-fit is evaluated by assessing the relation between EEG projection and audio power slope. (4) The regression filter is transformed into spatio-temporal patterns that can be subject to further neurophysiological interpretation.

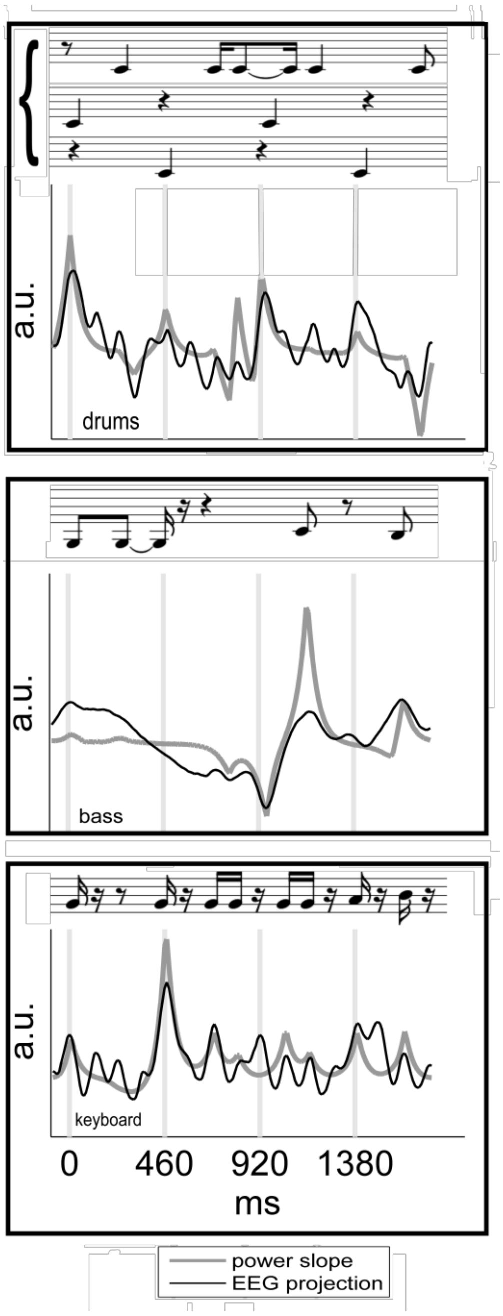

Figure 3: Solo clips: Grand Average (11 subjects) of extracted EEG projection (black line) and audio power slope (grey line), averaged across bars. The light grey vertical lines indicate the beats of the four-four time.

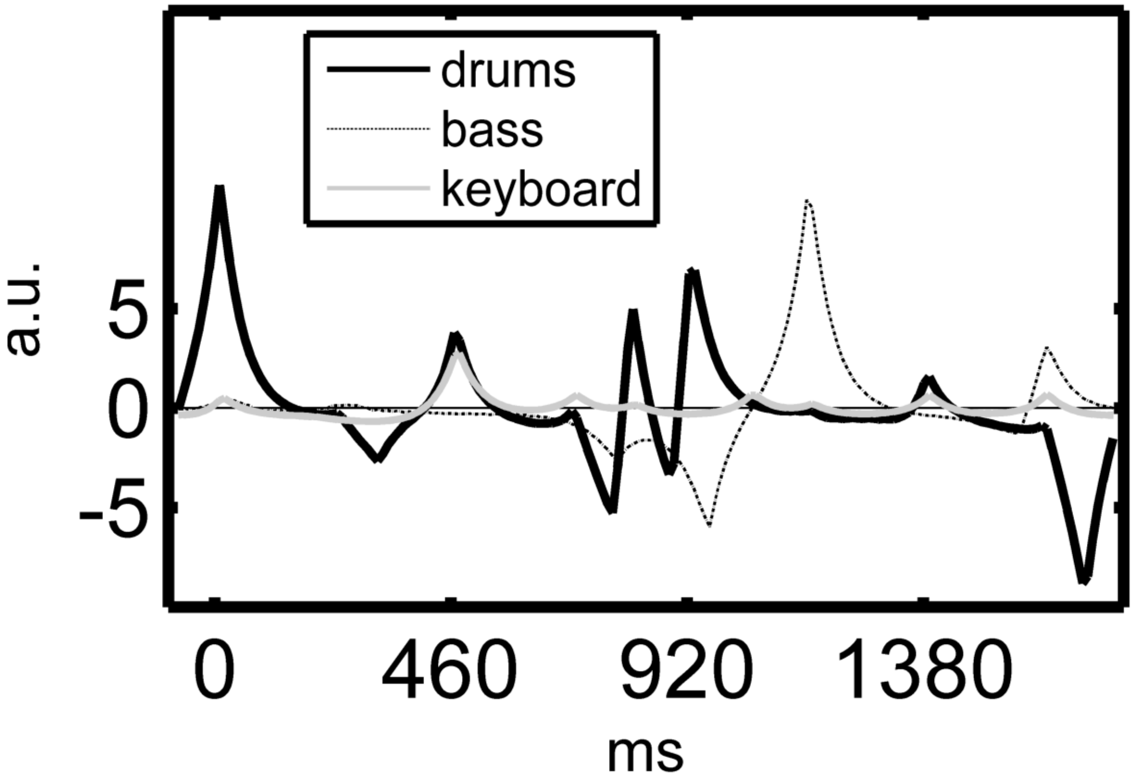

Figure 4: Audio power slopes of solo stimuli, displayed with identical scale. Amplitudes range between -8.8 and 11.2 for drums, between -5.9 and 10.5 for bass and between -0.7 and 2.8 for keyboard.

Table 1: Solo presentations: Percentage of 14 solo clips that were reconstructed with significant correlation from the EEG for the three instruments.

subject	drums	bass	keyboard
S1	100	75	67
S2	0	36	14
S3	31	100	21
S4	93	64	29
S5	57	36	64
S6	43	0	7
S7	57	79	21
S8	71	79	21
S9	71	57	50
S10	50	64	57

S11	29	64	7

Table 2: Solo clips: Correlation between EEG-reconstructed power slopes (averaged across 14 music clips) and audio power slope for single subjects and the three instruments. Significance of correlation was determined taking into account the effective degrees of freedom and applying a Bonferroni correction for N=11 subjects. Shaded cells indicate significant correlation at the level of alpha=0.05. GA: Grand average over 11 subjects.

subject	drums	bass	keyboard
S1	0.43	0.34	0.32
S2	0.23	0.26	0.21
S3	0.26	0.49	0.25
S4	0.52	0.39	0.17
S5	0.27	0.28	0.34
S6	0.22	0.13	0.08
S7	0.33	0.42	0.23
S8	0.35	0.45	0.24
S9	0.38	0.40	0.32
S10	0.32	0.33	0.30
S11	0.28	0.38	0.12
GA	0.60, p=0.00014, df_corrected=34,(df_u	0.52, p=0.00011,	0.54, p=0.0000004,

ncorrected=3968)	df_corrected=48	df_corrected=72

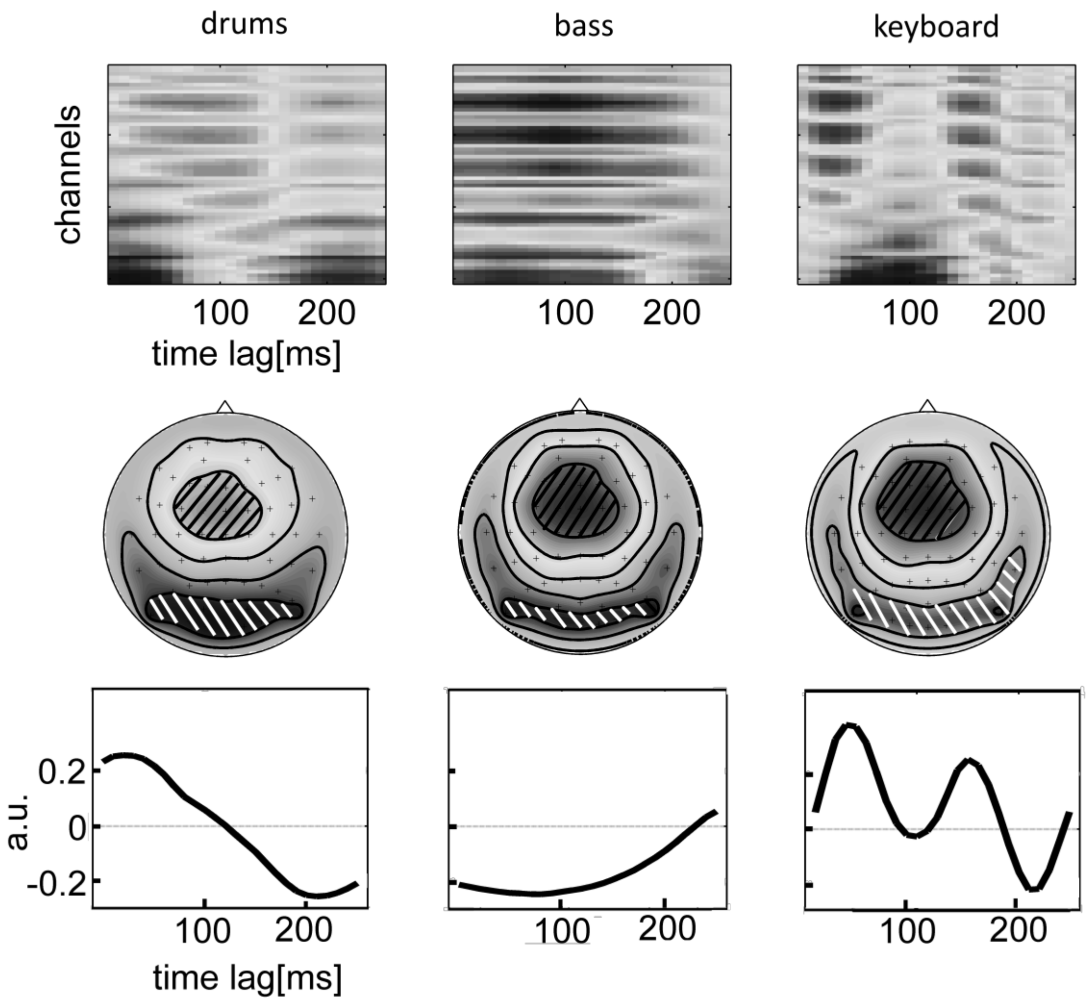

Figure 5: Spatio-temporal regression patterns and extracted MUSIC components for representative subject. Top: Regression patterns, middle: scalp pattern of first extracted MUSIC component: black hatching indicates positive weights, white hatching negative weights. Bottom: time course of first extracted MUSIC component.

Table 3: Ensemble version clips: Correlation between instrument-specific power slopes reconstructed from the EEG of the ensemble presentation (averaged across 63 music clips) and audio power slope of the respective single instrument for all 11 subjects and the three instruments. Significance of correlation was determined by estimating the effective degrees of freedom and applying a Bonferroni correction for N=11 subjects. Shaded cells indicate significance of correlation at the level of alpha=0.05.

subject	drums	bass	keyboard
S1	0.36	0.22	0.38
S2	-0.13	-0.06	0.25
S3	-0.07	-0.14	0.16
S4	0.0	-0.11	0.35
S5	-0.23	-0.06	0.47
S6	0.01	-0.12	0.25
S7	-0.01	0.23	0.20
S8	0.09	0.0	0.12
S9	-0.12	-0.09	0.36
S10	0.2	0.08	0.25
S11	0.26	0.09	0.20
GA	0.04	0.01	0.45,p=0.0001 df_corrected=69.

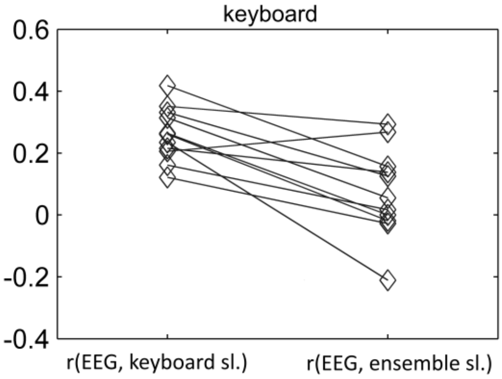

Figure 6: The EEG-reconstructed Keyboard power slope extracted from the EEG of the ensemble presentation by applying the keyboard-specific filter is correlated higher with the solo keyboard audio power slope than with the ensemble version audio power slope.

Table 4: Correlation between audio power slopes of solo and ensemble version stimuli

Correlation coefficient between power slopes	bass	keyboard	ensemble version
drums	-0.15	0.24	0.48
bass			-0.05
keyboard	0.06		0.26