Natural Stimuli Improve Auditory BCIs with Respect to Ergonomics and Performance

Johannes Höhne, Konrad Krenzlin, Sven Dähne, Michael Tangermann

Berlin Institute of Technology Machine Learning Group Franklinstr. 28/29 10587 Berlin, Germany

E-mail: j.hoehne@tu-berlin.de

Abstract. Moving from well-controlled, brisk artificial stimuli to natural and less controlled stimuli seems counter-intuitive for event-related potential (ERP) studies. As natural stimuli typically contain a richer internal structure, they might introduce higher levels of variance and jitter in the ERP responses. Both characteristics are unfavorable for a good single-trial classification of ERPs in the context of a multi-class Brain-Computer Interface (BCI) system, where the class discriminant information between target stimuli and non-target stimuli must be maximized.

For the application in an auditory BCI system, however, the transition from simple artificial tones to natural syllables can be useful despite of the variance introduced. In the presented study healthy users (N=9) participated in an offline auditory 9-class BCI experiment with artificial and natural stimuli. It is shown that the use of syllables as natural stimuli does not only improve the users' ergonomic ratings, also the classification performance is increased. Moreover, natural stimuli obtain a better balance in multi-class decisions, such that the number of systematic confusions between the nine classes is reduced. Hopefully, our findings may contribute to make auditory BCI paradigms more user-friendly and applicable for patients.

Submitted to: *J. Neural Eng.*

1. Introduction

Differences in brain responses evoked by either target or non-target stimuli during visual and auditory oddball paradigms are long known [1]. These differences, expressed in event-related potentials (ERP) of the electroencephalogram (EEG) upon visual stimuli, have been used to control visual multi-class Brain-Computer Interfaces (BCI) since the eighties [2]. However, recent work [3, 4] suggested that a standard visual ERP paradigm (as the matrix speller) might suffer from a drastic performance breakdown, if a user is not able to control gaze direction, stabilize gaze direction, or even lacks control over eye lid closure.

Increasing awareness of the BCI community that traditional visual ERP paradigms have limited use for that population of severely impaired users gave rise to a number of new auditory BCI paradigms [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], being independent of the visual domain. While earlier approaches [16] could show the basic feasibility of auditory BCIs, but suffered from very low theoretical or practical information transfer rates (ITR), some of the recent paradigms showed a break-through in performance [14, 13, 17]. In online studies with healthy subjects, their communication rate came close to that of covert visual attention paradigms. Furthermore new auditory user interfaces for text spelling were introduced, that in principle could be used without any visual components at all – be it for stimulation or feedback [14].

The AMUSE paradigm [14] and the PASS2D paradigm [13] both utilize rather brisk and artificially generated tones to elicit auditory ERP responses, with 6 tones of 40 ms duration (AMUSE) and 9 tones of 100 ms duration (PASS2D). The spatial direction of stimulus presentation as well as the pitch of stimuli were used to code for the multi-class paradigm. Though suited for relatively fast text entry, two practical drawbacks were observed that were related to this choice of stimuli.

Firstly, these highly controlled and very uniform tone sets were perceived as little intuitive and were – by single users – even described as unpleasant [13]. Taking into consideration that such ratings might indicate a limited overall acceptance of a final BCI spelling system, but also that the motivation of users is correlated with BCI performance [18, 19], an improvement of such subjective user ratings must be sought.

Secondly, a posterior analysis of the online spelling performance in both paradigms revealed a number of systematic multi-class confusions in the classification of targets vs. non-targets stimuli. Depending on the paradigm, the mis-classifications were related to front-back confusions (AMUSE) or confusions with stimuli that share some characteristics (PASS2D).

Even though visual ERP paradigms have undergone improvements by stimulus optimization [20, 21, 22, 19, 23], the stimulation principles for auditory BCI paradigms – as a relatively young line of research – were only rarely investigated. Initial attempts to compare different auditory stimulation principles can be found in [24, 12].

The goal of this work is to tackle both of the above mentioned problems (low user acceptance and confusion) simultaneously by improvements on the level of stimulation.

For this purpose a comparison between three auditory stimulus sets is performed within the PASS2D paradigm.

2. Methods

2.1. Participants

Nine healthy subjects (age: 24–26) participated in an offline BCI experiment comprising a single session of EEG recording. Two of the participants (*VPmg* and *VPlg*) had already participated in earlier BCI experiments. Each participant provided written informed consent, did not suffer from a neurological disease and had normal hearing. Subjects were not paid for participation. Prior to any processing, the data of the nine subjects were anonymized with a random code.

2.2. Experimental design

Within a single session, three conditions (i.e. three different sets of stimuli, see Sec. 2.4) were compared. The session was divided into several blocks that lasted approx. 10 min including a short break. Subjects were asked to perform six blocks at least, but they could decide to extend the recording in steps of three blocks. Two subjects performed nine blocks, while seven subjects chose to perform 6 blocks only.

Every block consisted of nine trials, with three consecutive trials showing the same type of stimulus (same condition). The order of conditions within trials was block-randomized. A trial was defined as a sequence of 135–144 auditory stimuli, subdivided into 15–16 iterations. As one iteration contained a complete set of nine stimuli in random order, each trial contained 15–16 target stimuli and 120–128 non-target stimuli.

During data preprocessing (see Sec. 2.5) only the last 14 iterations of each trial were considered. Removing the initial one or two iterations compensated for starting effects, such as orientation time necessary to direct spatial auditory attention to the target tone.

Participants were asked to concentrate on the occurrences of the target-stimulus and to neglect all other (non-target) stimuli. In addition, they were asked to count the targets and to report the number of occurrences at the end of each trial. Prior to the start of a new trial, the target stimulus was cued by three repetitive presentations. Targets were pseudo-randomized between trials, such that the number of targets was balanced between the nine classes.

2.3. Behavioral data

After the EEG recording, the participant filled out a questionnaire and rated each condition on a visual-analog scale, answering six questions per condition (translated from German):

(i) **Q1**: “How motivating does condition x appear to you?” (**motivation**)

(ii) **Q2**: "How do you judge your concentration while attending to stimuli in condition x?" (**concentration**)

(iii) **Q3**: "How tiring is condition x?" (**tiring**)

(iv) **Q4**: "How difficult was it to discriminate the stimuli in condition x?" (**discrimination**)

(v) **Q5**: "How exhausting is condition x?" (**exhaustion**)

(vi) **Q6**: "What is your overall impression of condition x?" (**overall**)

The scales were designed such that negative features (such as "hard to discriminate", "very exhausting" or "very tiring") were assigned low scores. To deliver a rating, subjects had to set a mark on a line of 10 cm length, which represented a continuous scale between the most negative and positive outcomes of each question.

2.4. Stimuli

Three different sets of auditory stimuli were used and formed the three conditions: (1) artificially generated tones, (2) spoken syllables and (3) sung syllables. Each set consisted of nine stimuli with spatial characteristics.

- The stimuli of condition 1 had already been successfully applied in an online study [13]. The nine artificially generated stimuli consisted of three tones with different pitch (high/medium/low) and also a varying tonal character. Each of the three tones was presented from three different directions (left/middle/right), leading to the 3×3 design shown in Fig. 1a.
- For condition 2, short spoken syllables were recorded by three speakers, visualized in Fig. 1b. Each speaker recorded three stimuli: syllables that either contained the vowel "i", an "æ" or an "o", like {ti, tæ, to, it æt, ot}. To obtain an intuitive separation of the stimuli, every speaker was presented only from one fixed direction (base: from the left, tenor: from the middle, soprano: from the right) . Thereby the 3×3 design of the PASS2D paradigm [13] was maintained since a column represented a speaker/direction and each row represented the vowel {"i", "æ" or "o"}, see Fig. 1b. The three different vowels lead to an intrinsic difference in the higher order harmonics, but the stimuli in condition 2 were all spoken and had no explicit pitch differences.
- For condition 3, the stimuli were recorded similar to condition 2. The only difference was that the syllables were not spoken, but sung by the same voices as in condition 2. Syllables with an "i" were sung with high pitch ($A\#$), syllables with an "æ" were sung with medium pitch (F), syllables with an "æ" were sung with low pitch ($C\#$). (The chosen pitches would result in a consonant chord, when they were played together.)

All stimuli were generated/recorded such that they had a duration of 100 ms (condition 1) or 125 ms (condition 2-3). Condition 2 was considered as an intermediate

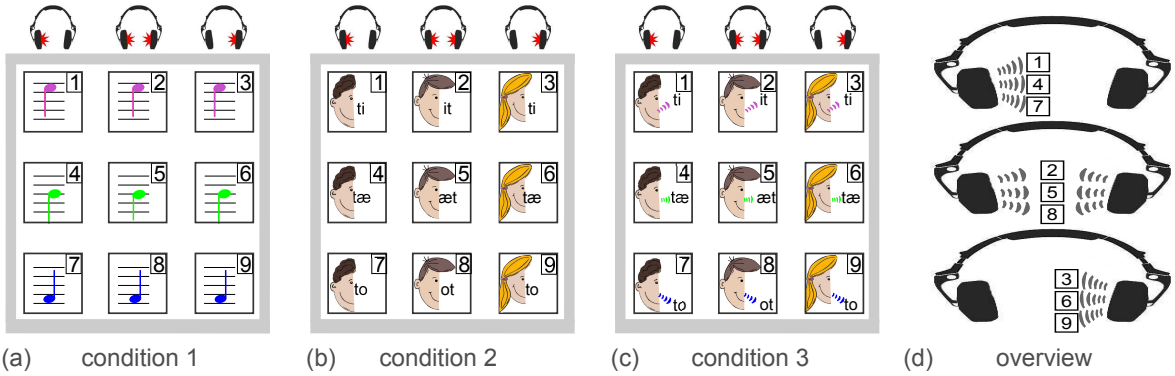

Figure 1. Graphical representation of the three sets of auditory stimuli.

condition in the following and we focused on the transition from condition 1 to condition 3, denoting the step from the maximally standardized artificial stimuli to the most complex natural stimuli.

Stimuli for multi-class auditory BCI paradigms are generally designed such that they are easy to discriminate on the one hand, but also similar enough to evoke at least similar target and non-target responses for each stimulus. In contrast to the artificial stimuli with a well defined onset (condition 1), the natural stimuli used in condition 2 and 3 had an intrinsic temporal diffuse characteristic, as shown in Fig. 2. Thus, the uniform and artificial stimuli in condition 1 didn't vary over time, while the stimuli in condition 3 (syllables) had a rather complex and heterogeneous temporal structure. However, the syllables in condition 3 were recorded and aligned such that vowels started at the same time (i.e. 30 ms) in each stimulus. This alignment had two advantages, as a sequence of stimuli was then perceived to be rhythmic and the class discriminative information in the stimuli was aligned. The time-frequency spectrograms in Fig. 2b show this alignment.

The stimuli were presented with a stimulus onset asynchrony (SOA) of 130 ms. A Terratec DMX 6Fire USB soundcard was utilized for stimulation, and light neckband headphones (Sennheiser PMX 200) enabled a comfortable audio perception. The mean latency of 51.4 ms (median: 50.5 ms, std: 4.46 ms, min: 41.2 ms, max: 61.8 ms) was corrected before the start of the data analysis. Pseudo-random sequences of stimuli were generated such that two subsequent stimuli were neither in the same row, nor in the same column (cp. to the 3 × 3 design shown in Fig. 1). As an example, none of the stimuli {5,6,1,7} was presented immediately after stimulus 4 had been presented. This constraint was implemented to prevent a consecutive presentation of two stimuli that share either speaker identity, pitch or direction. The stimulus presentation was programmed in Python and embedded in the PyFF framework [25].

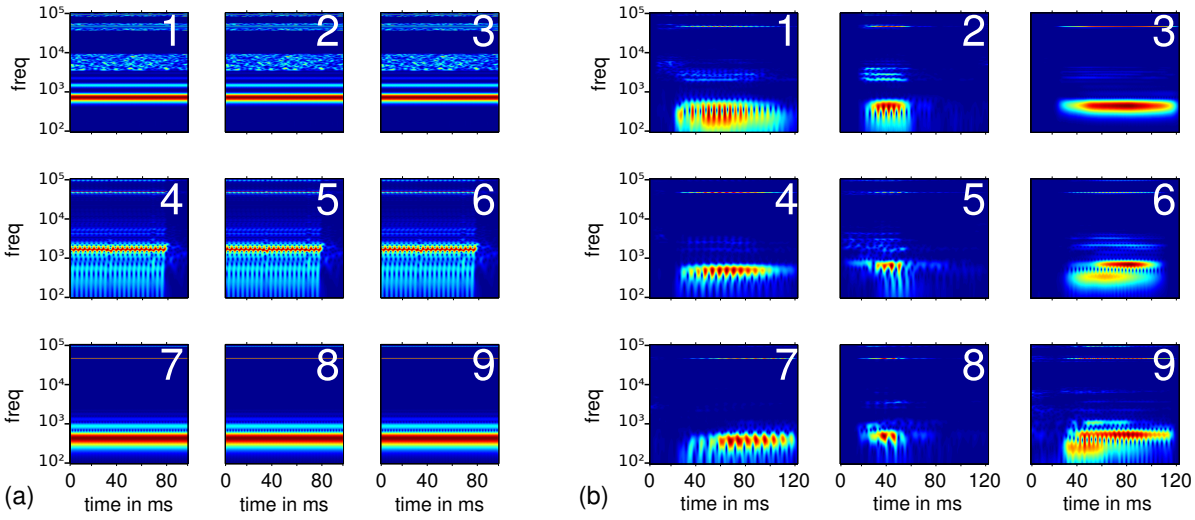

Figure 2. Spectrograms of auditory stimuli. Subplot (a) shows the spectrograms of three artificial tone stimuli used for condition 1. Stimuli were the same for the left, right and binaural presentation. Subplot (b) shows nine different stimuli used in condition 3, which consisted of sung syllables. In this condition, the directional presentation was supported by the use of different singers for the left, right and binaural stimuli.

2.5. Data Acquisition and Preprocessing

EEG signals were recorded with a Fast'n Easy Cap (EasyCap GmbH) using 63 monopolar, wet Ag/AgCl electrodes placed at symmetrical positions based on the extended international 10-20 system. Channels were referenced to the nose. Electrooculogram (EOG) signals were recorded via bipolarly referenced electrodes (vertical EOG: electrode Fp2 vs. an electrode directly below the right eye; horizontal EOG: F9 vs. F10). Two 32-channel amplifiers (Brain Products BrainAmp) processed the signals by an analog bandpass filter between 0.1 Hz and 250 Hz before digitalization (sampling rate 1 kHz). After applying the analog filter, the EEG raw data was first high-pass filtered at 0.2 Hz, then low-pass filtered at 25 Hz, both by a causal Chebyshev filter.

The EEG response to one stimulus is called subtrial in the following and comprises the most informative time period of 800 ms starting with the stimulus onset. DC offsets were subtracted based on the mean offset in a baseline interval of -150 ms to 0 ms relative to the stimulus onset. Thus, a stimulus was epoched at [-150 ms, 800 ms]. As 14 iterations of nine stimuli were contained in one trial, and three trials of each condition belonged to one block, the number of subtrials (before artifact rejection) was $14 * 9 * 3 = 378$ per block and condition, summing up to $6 * 378 = 2268$ subtrials for seven of the subjects, and $9 * 378 = 3402$ for two subjects.

Eye-artifacts were excluded by applying a moderate min/max-threshold criterion: subtrials were rejected if their peak-to-peak activity in at least one of the EOG channels exceeded 80 μV. On average over the subjects, this criterion lead to a rejection of 5.5 % of artifactual subtrials, while approximately maintaining the 1:8 ratio of targets and

non-targets.

2.6. Features and Classification

For each subtrial of the preprocessed EEG signals, a feature vector was obtained by computing the average amplitudes of 19 predefined intervals for all electrodes, resulting in 19 intervals × 63 channels = 1197 features per subtrial. The intervals are marked in the top plot of Fig. 5. Short time intervals of 30 ms length were chosen to cover earlier ERP components, while broader late components are sampled more coarsely by intervals of 60 ms length.

Binary classification of target and non-target epochs was performed using a linear Fisher Discriminant Analysis (FDA) regularized with covariance shrinkage [26]. The shrinkage is necessary due to the large dimensionality of the features in order to regularize the covariance matrix estimated for the FDA classifier. All subtrial epochs that survived the previous artifact rejection step (see Sec. 2.5) were used to estimate the classification performance. In order to account for the imbalance between targets and non-targets (ratio 1:8), the class-wise balanced accuracy is reported. It describes the average decision accuracy across classes (target vs. non-target) and has a chance level of 50 %. The binary classification accuracy was estimated by a 5-fold cross validation procedure, which itself was repeated five times with random shuffling of the epochs (5×5 CV).

Any performance comparison between the three stimulation conditions is based on EEG channels only. In addition, classification performance is estimated exclusively for EOG channels and for EOG combined with EEG. The latter two combinations are used only to upper-bound the unwanted influence of potential eye-related artifacts to an EEG-based system.

2.7. Simulation of Information Transfer Rates

It is noteworthy that the stimuli were presented in a rapid sequence (SOA of 130 ms), thus already lower binary classification accuracies may result in sophisticated communication rates. To compare the communication speed across several BCI paradigms, the Information Transfer Rate (ITR) is widely used. This measure has the advantage that it takes varying stimulation speeds into account. However, more timing issues such as inter-trial pauses have to be included in the ITR calculation, to come closer to a realistic value.

Our simulation targeted the ITR of a single block of online use of a BCI system. An online multi-class BCI experiment of 100 hours duration was simulated for each subject and each condition. Therefore, classifier outputs for target and non-target events were generated according to the binary accuracy, which was derived from the offline data analysis (see Sec. 3.3). Based on generated classifier outputs, trials were simulated and a multiclass decision was made as soon as an early-stopping criterion was fulfilled, at the latest after 20 iterations. For details on the dynamic stopping method (here: t-test),

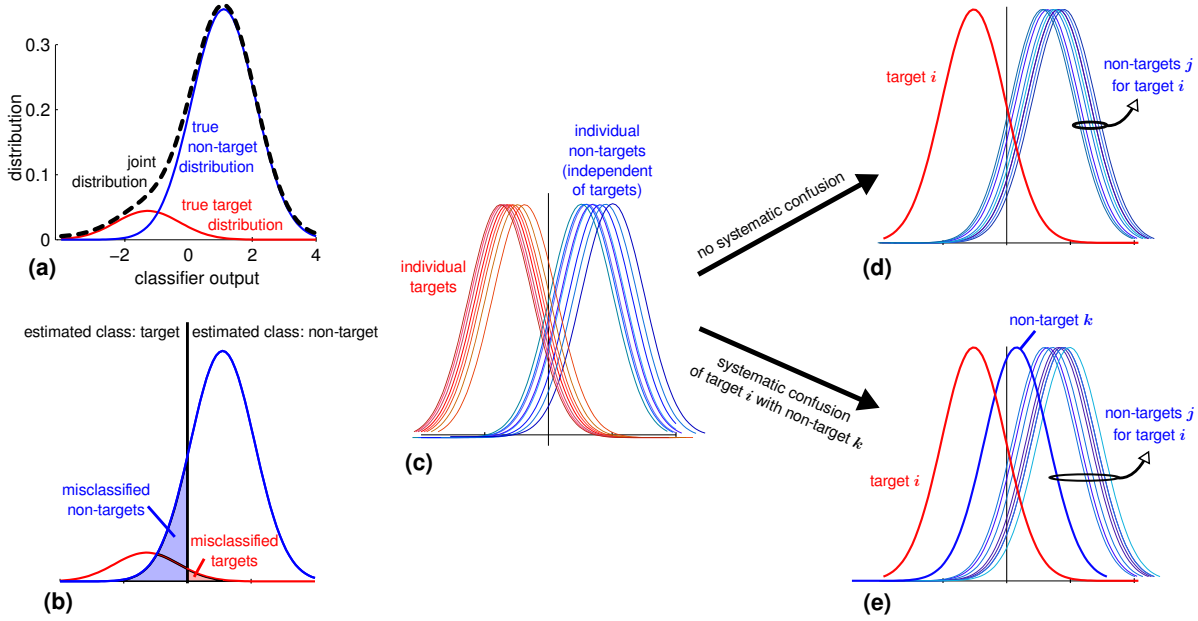

Figure 3. Schematic visualization of distributions of classifier outputs. Plot (a) depicts the distributions of classifier outputs for targets and non-targets, when all nine stimuli are pooled together. The distributions of misclassified stimuli are visualized in plot (b). Plot (c) shows the (rescaled) distributions for the 9 possible target and non-target stimuli. These distributions of classifier outputs disregard the trial structure, i.e. the distribution of non-targets j are relatively independent of a specific choice of a target i. In contrast, the plots (d) and (e) consider the trial structure. Here, distributions of non-targets j do depend on the choice of a specific target i. Plot (d) depicts a situation where there is no systematic confusion (no increased probability of a misclassification) between target i and any of the non-targets j. Plot (e) shows another example, where there is a systematic confusion between target i and the non-target k.

see [27]. A fixed inter-trial pause of 7 seconds was added after each trial, assuming that subjects need time to shortly relax and re-orient their attention to the next tone. The ITR (as defined in [28]) was then computed based on the number of correct and incorrect decisions after the simulated online BCI experiment.

2.8. Quantification of Systematic Confusions

This section deals with the question, whether or not there are pairs of stimuli that are more difficult to discriminate than others. One can pose the same question from the classifier's point of view by asking, whether or not there are pairs of stimuli that are more likely to be confused by the classifier than others. This phenomenon will be called "systematic confusion" in the following.

The problem of systematic confusions cannot be investigated with a measure for binary (target vs. non-target) classification accuracy alone, as shown in Fig. 3: in the depicted simulations, a binary classification accuracy of 90% is simulated in a 9-class paradigm. This is visualized in plot (a-c), where the red and blue curves show the distributions of classifier outputs (cl_{out}) for target and non-target stimuli. The shaded areas in (b) depict

the fraction of binary misclassifications which is 10% for both classes, with value 0 being the classification threshold. Investigating only plots (a-c) – which is a visualization of the binary classification accuracy – it is not possible to evaluate systematic confusions, since both situations plotted in (d) and (e) can evolve from distributions described in Fig. 3a-c.

In order to evaluate systematic confusions, a multi-class confusion matrix might be of limited help. It could only reflect the worst cases (misclassifications), but is unable to provide information about systematic similarities between some target- and non-target subclasses. Instead, we propose an introspection of the distributions of cl_{out}: the cl_{out} for all non-targets j have to be considered, when the user was focusing on target i. If there are no systematic confusions, then the distributions of cl_{out} for any non-target j is equally distributed and independent of the target stimulus i, as shown in Fig. 3d. If there are systematic confusions, then the distributions of cl_{out} for a non-targets j depend of the target stimulus i. Thus, when the user is attending to target i, there will be a non-target k which the BCI will classify more likely as target than other non-targets j (see Fig. 3e).

In the following, we will describe, how to statistically quantify the systematic confusions that were described above. In a typical BCI scenario, stimuli are presented in iterations, where in one iteration, each stimulus is presented exactly once in a pseudo-random order. Thus, in the given 9-class scenario, there is one classifier output for a target stimulus i and a classifier output for each of the 8 non-target stimuli in every iteration. We now denote the non-target j with the smallest (i. e. most negative) classifier output as the "worst non-target"($wNT_{j|i}$), since it is seen by the classifier most likely as the target. In the ideal case without systematic confusions, $wNT_{j|i}$ is independent of target stimulus i, as shown in plot 3c and the probability of being the "worst" non-target ($wNT_{j|i} = 1$) is 1/8 for each pair {i,j} and in each iteration. This can be described with a Bernoulli distribution. Accumulating the $wNT_{j|i}$ across iterations, we obtain the number of times that non-target j is the "worst" for target i, being referred to as $nNT_{j|i}$ in the following. In the situation without systematic confusions, $nNT_{j|i}$ is a binomial distributed random variable with $p = 1/8, k = nNT_{j|i}, n = nT_i$, where nT_i denotes the number of sequences with i being target.

Thus, if there are systematic confusions between target m and non-target l, then $nNT_{m|l}$ does not follow a binomial distribution. It is tested across all iterations and all subjects, whether or not $nNT_{i|j}$ follows a binomial distribution for any pair {i,j}. This test is independent of the overall binary accuracy.

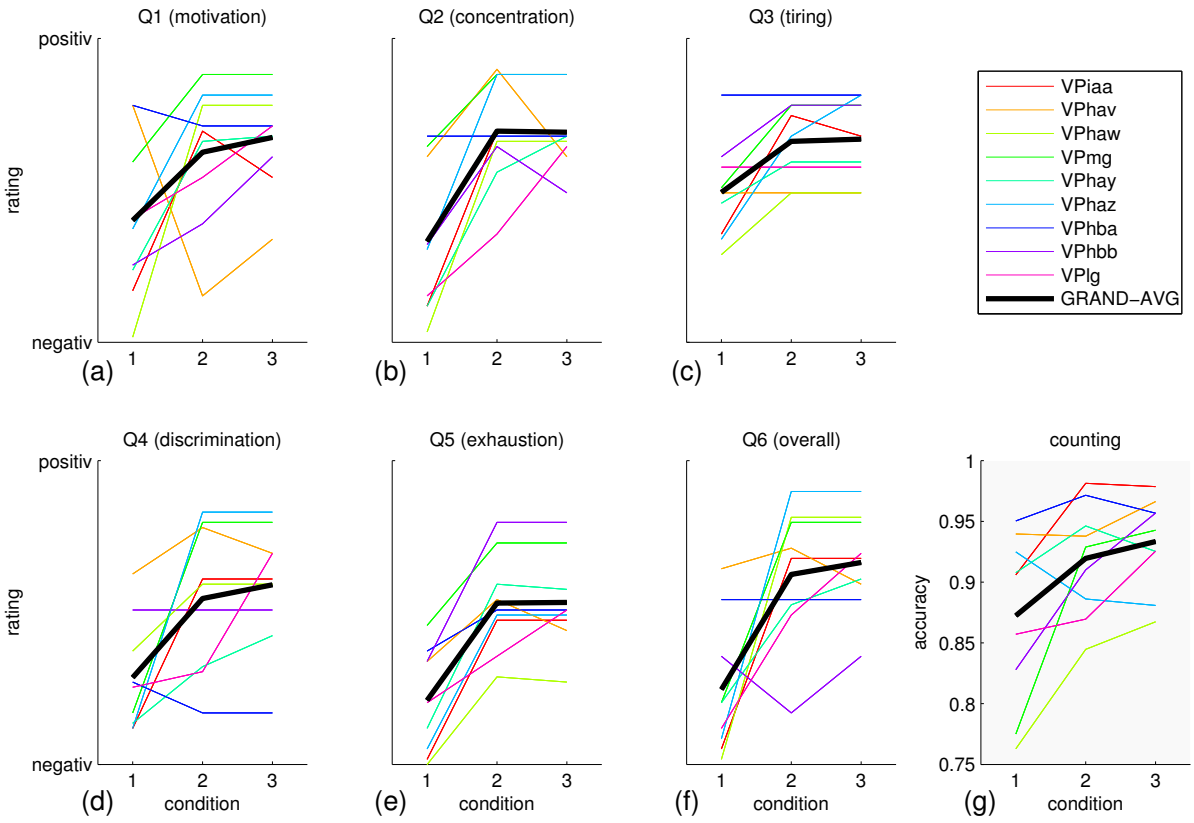

Figure 4. Overview over the behavioral data collected from each subject and the grand average. The subjective ratings of ergonomic aspects of three conditions are shown in subplots (a)–(f). Relative differences between reported counts and the true number of target stimuli are depicted in subplot (g).

3. Results

3.1. Behavioral Data

The subjective ergonomic ratings for the three stimulus conditions were assessed by questions Q1–Q6. These ratings as well as the objective counting accuracy show a clear trend: it was easier for the subjects to concentrate on natural stimuli (conditions 2, 3) than on artificial stimuli (condition 1). Participants rated the stimuli of condition 1 significantly more positive than condition 3 for each question (paired t-test, $p < 0.05$). As an example (Fig. 4d), all participants except *VPhba* rated the stimuli of condition 3 to be easier to discriminate than the stimuli of condition 1. The same trend can be seen for all other ergonomic ratings and also in the counting performances for the three conditions: the participants gave better ergonomic ratings and reported the number of targets more accurately for natural stimuli than for artificial stimuli.

3.2. ERPs

Fig. 5 shows time series of the event related potentials (ERPs), averaged over all participants for the three conditions. As expected, ERPs for non-target stimuli (gray

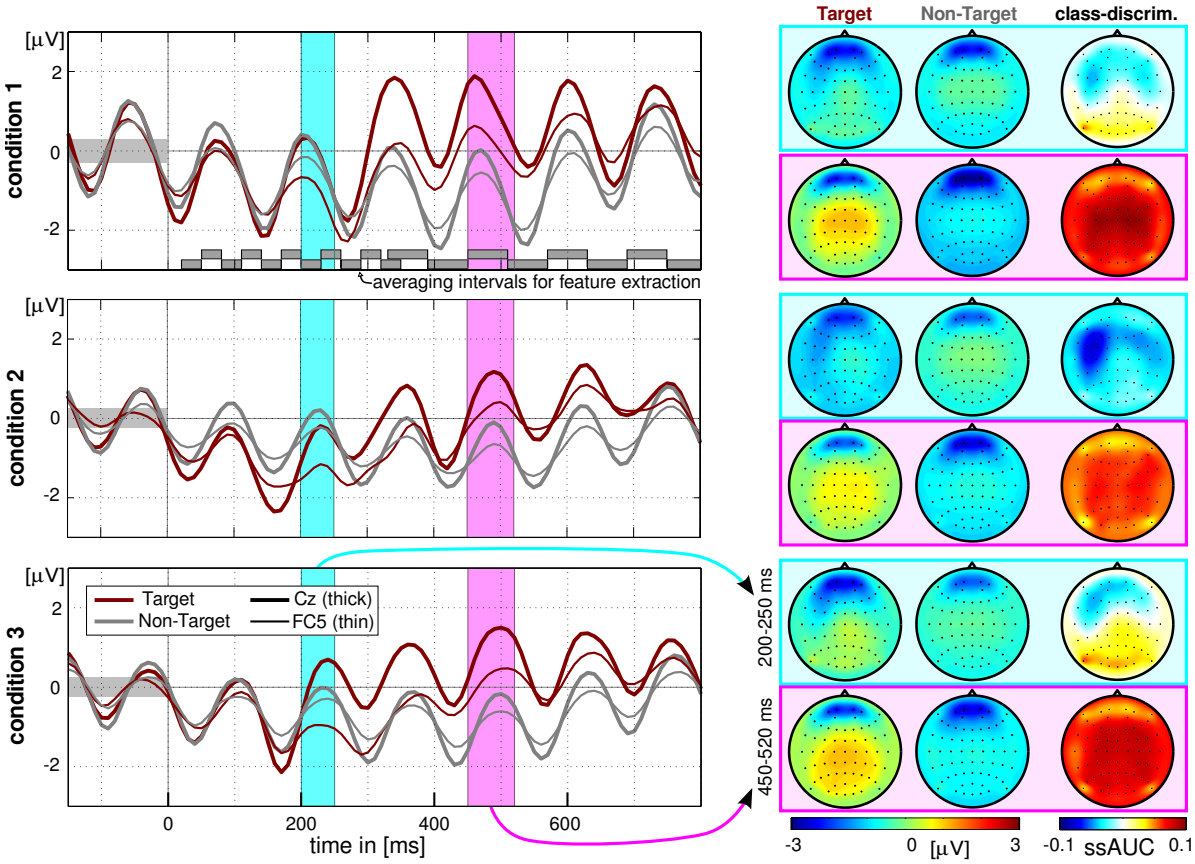

Figure 5. Grand average ERPs for target and non-target responses of the 3 conditions. *Time series plots (left):* From top to bottom, conditions 1–3 are visualized. For each condition, the average target and non-target responses are depicted for two EEG channels (FC5 and Cz). Two time intervals were marked in the time series plots: light blue intervals with a range of 200–250 ms after stimulus onset and light magenta intervals ranging 450–520 ms. The gray blocks in the top plot mark the 19 time intervals per channel used to calculate features for the classification task.
Scalp plots (right): For each condition and both colored intervals, three scalp plots are provided. They depict the average ERP activity for targets, non-targets, and of the distribution of class-discriminative information (ssAUC).

lines) in all three conditions show a regular pattern that mainly reflects early processing of the auditory stimuli. However, this regular pattern is not the same between the conditions, as a phase shift can be observed: regular responses for condition 1 have a shorter latency (approx. 30 ms) than responses for conditions 2 and 3. As a result, the peaks of steady state responses are slightly shifted to the left in condition 1.

The grand average spatial distribution of target- and non-target responses are shown in the scalp maps of Fig. 5. In addition, the class-discriminant information between target- and non-target responses was quantified with a signed and scaled measure of *area under the ROC curve*, called ssAUC. It is visualized as a third scalp map per condition and interval.

A common observation for all conditions is the appearance of a class-discriminant early

negative component. It is centered in fronto-temporal areas around 200 ms post stimulus onset. Recent psychophysiological literature [29] describes the same – or similar – early negative discriminative components in a spatial auditory multiclass paradigm as N2ac components. Common is also the existence of a class-discriminative late positive component. It shows a centro-parietal distribution starting around 300 ms and extends up to 700 ms. Its distribution resembles that of a P3b component, but appears much later than in standard oddball paradigms with slower stimulus presentation and less different classes.
Although largely similar, the scalp plots vary in details between conditions. We observe a trend of increased lateralization of class-discriminative early negative components to the left hemisphere for the natural conditions 2 and 3. The rightmost column of Fig. 5 suggests, that this effect is strongest for spoken syllables (condition 2).

3.3. Classification

The binary classification accuracy was computed for each subject and each condition with the finding that stimuli of condition 3 obtained a higher average accuracy than condition 1 and 2, see Fig. 6a. Over all participants, the class-wise balanced accuracy was between 50% (chance level) and 78%. Among the nine tested subjects, *VPmg* would profit in a special way from the two new conditions: evoked potentials can be classified clearly above chance level using natural tones, while this was not possible for the artificial tones. It can be observed that the performance curve of subject *VPhba* behaves against the general trend. This participant was also the only one who reported that it was easier for him/her to concentrate on stimuli of condition 1 than on the natural stimuli in conditions 2 and 3 (see Fig. 4d).

Individual scalp maps of class-discriminant activity did not give rise to conjecture a substantial influence of EOG activity to the classification results. However, since unconscious saccades and head movements in response to spatial auditory targets were already discussed in [30], we double-checked this by estimating (1) the classification performance on the two EOG channels only, and (2) on the combined EEG+EOG channels.
In scenario (1), average classification performances were 54.1 %, 53.8 % and 54.6 % for the three conditions. Although located close to chance level (50 %), the two EOG channels seemed to contain a small amount of task-relevant information. For comparison, the average absolute performance for EEG channels was about 10 % higher (63.4 %, 64.7 % and 65.8 %). The difference between scenario (1) and only EEG channels is significant.
In scenario (2), the difference in classification performance between using EEG channels only, and EEG channels plus EOG channels was very small and not significant for any condition.
Taken together, these results point out that EOG channels did not provide any additional information compared to EEG channels. The small amount of class-

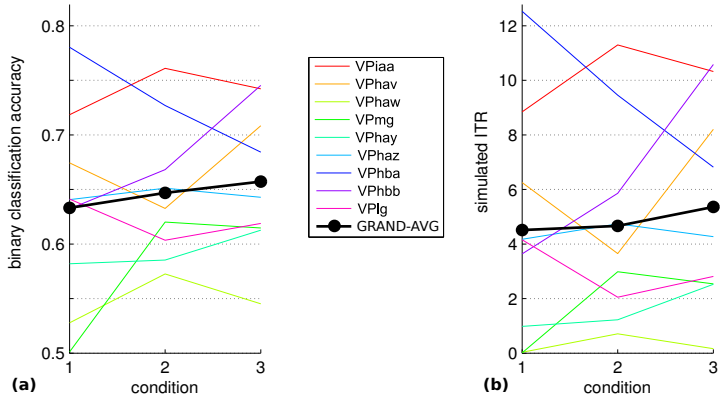

Figure 6. In subplot (a) the estimated binary classification accuracy is depicted for each subject (thin lines) and for the grand average (thick black line) for three conditions. Subplot (b) compares the resulting simulated Information Transfer Rate (ITR) in bits per minute for each subject and the grand average.

discriminative information contained in the EOG channels probably represents EEG activity that is picked up by the electrodes Fp2, F9 and F10.

3.4. Simulated Information Transfer Rate

The simulated ITR values for each condition were based on two assumptions: (1) the binary classification is constant over time and (2) there are no systematic confusions. Fig. 6b depicts the outcome of the simulation. We find that the average simulated ITR increases from 4.51 bits/min to 5.31 bits/min by the transition from artificial stimuli (condition 1) to natural stimuli (condition 3), being highly competitive for gaze-independent BCIs.

3.5. Temporal and Spatial Distribution of Discriminative Components

Fig. 7 shows, how class-discriminative information is distributed over time. A comparison between the shapes of lines reveals that the time structures are more similar between the two types of natural stimuli (red vs. blue) than between the artificial and the natural stimuli (black vs. red/blue). The blue line is above the red line for most subjects, which indicates a generally increased class-discrimination for condition 3 compared to condition 2. The black curve shows different peaks than the blue and red curve for several participants. This either indicates a temporal shift in components or the existence of different components when comparing the artificial stimulus condition to natural stimulus conditions. Noteworthy are the differences visible in the curves of subjects *VPiaa*, *VPhbb*, and *VPhba*.

In Fig. 7, it can be seen that *VPiaa* shows two distinct discriminative peaks for the natural sound condition. The first peak is centered at approximately 200 ms and the second about 150 ms later, at 350 ms. The second peak is strongly attenuated in the artificial sound condition and also about 50 ms delayed compared to the natural

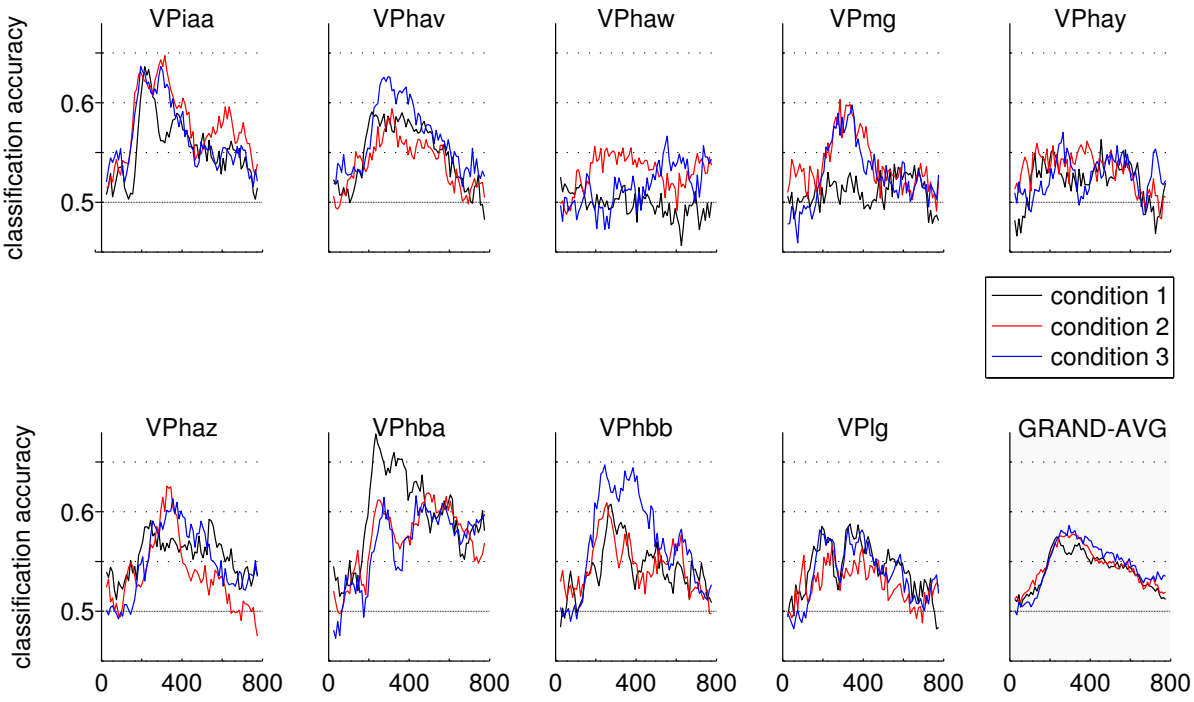

Figure 7. Distribution of discriminative information over time. For each subject and the grand average, class-discriminative information contained in all channels is estimated over time and compared for the three conditions. Classification accuracy is estimated within a sliding window of 50 ms width, which is used to scan the epoch. Most information is contained in the time windows around 300 ms after stimulus onset.

sound conditions. Subject *VPhbb* also shows two distinct peaks in the time-resolved classification plots in Fig. 7. For this subject it is the earlier peak that is attenuated in the artificial sound condition. Finally, in subject *VPhba*, we observe an effect that is contrary to that of other subjects: the transition from condition 1 to condition 3 leads to a weakening of both components.

For the three mentioned subjects, we plot the individual spatio-temporal dynamics of class-discriminative information in Fig. 8. The plots show r-square values for each channel and each time point. For two selected time intervals, we plot the temporal average of r-square values as scalp plots.

For subject *VPhbb*, an early negative discriminative component is found in condition 3, which is entirely absent in condition 1. While a positive component (P300) was found in both conditions, the spatial distributions vary. Thus, changing from condition 1 to 3, neither the position on the scalp, nor timing or intensity of class discriminant components are maintained for subject *VPhbb*.

For subject *VPiaa*, the transition from condition 1 to 3 leads to an earlier appearance of discriminative components. While the approximate spatial distribution is maintained for both components, the intensity of class discrimination varies for the two conditions: the early component is slightly weakened in condition 3, while the positive component is considerably increased in condition 3 compared to condition 1. For *VPhba*, the intensity of both components decreases by far, while the temporal and

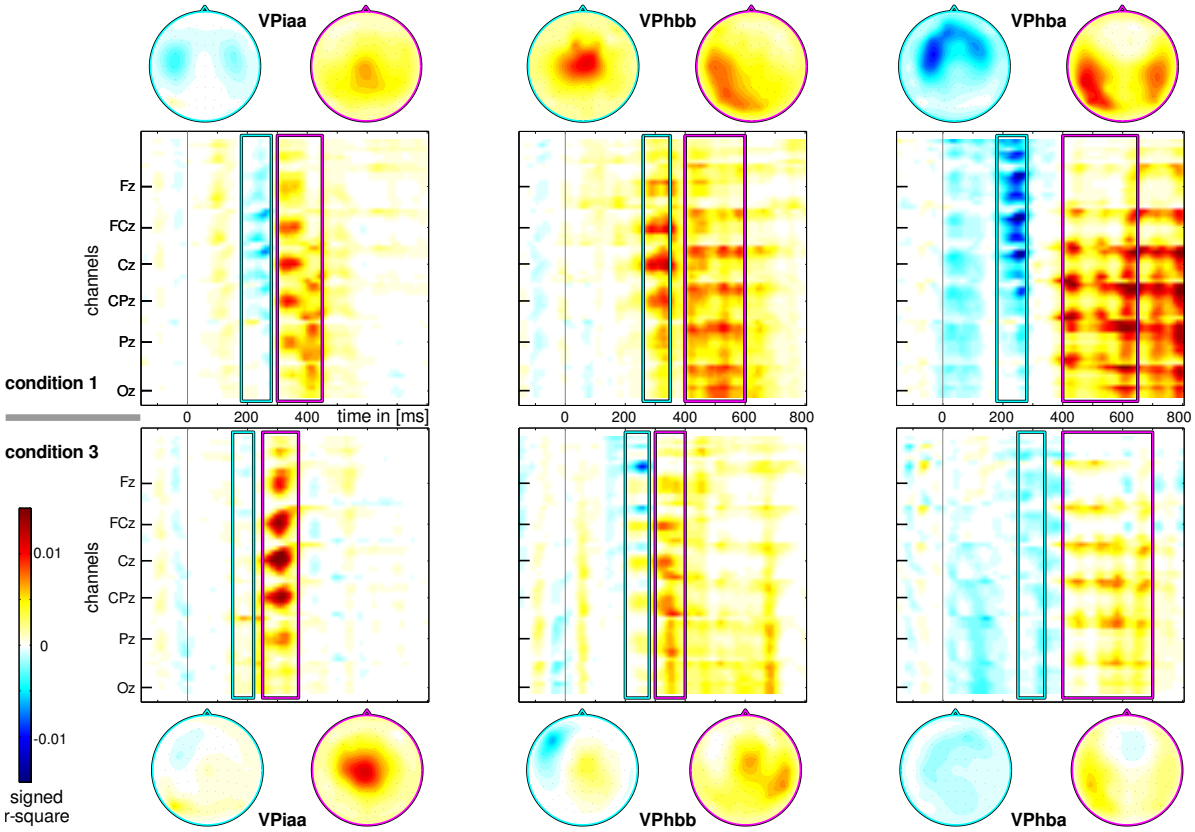

Figure 8. Spatio-temporal distribution of class-discriminative information for three selected subjects (arranged in columns) and for two conditions (arranged in rows). For each combination, one matrix plot and two scalp plots are provided. All plots share the same color scale.
A matrix plot shows signed r-square values for each EEG channel (y-axis) and time bin (x-axis). Channels are sorted from front to back and left to right, with occipital channels located in the bottom rows.
The two scalp plots depict averaged r-square values for two individually chosen time intervals, capturing early and late class-discriminative components. Their positions in time are marked by light blue and light magenta rectangles in the corresponding matrix.

spatial characteristics are maintained.

Another interesting aspect that becomes evident in the scalp maps of Fig. 8 is a change of shape in early discriminative components. The spatial distribution of r-square values in the early interval is rather symmetric in all subjects in condition 1. However, in condition 2 and condition 3 the maps are more asymmetric, as the center of mass of the early negative components is shifted towards left fronto-temporal regions. This shift is also visible in the grand-average ssAUC maps of early time intervals in Fig. 5, right-most column.

3.6. Joint Effects on Classification Performance and Behavioral Data

Thus far we have shown that classification performance and ergonomic rating of the stimuli increased when making the transition from the artificial stimuli in condition 1 to more natural stimuli in condition 3. It remains to be shown however, that this effect occurs *simultaneously* in the majority of subjects.

Fig. 9 shows the classification performance as well as the stimulus ratings for condition 1 and 3, pooled over subjects and rating questions. In order to preserve visibility, we only included two conditions (1 and 3) and do not differentiate between individual subjects or questions. It is difficult to make assertions about joint effects in the ratings, because their respective ranges differ. Thus, we standardized the classification performance and the stimulus ratings by z-scoring them (i.e. removing the mean and dividing by the standard deviation). Fig. 9a shows that for condition 1 the majority of subjects rated the stimuli less ergonomic and the classification performance was lower, compared to condition 3. Subjects rated each condition with respect to six categories, as shown in Fig. 4. This resulted in six data points for each subject and condition. Having the same classification performance, those sample points appear on a horizontal line in Fig. 9a. Two sample points (classification/stimulus rating) of condition 1 are connected with their corresponding sample point in condition 3. The arrows always point from condition 1 to condition 3 and thereby mark the effect of the transition from artificial to more natural stimuli. In Fig. 9b, such transition arrows are plotted for all subjects and rating questions, with the color indicating the identity of the subjects (same color code was used as in Fig. 6). The vast majority of transition arrows points into the upper right quadrant, which represents a simultaneous increase in perceived stimulus ergonomics and classification performance.

3.7. Systematic Confusions

The systematic confusions by the classifier were analyzed according to the method described in Section 2.8. Significant systematic confusions are present in all three conditions, but the number of confusions is reduced by natural stimuli in conditions 2 and 3 compared to artificial stimuli in condition 1 (see Fig. 10). Condition 3 (as the condition with the best neuroergonomic design) exhibits the smallest number of confusions. It should be noted that the number of systematic confusions is independent of the binary classification accuracy, as shown in Fig. 3.

4. Discussion

Auditory BCI paradigms are a potential solution for severely motor-disabled patients, as they can be utilized independent of gaze control or eye blinks. Aiming to improve existing auditory BCI paradigms with respect to usability and performance, this study investigates the use of natural auditory stimuli.

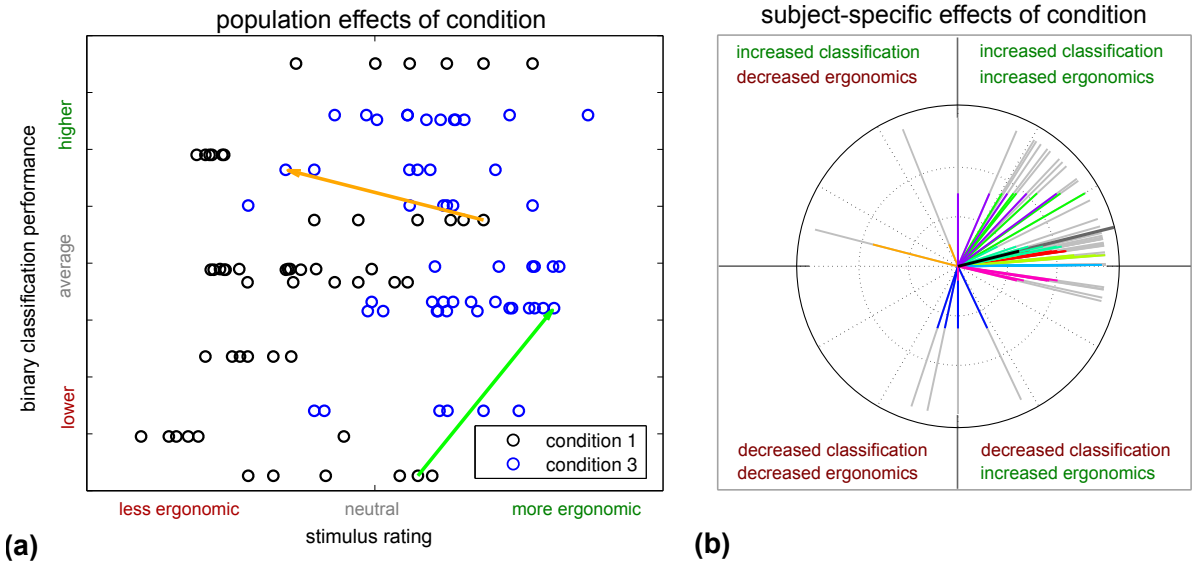

Figure 9. Joint effects of the stimulus condition (1 and 3) on both, classification accuracy and ergonomic ratings. One classification accuracy and six ergonomic ratings expressed by the VAS scores for questions Q1–Q6 were available per subject and condition. All values were standardized by z-scoring and entered into the scatter plot (a). The subject-specific changes from condition 1 to condition 3 are depicted in plot (b). Gray bars indicate directions of change, while the colored portion of the bar indicates the its magnitude (relative to the maximum over all subjects). Subject identity is color-coded as in Figures 4 and 6, with black representing the grand average.

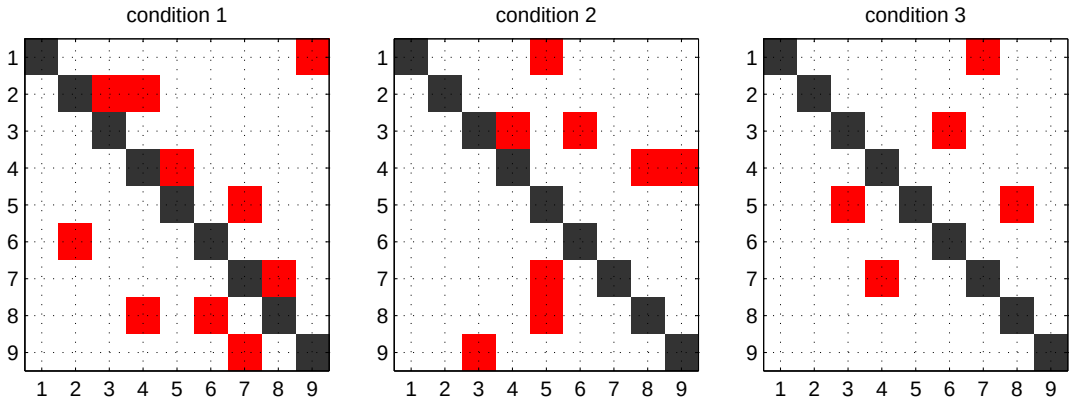

Figure 10. Systematic confusions of stimuli for each condition. A row of a confusion matrix corresponds to one stimulus in the role of a target. A red square with index i, j marks systematic confusions of the non-target j with this target i (p-value of binomial distribution ≤ 0.05). Example: in condition 1, target stimulus 4 is systematically confused with non-target stimulus 5.

The transition from artificial to natural stimuli was motivated by the idea to utilize our over-trained ability of speech processing. First, this comprises the decomposition of a complex auditory stream into relevant components, such as syllables. Second, we are trained to focus on one out of several voices that we hear from different directions.

Tones produced by singers offer a large number of class-discriminant cues for the BCI user (e.g. harmonics, pitch, direction, voice-characteristics). Even though the syllables used in this study are more complex and less standardized than the artificial tones, they allow for better classification rates and lead to increased subjective ergonomic ratings. In short, the auditory BCI became "faster" and was considered "more pleasant" when using these more natural stimuli.

Of course it is an interesting question, whether or not the syllables evoke ERP components, that are different from those evoked by artificial tones. In fact we observed considerable changes of individual ERP responses (Sec. 3.5) as well as in the grand average responses (Sec. 3.2) by changing the stimulus condition. The delay of 30 ms in the grand average for the time series of natural stimuli might best be explained best by a delayed perception of natural stimuli. An alternative explanation would be based on an increased mental processing demand for the natural stimuli due to their higher complexity. In the grand average, the trend of increased lateralization of early negative components to the left hemisphere was observed especially for the spoken syllables (Fig. 5). This lateralization of language-related processing in the human brain was observed before [31] and it is plausible that language-related brain areas become increasingly involved in the processing syllables compared to tones. As the lateralization is best reflected in the ssAUC scalp maps (see rightmost column of Fig. 5), this suggests an active role of language-related areas during the discrimination of target and non-target stimuli. Latencies and amplitudes of late positive class-discriminative components were rather unstable between conditions, when compared on an individual basis. Assuming that these components might represent P3b components, which are known for their stability, this variation comes unexpected on the one hand. On the other hand, our multiclass setup with short SOA is far from the standard oddball paradigm.

Even though a rather fast stimulation speed (SOA: 130 ms) was applied, our results show that users can handle such a rapid sequence of adequately designed stimuli. We assume, that auditory stimuli can in principle be presented with at least the speed as visual BCI paradigms.

Moreover, this study demonstrates the problem of systematic confusions, which was mostly disregarded by the BCI community so far. A data driven approach to identify and quantify those confusions is presented. Based on this method, it is shown that next to increasing the classifier performance, also the number of systematic confusions can be reduced by a design of stimuli that follows neuroergonomic principles.

Several auditory BCI paradigms for text spelling were recently developed and successfully tested with healthy subjects [8, 13, 14]. A systematic study of whether or not auditory BCIs are applicable with end-users such as patients suffering from ALS for daily use, remains an open question. The improvement in experimental paradigm

presented here are is important step for the transfer of multi-class auditory BCIs from the lab into the real world and into patients' homes.

Acknowledgments

The authors are grateful for the financial support by several institutions: This work was partly supported by the European Information and Communication Technologies (ICT) Programme Project FP7-224631 and PASCAL2 Network of Excellence, ICT-216886. This publication only reflects the authors' views. Funding agencies are not liable for any use that may be made of the information contained herein. Furthermore, the authors thank Friederike von Möllendorff, Judith Tangermann and Aaron Dan for their support with voice recordings.

References

[1] S. Sutton, M. Braren, J. Zubin, and E. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, pp. 1187–1188, Nov 1965.

[2] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr Clin Neurophysiol*, vol. 70, pp. 510–523, 1988.

[3] 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.

[4] 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.

[5] F. Nijboer, A. Furdea, I. Gunst, J. Mellinger, D. McFarland, N. Birbaumer, and A. Kübler, "An auditory brain-computer interface (BCI)," *J Neurosci Methods*, vol. 167, pp. 43–50, Jan 2008.

[6] S. Kanoh, K. Miyamoto, and T. Yoshinobu, "A brain-computer interface (BCI) system based on auditory stream segregation," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of theIEEE.* IEEE, 2008, pp. 642–645.

[7] 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.

[8] D. S. Klobassa, T. M. Vaughan, P. Brunner, N. E. Schwartz, J. R. Wolpaw, C. Neuper, and E. W. Sellers, "Toward a high-throughput auditory P300-based brain-computer interface," *Clin Neurophysiol*, vol. 120, pp. 1252–1261, Jul 2009.

[9] R. S. Schaefer, R. J. Vlek, and P. Desain, "Decomposing rhythm processing: electroencephalography of perceived and self-imposed rhythmic patterns," *Psychol Res*, vol. 75, no. 2, pp. 95 – 106, 2011.

[10] 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.

[11] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann, "Two-dimensional auditory P300 speller with predictive text system," in *Engineering in Medicine and Biology Society, 2010. EMBS 2010. 32th Annual International Conference of the IEEE*, vol. 1, 2010, pp. 4185–4188.

[12] S. Halder, M. Rea, R. Andreoni, F. Nijboer, E. M. Hammer, S. C. Kleih, N. Birbaumer, and A. Kübler, "An auditory oddball brain-computer interface for binary choices." *Clin Neurophysiol*, vol. 121, no. 4, pp. 516–523, Apr 2010.

[13] 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.

[14] M. Schreuder, T. Rost, and M. Tangermann, "Listen, you are writing! speeding up online spelling with a dynamic auditory BCI," *Front Neuroscience*, vol. 5, p. 112, 2011.

[15] D.-W. Kim, H.-J. Hwang, J.-H. Lim, Y.-H. Lee, K.-Y. Jung, and C.-H. Im, "Classification of selective attention to auditory stimuli: Toward vision-free brain-computer interfacing," *J Neurosci Methods*, vol. 197, no. 1, pp. 180 – 185, 2011.

[16] N. Hill, T. Lal, K. Bierig, N. Birbaumer, and B. Schölkopf, "An auditory paradigm for brain–computer interfaces," in *Advances in Neural Information Processing Systems*, Y. W. Saul, L.K. and L. Bottou, Eds., vol. 17. Cambridge, MA, USA: MIT Press, 2005, pp. 569–576.

[17] J. 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.

[18] S. C. Kleih, F. Nijboer, S. Halder, and A. Kübler, "Motivation modulates the P300 amplitude during brain-computer interface use," *Clin Neurophysiol*, vol. 121, pp. 1023–1031, 2010.

[19] M. Tangermann, M. Schreuder, S. Dähne, J. Höhne, S. Regler, A. Ramsay, M. Quek, J. Williamson, and R. Murray-Smith, "Optimized stimulation events for a visual ERP BCI," *Int J Bioelectromagnetism*, vol. 13, no. 3, pp. 119–120, 2011.

[20] B. Z. Allison and J. A. Pineda, "ERPs evoked by different matrix sizes: implications for a brain computer interface (BCI) system," *IEEE Trans Neural Syst Rehabil En*, vol. 11, pp. 110–113, 2003.

[21] E. Sellers, D. Krusienski, D. McFarland, T. Vaughan, and J. Wolpaw, "A P300 event-related potential brain-computer interface (BCI): the effects of matrix size and inter stimulus interval on performance," *Biological Psychology*, vol. 73, pp. 242–252, Oct 2006.

[22] J. Hill, J. Farquhar, S. Martens, F. Bießmann, and B. Schölkopf, "Effects of Stimulus Type and of Error-Correcting Code Design on BCI Speller Performance," *Advances in Neural Information Processing Systems*, vol. 21, 2009.

[23] T. Kaufmann, S. Schulz, C. Grünzinger, and A. Kübler, "Flashing characters with famous faces improves ERP-based brain-computer interface performance," *J Neural Eng*, vol. 8, no. 5, p. 056016, 2011.

[24] M. Schreuder, M. Tangermann, and B. Blankertz, "Initial results of a high-speed spatial auditory BCI," *Int J Bioelectromagnetism*, vol. 11, no. 2, pp. 105–109, 2009.

[25] B. Venthur, S. Scholler, J. Williamson, S. Dähne, M. S. Treder, M. T. Kramarek, K.-R. Müller, and B. Blankertz, "Pyff – a pythonic framework for feedback applications and stimulus presentation in neuroscience," *Front Neuroscience*, vol. 4, p. 179, 2010.

[26] 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.

[27] M. Schreuder, J. Höhne, M. S. Treder, B. Blankertz, and M. Tangermann, "Performance optimization of ERP-based BCIs using dynamic stopping," in *Engineering in Medicine and Biology Society, 2011. EMBS 2011. 33th Annual International Conference of the IEEE*, 2011, pp. 4580–4583.

[28] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin Neurophysiol*, vol. 113, pp. 767–791, 2002.

[29] M. L. Gamble and S. J. Luck, "N2ac: An ERP component associated with the focusing of attention within an auditory scene," *Psychophysiology*, vol. 48, no. 8, pp. 1057–1068, 2011.

[30] S. Röttger, E. Schröger, M. Grube, S. Grimm, and R. Rübsamen, "Mismatch negativity on the cone of confusion," *Neurosci Lett*, vol. 414, no. 2, pp. 178 – 182, 2007.

[31] A. Friederici and K. Alter, "Lateralization of auditory language functions: A dynamic dual pathway model," *Brain and Language*, vol. 89, no. 2, pp. 267–276, 2004.