Pattern analysis of EEG responses to speech and voice: Influence of feature grouping

NeuroImage 59 (2012) 3641–3651

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Pattern analysis of EEG responses to speech and voice: Influence of feature grouping

Lars Hausfeld [a,b,*], Federico De Martino [a,b,c], Milene Bonte [a,b], Elia Formisano [a,b]

[a] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands
[b] Maastricht Brain Imaging Center, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands
[c] Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, USA

ARTICLE INFO

Article history:
Received 25 July 2011
Revised 24 October 2011
Accepted 16 November 2011
Available online 30 November 2011

Keywords:
EEG
ERP
Machine learning
Classification
Audition
Speech

ABSTRACT

Pattern recognition algorithms are becoming increasingly used in functional neuroimaging. These algorithms exploit information contained in temporal, spatial, or spatio-temporal patterns of independent variables (features) to detect subtle but reliable differences between brain responses to external stimuli or internal brain states. When applied to the analysis of electroencephalography (EEG) or magnetoencephalography (MEG) data, a choice needs to be made on how the input features to the algorithm are obtained from the signal amplitudes measured at the various channels. In this article, we consider six types of pattern analyses deriving from the combination of three types of feature selection in the temporal domain (*predefined windows*, *shifting window*, *whole trial)* with two approaches to handle the channel dimension (*channel wise*, *multi-channel*). We combined these different types of analyses with a Gaussian Naïve Bayes classifier and analyzed a multi-subject EEG data set from a study aimed at understanding the task dependence of the cortical mechanisms for encoding speaker's identity and speech content (vowels) from short speech utterances (Bonte, Valente, & Formisano, 2009). Outcomes of the analyses showed that different grouping of available features helps highlighting complementary (i.e. temporal, topographic) aspects of information content in the data. A *shifting window/multi-channel* approach proved especially valuable in tracing both the early build up of neural information reflecting speaker or vowel identity and the late and task-dependent maintenance of relevant information reflecting the performance of a working memory task. Because it exploits the high temporal resolution of EEG (and MEG), such a shifting window approach with sequential multi-channel classifications seems the most appropriate choice for tracing the temporal profile of neural information processing.

Introduction

Electroencephalography (EEG) and magnetoencephalography (MEG) are commonly used to study the time course of neural information processing in the human brain with high temporal resolution. In most cases, EEG/MEG studies rely on the comparison of averaged responses to repeated presentations of experimental conditions either in the temporal domain (event-related potentials [ERPs] or fields [ERFs], respectively) and/or in the frequency domain (event-related desynchronization and synchronization) (Pfurtscheller and Lopes Da Silva, 1999). Often, the statistical analyses (and related inferences on neural processing) are limited to a-priori specified (spectro-) temporal windows of interest – at channel or estimated source level – and therefore only a small subset of the measured signal is actually utilized.

This article illustrates several approaches to EEG data analysis based on *pattern recognition* (e.g. Bishop, 2007; Duda et al., 2001). In contrast to the conventional approach where a single dependent variable is examined (univariate statistics), these techniques exploit the information content in patterns of dependent variables (features), which are extracted from the measured signals. Pattern recognition allows analyzing EEG data in a more exploratory and data-driven manner and – similar to the recent developments in fMRI (e.g. Haynes and Rees, 2006) – promises to complement conventional approaches for EEG/MEG analysis.

A typical application of pattern recognition methods includes three steps, (1) extracting and selecting features (i.e. dependent variables), (2) learning a model with a machine-learning algorithm, and (3) determining the generalization ability of the learnt model using an independent evaluation dataset. In EEG/MEG, various *types* of features can be considered, ranging from signal amplitude in the temporal domain (e.g. Rieger et al., 2008) to power or phase information in the frequency domain (Kerlin et al., 2010; Luo and Poeppel, 2007; Rieger et al., 2008). Specific transformations, such as wavelet coefficients (Åberg and Wessberg, 2007; Rieger et al., 2008), and coherence measures (Besserve et al., 2007) can also be used. Furthermore,

* Corresponding author at: Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Fax: +31 43 3884125.
E-mail addresses: lars.hausfeld@maastrichtuniversity.nl (L. Hausfeld), f.demartino@maastrichtuniversity.nl (F. De Martino), m.bonte@maastrichtuniversity.nl (M. Bonte), e.formisano@maastrichtuniversity.nl (E. Formisano).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.11.056

features can be differently *grouped* in the (spectral-) temporal and spatial domain. For example, limiting the information to pre-defined temporal windows of interest is essential to many realizations of EEG-based *brain-computer interface* (BCI) systems (e.g. Birbaumer, 2006; Blankertz et al., 2011; Wolpaw et al., 2002). Alternatively, the information contained in a sliding time interval of EEG data can be used, e.g. to detect the occurrence of seizures in epileptic subjects (Schad et al., 2008). Concerning the spatial (channel) domain, many BCI systems employed spatial filters (i.e. linear combinations of channels; see Blankertz et al., 2011) to enhance performances. For the same reason sophisticated feature selection or reduction methods were applied in BCI systems (see Bashashati et al., 2007).

Several machine-learning algorithms have been used to learn the relation between selected features of the EEG/MEG data and experimental labels. These algorithms include simple correlation (e.g. Luo and Poeppel, 2007), support vector machines (SVMs) (Vapnik, 1995), linear discriminant analysis (LDA) (e.g. Duda et al., 2001), and neural networks or Bayesian approaches (Bishop, 2007). Most frequently, learning algorithms are based upon linear models (e.g. Lotte et al., 2007; Rieger et al., 2008; van Gerven et al., 2009) due to their fast computation, robustness and simplicity of results interpretation.

To determine the generalization ability of the computed model, an independent set of test data is required. This can be done at single-subject level, splitting the measured data into training and testing sets (e.g. Luo and Poeppel, 2007) or across subjects, using a subset of subjects for training and the other for evaluating the generalization performance (e.g. Kerlin et al., 2010).

In this study, we consider and evaluate the effects of differently combining and grouping the features in the temporal (*predefined windows*, *shifting window*, *whole trial*) and channel domain (*single channel*, *multichannel*) in the context of a neuro-cognitive EEG paradigm. Using *Gaussian Naïve Bayes* (GNB; Mitchell, 1997) classification, we analyze data from an auditory EEG study aimed at understanding the task dependence of the cortical mechanisms underlying the processing of voice and speech identification (Bonte et al., 2009) and illustrate the results of each possible feature combination in the temporal and channel domain.

Materials and methods

Machine-learning approaches for the analysis of neuroimaging data require single trials to be described by an *n*-dimensional vector of features. In our approach, basic features are defined as EEG voltages and include time (samples) and measurement channels (electrodes). In particular, we consider six types of classification analyses derived from combining three types of features grouping in the temporal domain (*predefined windows*, *shifting windows*, *whole trial)* with two approaches to handle the channel dimensions (*single channel*, *multichannel*, see Fig. 1). These different types of analyses can be combined with any classification algorithm (e.g. LDA classifier or SVMs). Here, we use a modified *Gaussian Naïve Bayes classification,* because of its simplicity which implies lower computational costs (e.g. compared to SVM classification) and interpretability of model parameters. We examine the case of pairwise classifications of EEG responses to simple vowels (/a/, /i/, /u/) spoken by three speakers (sp1, sp2, sp3) (see EEG experiment and data section).

Predefined windows

In the first approach, we use prior hypotheses (e.g. typical ERP windows) to select the temporal windows entering the analysis. As depicted in Fig. 1.a, the temporal samples within a specific interval are used as features to classify single trials either for each of the *K* channels (right upper panel) or for all channels (right lower panel). In the latter case, the feature set is defined by concatenating sampling points of multiple channels. In the case of a channel-by-channel analysis accuracy values are obtained for each electrode. This allows creating a topographic map of classification performance for the predefined intervals. Classifying based on features from multiple channels results in one classification accuracy value. In this case, a topographic map is created from the weights estimated during model training (see Eq. (4)) that indicate the relevance of each electrode contribution to the classification.

Shifting windows

In the second approach (Fig. 1b), the analyses are not restricted to specific latencies and are based upon features from *shifting windows* either on a channel-by-channel basis (right upper panel) or by concatenating features from multiple channels (right lower panel). Results of the single-channel approach can be depicted as a time series of topographic plots indicating classification performance.

The multi-channel classification allows retrieving the information content over time (information time-course). A weight vector – indicating the relevance of individual channels – is obtained for each time window.

Whole trial period

In the third temporal approach (Fig. 1.c) all temporal samples within a trial period are used. Classifications are performed either using the channel-wise (right upper panel) or multi-channel (right lower panel) approach. Results for the channel-wise approach may be used to create a topographic map of the information content within the entire trial period. For the multichannel approach, the analysis returns an overall accuracy value. Weights are defined for each sampling point and channel and thus indicate the temporal and topographical variation of the information content.

Gaussian Naïve Bayes classification

We report below a short description of GNB classification with reference to EEG data; see Mitchell (1997), for a more complete and general formulation of this algorithm.

Let us consider a supervised learning problem in which we wish to approximate the function $f: X \rightarrow C$ or equivalently $P(C|X)$, where C is a Boolean random variable representing the categories in our classification problem and $X = \langle x_1, \ldots, x_n \rangle$ is a n-dimensional feature vector obtained from the EEG data. Using Bayes rule we can write:

$$P(C = c_m|X) = \frac{P(X|C = c_m)P(C = c_m)}{\sum_j P(X|C = c_j)P\left(C = c_j\right)} \quad (1)$$

where c_m represents the mth category. One way to learn $P(C|X)$ is to use the training data to estimate $P(X|C)$ and $P(C)$ and then use Eq. (1) to classify any new instance of X.

The Naïve term is introduced when in the estimation of $P(X|C)$ the n features are assumed to be conditionally independent and Eq. (1) can be written as:

$$P(C = c_m|X) = \frac{\prod_{i=1}^{n} P(x_i|C = c_m)P(C = c_m)}{\sum_j \prod_{i=1}^{n} P(x_i|C = c_j)P\left(C = c_j\right)}. \quad (2)$$

Following Eq. (2) and having estimated $P(x_i|C)$ and $P(C)$ from the training data, any new EEG trial $Y_{new} = \langle y_1, \ldots, y_n \rangle$ can be classified following:

$$C \leftarrow \arg\max_{C_m} P(C = c_m) \prod_{i=1}^{n} P(y_i|C = c_m), \quad (3)$$

where argmax returns the class (c_m) with highest probability given Y_{new}. In spite of the naïve assumption, the GNB was shown to perform well for many examples of neuroimaging datasets and to be fast and robust (e.g. Pereira et al., 2009). To solve the multi-class classification problem we used a one-versus-one approach which reduced the problem to a series of binary (2-class) classifications (see EEG data classification section). We furthermore assumed equal covariance matrices of the two classes in the estimation of $P(x_i|C)$. This allowed us to pool the training data set of the two classes in the estimation of the variance of the classes. In order to derive the relative importance of features in the classification problems we estimated weights for each of the n features as:

$$w_i = \frac{1}{\sigma_i^2}\left(\mu_i^+ - \mu_i^-\right) \quad (4)$$

where σ_i^2 represents the estimated variance and μ_i^+ (μ_i^-) represent the mean of the two classes (+,−) for the *i*th feature (Pereira et al., 2009). For visualization purposes, weights were further transformed by a ranking procedure (values ranged from 1 to 100 with 1 representing the lowest and 100 the highest weight).

EEG experiment and data

We illustrate the different types of classification analyses in the context of a recent auditory EEG study aimed at understanding the timing and mechanisms of cortical processing of voice and speech (Bonte et al., 2009). For reader's convenience, essential information on experimental design, EEG measurements and data pre-processing are reported below. A more detailed description can be found in Bonte et al. (2009).

Participants

Fourteen Dutch undergraduate students (8 female; 1 left-handed) took part in this study. No history of hearing losses or neurological abnormalities was reported. Participants gave their informed consent and received course credits or payment for participation. The study was approved by the Ethical Committee of the Faculty of Psychology at the University of Maastricht.

Stimuli

Stimuli were speech sounds of three Dutch vowels (/a/, /i/, and /u/) uttered by three native Dutch speakers (sp1: female, sp2: male, sp3: female). To introduce acoustic variability, for each vowel and each speaker three different tokens were recorded. Stimulus length was equalized to 230 ms. Sound intensity levels were equalized by matching RMS values. For analysis, stimuli were either grouped according to speaker identity (*speaker grouping*) ignoring the vowel dimension or according to vowels (*vowel grouping*) ignoring the speaker dimension.

Experimental design and procedure

Task dependent processing was induced by introducing one-back tasks on either speaker or vowel identity (*speaker and vowel task*). A passive task denoting passive listening of the stimuli was also included but not used in our analyses. For the active tasks, subjects were instructed to respond with a button press every time that the same vowel (vowel task) or the same speaker (speaker task) was presented in two subsequent trials (target trials), which occurred in 6.25% of all trials. Trials including targets, and/or button responses (correct responses, omissions, false positives) were not included in the analysis. Each task involved two blocks amounting to a total of 450 non-target trials. Stimulus onset asynchrony varied between 3.0 and 3.5 s. All subjects participated for two EEG sessions and performed either two passive blocks followed by two speaker task blocks or two passive blocks and two vowel task blocks. The order of sessions was counterbalanced across subjects. Before the speaker and vowel tasks a short practice session assured that participants understood the task.

EEG recording and preprocessing

Data were recorded (sampling rate: 250 Hz) in an electrically shielded and sound attenuating room from 61 equidistant electrode positions (Easycap, Montage No.10) relative to left mastoid reference. Impedance levels were kept below 5 kΩ. Artifacts were removed in two steps. First, artifacts like high-amplitude, high frequency muscle noise, swallowing, or electrode cable movements were rejected. Second, eye-blinks, eye movements, heartbeat effects were corrected by using ICA as implemented in EEGlab (Delorme and Makeig, 2004; Makeig et al., 2002). For each task, ICA components were decomposed into brain-related activity and non-brain artifacts by visual inspection. Electrode signals were recreated by using all brain-related components (speaker task: 24 ± 4 components; vowel task: 23 ± 4; passive task: 20 ± 4) and baseline corrected (1 s before stimulus onset) (see Bonte et al., 2009). Finally, signals were recomputed using the average reference.

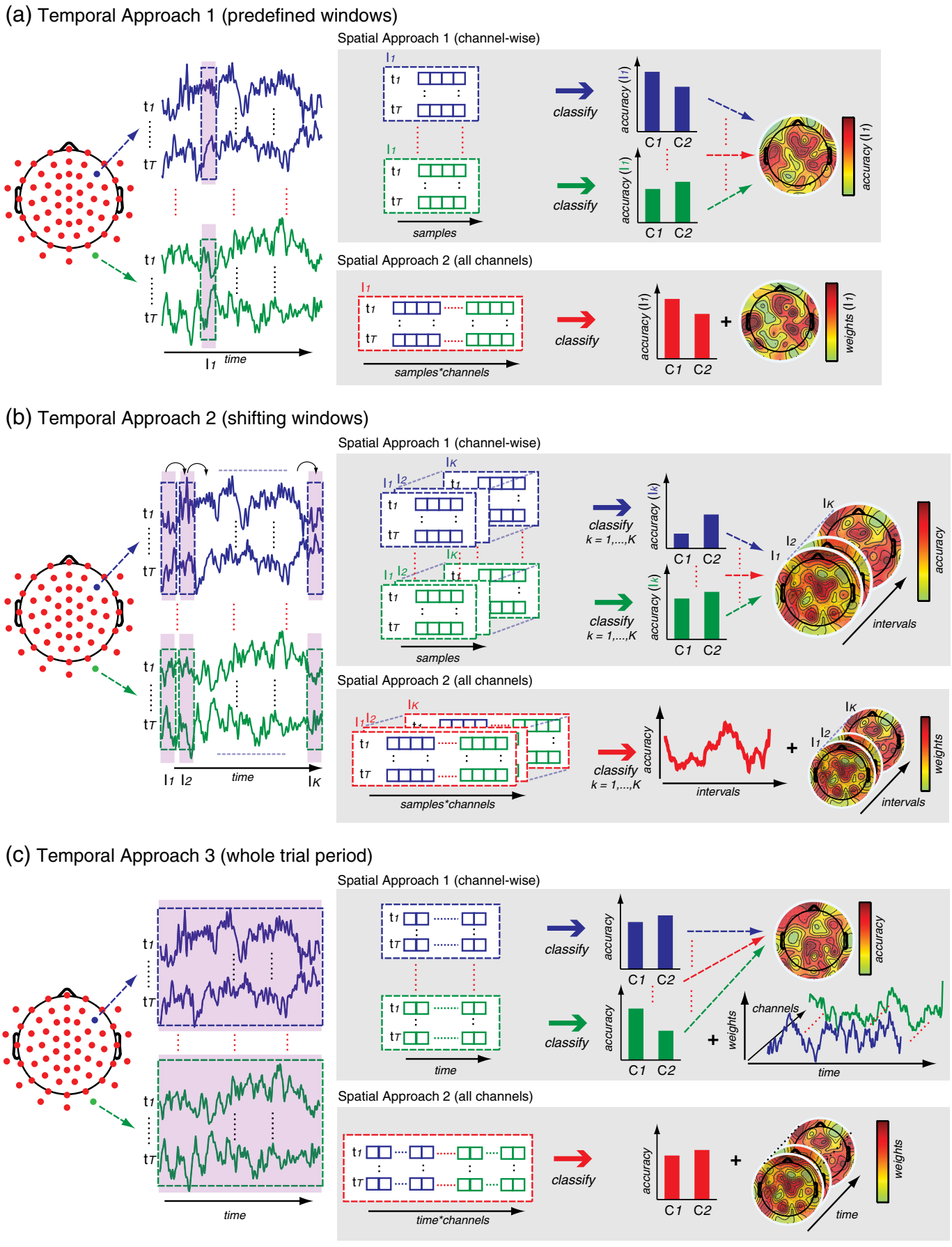
EEG data classification

For all different types of classification we followed a 20-fold cross-validation procedure by assigning randomly selected trials to non-overlapping training ($Train_l$; $l = 1,\ldots,20$) and testing sets ($Test_l$). In order to prevent model learning to be affected by the number of training examples, we made use of a leave-in procedure (i.e. resulting in a constant number of training trials). For each iteration l the training set $Train_l$ consisted of 30 trials per condition whereas the amount of trials in $Test_l$ varied (~15) due to trial rejection. Three binary comparisons were performed for each grouping (i.e. Speaker Grouping: sp1 vs. sp2, sp1 vs. sp3, sp2 vs. sp3; Vowel Grouping: /a/ vs. /i/, /a/ vs. /u/, /i/ vs. /u/).

To evaluate classification performance, we computed the accuracy of predicting class labels for the independent test set for each binary comparison. Accuracy was defined as the percentage of correctly classified trials.

Classification performances and feature weights were averaged over the 20 folds. Accuracies and weights for the speaker and vowel grouping were obtained by averaging results of the respective three binary comparisons. Significance of classification accuracies on individual subject level was obtained with permutation testing (Golland and Fischl, 2003). The empirical null distribution was derived for each classification strategy and subject by repeating the whole classification one thousand times with permuted labels of trials in $Train_l$. In the case of *shifting window* analysis, we made use of the permutation distribution obtained for the *predefined windows* approach to avoid massive computations by computing permutations for each window. We assessed the significance for each channel and window using the channel's most conservative chance level estimation of the five windows examined in the *predefined windows* approach for the respective task and grouping.

At group level, we calculated the significance of classification using a binomial test (e.g. Darlington and Hayes, 2000) with $n = 14$ (number of subjects), $p = 0.05$, and k expressing the number of subjects with a significant ($p < 0.05$) classification performance according to individual permutation tests. Differences between stimulus groupings [*speaker grouping–vowel grouping*] were examined by paired t-tests on the respective classification accuracies for each task (*grouping effect*). For visualization of scalp topographies, only significant channels with at least one significant neighboring channel were considered (i.e. significant but isolated channels were not displayed).


(a) Temporal Approach 1 (predefined windows)
Spatial Approach 1 (channel-wise)
classify
accuracy (I1)
C1 C2
samples
Spatial Approach 2 (all channels)
samples*channels
weights (I1)
time
(b) Temporal Approach 2 (shifting windows)
Spatial Approach 1 (channel-wise)
classify
k = 1,...,K
accuracy (Ik)
intervals
accuracy
Spatial Approach 2 (all channels)
samples*channels
weights
(c) Temporal Approach 3 (whole trial period)
Spatial Approach 1 (channel-wise)
channels
Spatial Approach 2 (all channels)
time*channels

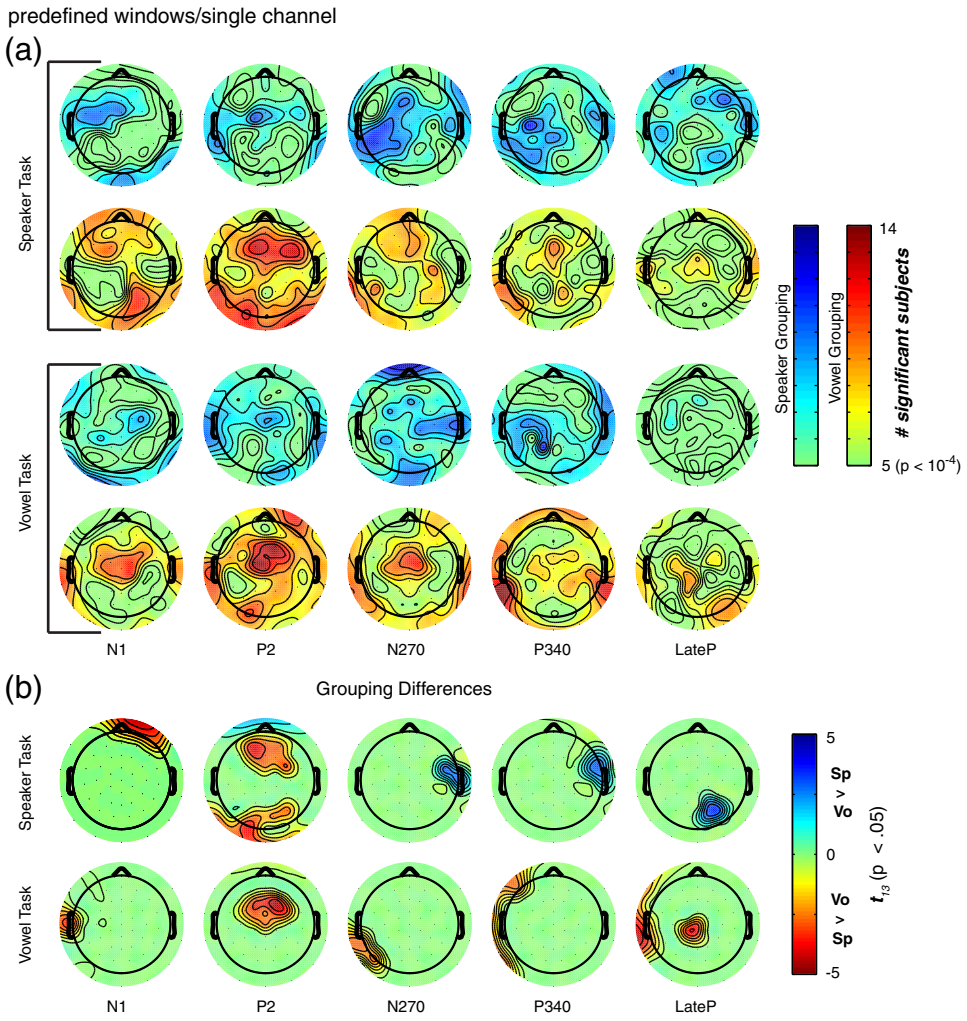


Fig. 2. Results of the predefined windows/single channel analysis. (a) Single channel classification performances are presented – for each interval, task and grouping – by scalp topographies depicting the number of subjects with significant classification accuracy (see text). Values for speaker and vowel grouping are depicted in blue and red colors, respectively. (b) Higher classification performance for speaker vs vowel grouping are indicated by blue and red colors for each channel. Tests were restricted to channels with a significant accuracy for one of the groupings. For visualization of scalp topographies, only channels with at least one significant neighboring channel were considered.

Parameters for feature extraction and grouping

For the analysis in pre-defined time intervals (see Predefined windows section), we selected five intervals of 60 ms (i.e. N1: 80–140 ms; P2: 170–230 ms; N270: 240–300 ms; P340: 310–370 ms; LateP: 500–560 ms) based on results from Bonte et al. (2009) that consisted of 15–16 time samples. Classification was performed following two different strategies: 1) separately for each participant, channel and window (the feature set reduced to either 15 or 16 values per trial); 2) considering all channels together for each subject and window (leading to 915 or 976 features for each trial). In both cases we obtained classification accuracies for each subject and interval. The relevance of a single channel was either accessed by its performance (single channel classification) or averaging feature weights (multichannel approach).

For the classification analysis with *shifting windows* (see Shifting windows section), we selected a window length = 60 ms, a sliding step = 10 ms and a trial period from −250 to 810 ms. Finally, the same temporal interval (−250 ms–810 ms) was used for the *whole trial*-based classification (see Whole trial period section). Both classifications were performed considering either single channels or multiple channels. Note that pre-stimulus data was included to compare results (classification performance and feature weights) of time windows containing no information to informative ones.

Results

Predefined windows

We first considered the classifications of speakers and vowels in five predefined temporal intervals (N1, P2, N270, P340, LateP). Fig. 2.a shows – for the single channel case – group classification results for *speaker* and *vowel grouping* during the *speaker* (top panels) and *vowel*

Fig. 1. Overview of the six different types of classification considered in this study. Different selections and groupings of data in the spatial and temporal dimension result in different types of classification. (a) *Temporal Approach 1*: Classifications are performed using signal amplitudes within predefined windows of interest, e.g. I_1, which need to be specified based on prior hypotheses. (b) *Temporal Approach 2:* Using *K* shifting windows (I_1,..., I_K), separate classifications are performed, which results in a time-course of the information content (accuracies) (c) *Temporal Approach 3:* The overall information content within a trial is estimated using a single classification that employs all samples within the trial period. In addition to the different types of temporal grouping, the spatial dimension can be accounted for either by performing separate classifications at each channel (*Spatial Approach 1, left panels*) or concatenating all channels (*Spatial Approach 2, right panels*), which results in a single classification. t_1, ..., t_T denote trials of EEG signals, measured at each recording channel. See text for detailed information.

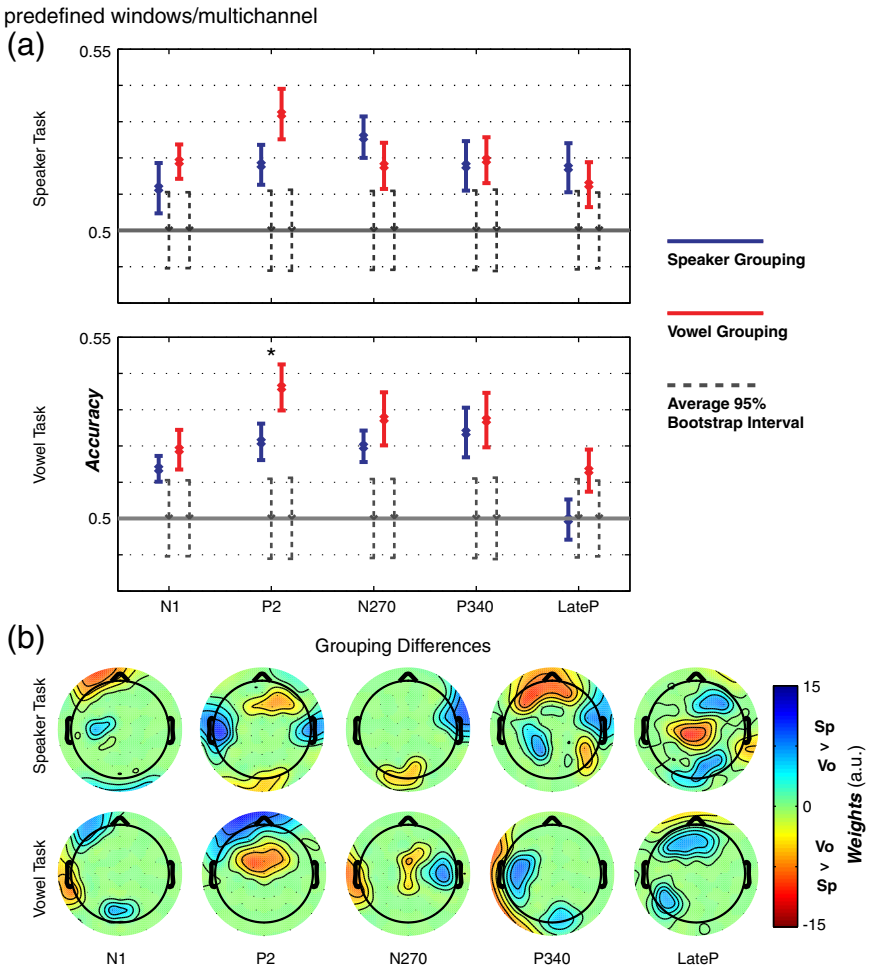


Fig. 3. Results of the predefined windows/multichannel analysis. (a) Average classification accuracies are shown for tasks and groupings (red: speaker grouping; blue: vowel grouping) for each window separately. Bars denote average accuracies and SEM. Grey bars show average 95% CIs for individual permutation tests. (b) Weight differences of speaker and vowel grouping for each time window and task are presented in the lower panel. Note the similarities to the results for the corresponding single channel analysis (see Fig. 2.b). Only channels with at least one significant neighboring channel were considered for visualization of scalp topographies.

task (lower panels), respectively. To estimate reproducibility across subjects, we created topographic maps depicting, at each channel, the number of subjects with a significant classification performance. For each subject, significance was assessed channel-by-channel by permutation testing and corrected to account for multiple testing [no. of channels] using false discovery rate (FDR [Benjamini and Hochberg, 1995], $q<0.05$).

Differences between *speaker* and *vowel groupings* are depicted in Fig. 2.b for the two tasks separately. The N1 and P2 topographic maps included several channels showing higher classification performance for the *vowel grouping* during both tasks, with early left lateralization n in the case of the *vowel task*. The later intervals N270, P340 and LateP were characterized by better classification performances for the dimension relevant for the task. During the *speaker task* better speaker discrimination was observed for right temporo-parietal (N270, P340) and right occipito-parietal channels (LateP). Higher accuracy values for the vowel grouping during the *vowel task* were found at left lateral (N270, P340, LateP) and parietal channels (LateP).

Fig. 3 shows results obtained when features extracted from all channels were employed. Group averaged accuracy values for *speaker* and *vowel groupings* and both tasks are presented in Fig. 3.a together with the average 95% confidence intervals, resulting from permutation tests at single-subject level. Corresponding weight differences between groupings are presented as topographic maps in Fig. 3.b for each of the two tasks. Classification performances for the two groupings and tasks for most of the windows were small but above chance. Largest average accuracies were found in the P2 interval for the classification of vowels both during the *speaker* and *vowel task*. Within this window, a significantly higher accuracy was observed during the *vowel task* in the classification of vowels compared to the classification of speakers (paired *t*-test, $p=0.026$). For both tasks the topographies of weight differences were comparable to single channel accuracy differences (Fig. 2.b) but possessed additional channels being more relevant during one of the groupings especially for the task irrelevant dimension (i.e. during the speaker task for vowel grouping and vice-versa).

Shifting windows

To obtain a detailed temporal profile of speaker and vowel discrimination we conducted classification analyses using *shifting windows*. Fig. 4.a shows the results for the *shifting window/single channel* analysis. For display, different channels are arranged along the y-axis of the plot; blue and red color-coding denotes significant differences between speaker and vowel grouping ($p<0.05$, uncorrected). In addition, topographic plots of accuracy differences (*speaker grouping–vowel grouping*) are shown below for relevant latencies. Statistical tests and color-coding were limited to channels and intervals with speaker and/or

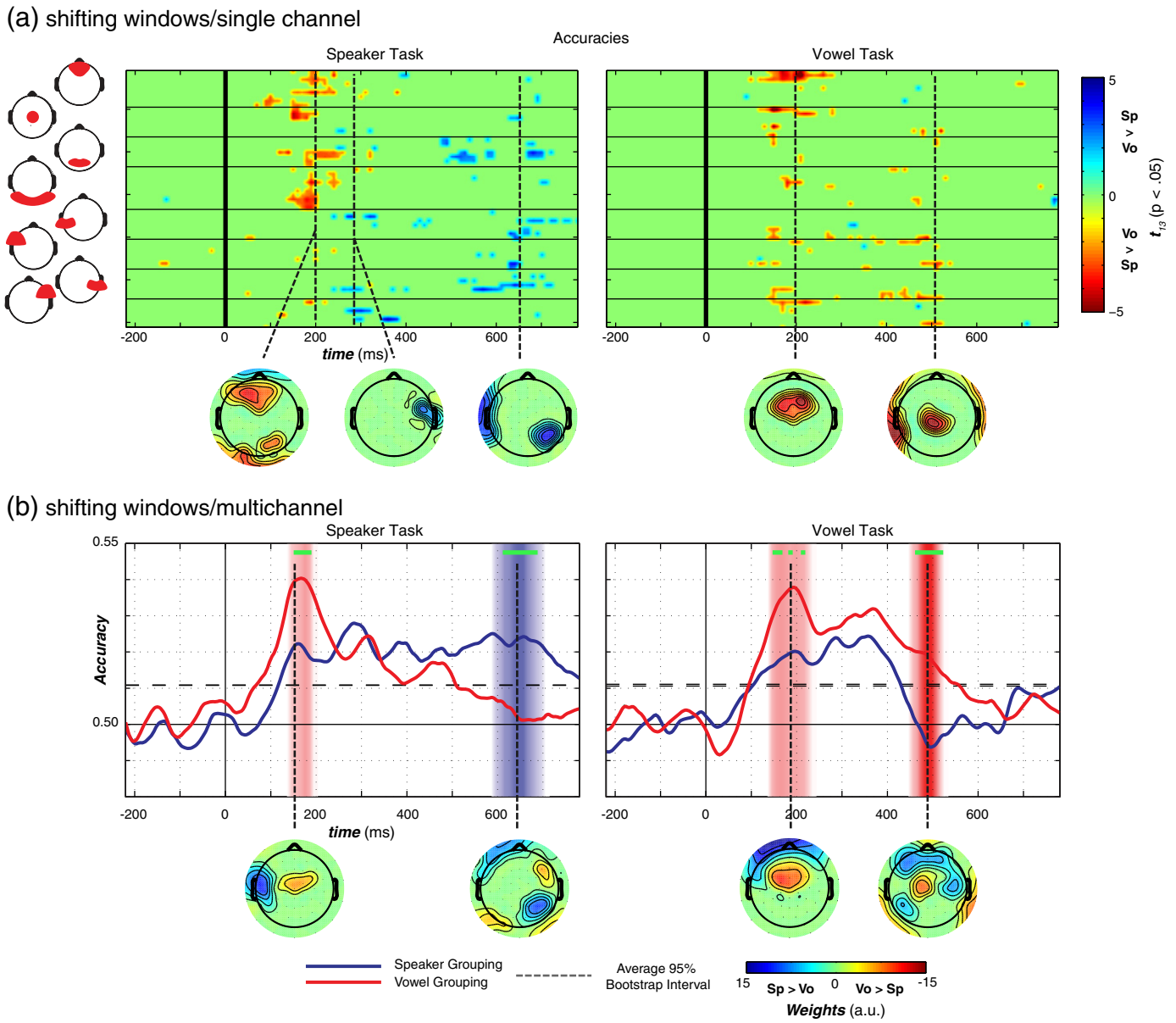


Fig. 4. Results of shifting window classifications. (a) Single channel analysis: accuracy differences between speaker and vowel grouping are presented for the speaker (left) and vowel task (right panel). For better visualization scalp topographies of accuracy differences at specific latencies indicated by dashed lines are shown below. (b) Multichannel analysis: Classification performances for shifting windows are shown for the speaker (left) and vowel task (right panel). Average accuracy for both speaker (blue) and vowel (red) grouping are depicted (blue and red shadings denote significance of higher speaker and vowel grouping, respectively). Green lines indicate grouping differences with FDR-corrected ($q<0.1$) significance levels. Horizontal dashed lines depict the most conservative estimation of the chance level in the *predefined windows/multichannel* analysis. Topographic plots show weight differences between the speaker and vowel task at latencies with significant performance differences. For visualization of scalp topographies, only channels with at least one significant neighboring channel were considered.

vowel classification performance above chance level (i.e. exceeding the most conservative 95% confidence interval in the previous analysis). During the *speaker task* enhanced classification accuracies for vowels were observed between 150 and 240 ms (frontal, central, posterior channels at [150–200 ms]; frontal, parietal channels at [200–240 ms]). At [230–400 ms] and [500–730 ms] higher classification accuracies for speakers were found. Right temporo-parietal channels discriminated better between speakers during the medium latencies interval and posterior and left lateral channels showed this effect during the later intervals. Accuracy differences for the *vowel task* were characterized by enhanced vowel discrimination at two intervals ([120–230 ms], [450–550 ms]). During the early interval vowels were better classified at central and frontal channels. Central, left temporal and lateral channels classified vowels better than speakers during the late interval.

Using a *multichannel* approach we extracted the overall information content over time. Fig. 4.b visualizes classification performance of the *speaker* and *vowel grouping* as a function of time for the speaker (right) and vowel task (left). Accuracies were defined to be above chance when the respective most conservative 95% confidence interval in the *predefined windows/multichannel* analysis was exceeded (see above). Analyses of task differences were limited to intervals with classification performance above chance for at least one grouping. Shadings denote latencies that showed significant differences between *speaker* and *vowel grouping* ($p<0.05$, uncorrected). Grouping differences that remain significant after correcting for multiple comparisons [FDR, $q<0.10$] are indicated by green lines. Enhanced classification of vowels compared to speakers can be noted for an early interval for both tasks (*speaker task* [190–240 ms], *vowel task* [150–240 ms]) with similar

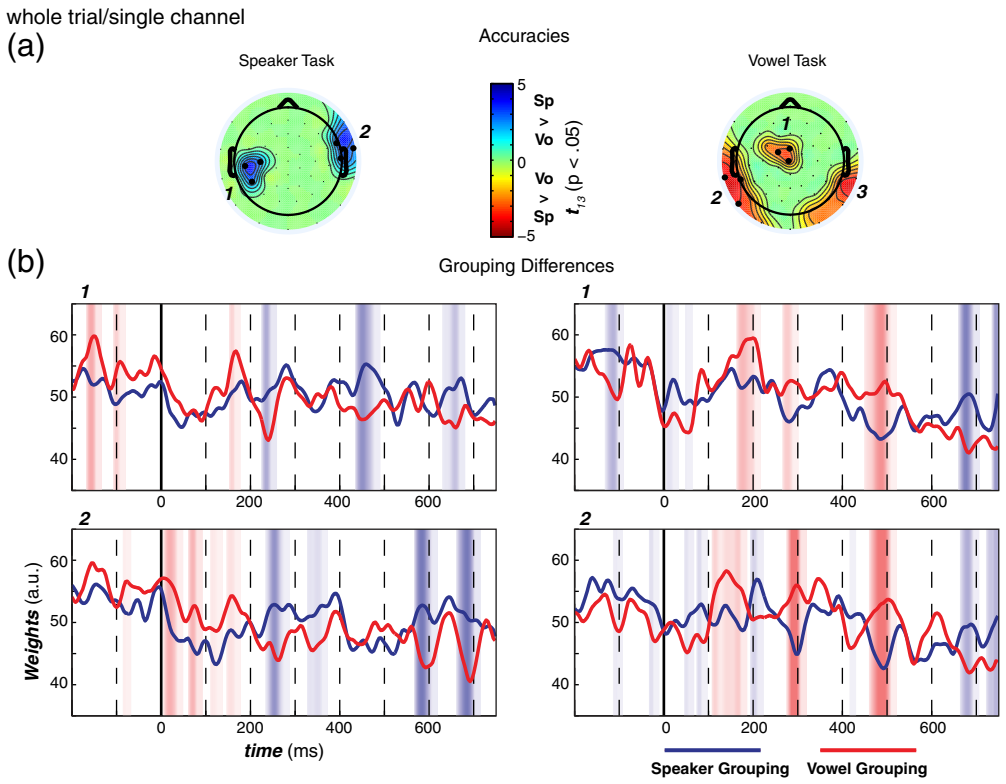


Fig. 5. Results of the whole trial/single channel analysis. (a) Channel-wise differences of classification performances for each task are shown in the upper panel. Tests were restricted to channels that had a significant accuracy value for one of the groupings. (b) For clusters showing significant performance differences in (a), average weight differences of speaker and vowel grouping are shown. Weights for speaker and vowel groupings are depicted in blue and red, respectively. Shadings denote larger speaker (red) and vowel (blue) weights. Only channels with at least one significant neighboring channel were considered for visualization of scalp topographies.

topographies of weight differences. Task dependent effects as shown by higher speaker classification during the *speaker task* and higher vowel classification during the *vowel task* are observed at late intervals (*speaker task* [580–700 ms], *vowel task* [480–550 ms]). For the *speaker task* the enhanced speaker discrimination was accompanied by higher weights in right parietal channels whereas the enhanced vowel discrimination during the *vowel task* was characterized by higher weights at central and lateral channels.

Whole trial period

Next, we considered all samples within the trial period and performed *single channel* and *multichannel* classifications. Results for the single channel analysis are depicted in Fig. 5.a by means of topographic maps of accuracy differences between *speaker* and *vowel grouping* ($p<0.05$, uncorrected) for both tasks. Maps were restricted to channels that were significant for at least one grouping (FDR-corrected, $q<0.05$). Fig. 5.b shows – for selected channel clusters – the temporal profile of the weights resulting for the classification of speakers and vowel. Time intervals with high values for the weights are those mostly contributing to the classification. Accuracy differences during the *speaker task* showed a left parieto-temporal and a right temporal cluster with enhanced classification performance for speakers. Weight differences for these two clusters were found to be larger for speakers at an early interval [220–260 ms] for both clusters (Fig. 5.b). A later interval [440–480 ms] showed larger weights for the *speaker grouping* for the left parieto-temporal cluster. After ~570 ms larger weights for speakers were observed for both clusters but differences were more pronounced within the right lateral cluster. For the *vowel task* three clusters (a central cluster and both a left and right posterior lateral cluster) showed enhanced accuracies for vowels compared to speakers. Early intervals showing higher weights for vowels were found for two clusters (central at [160–210 ms]; left lateral at [100–190 ms]). For later intervals at [260–300 ms] and [450–500 ms] higher weights for the vowel grouping were observed for the central and left lateral clusters (for the right lateral cluster higher weights for vowels were found at [380–500 ms]; results not shown).

Finally, classifications were performed by employing the full spatio-temporal set of features. Accuracy values for the two tasks and effects (top panel) and corresponding weight differences of selected channels between the two groupings (lower panel) are shown in Fig. 6. For each task both types of groupings were above chance level ($p<10^{-11}$, for all task by grouping combination). When comparing speaker and vowel groupings, larger classification performances could be found for the vowel grouping during the vowel task ($p<0.001$) but not during the speaker task, during which there was a trend for enhanced speaker classification (Fig. 6.a). Weight differences between speaker and vowel groupings (Fig. 6.b) revealed similar results compared to the accuracies obtained in the *shifting windows/single channel* classification with some differences.

In sum, outcomes of the *single channel* analysis for the *whole trial period* produced maps revealing the spatial distribution of classification differences between speaker and vowel groupings. These were characterized by higher classification performances for the task-relevant stimulus dimension (i.e. grouping) compared to the dimension that was not relevant for the task. A similar task-dependent effect was also found when the whole set of data was analyzed by means of the *multichannel* analysis.

Discussion

Pattern recognition and EEG data

We have illustrated different strategies for analyzing EEG data using a pattern recognition algorithm. We have shown that it is feasible to

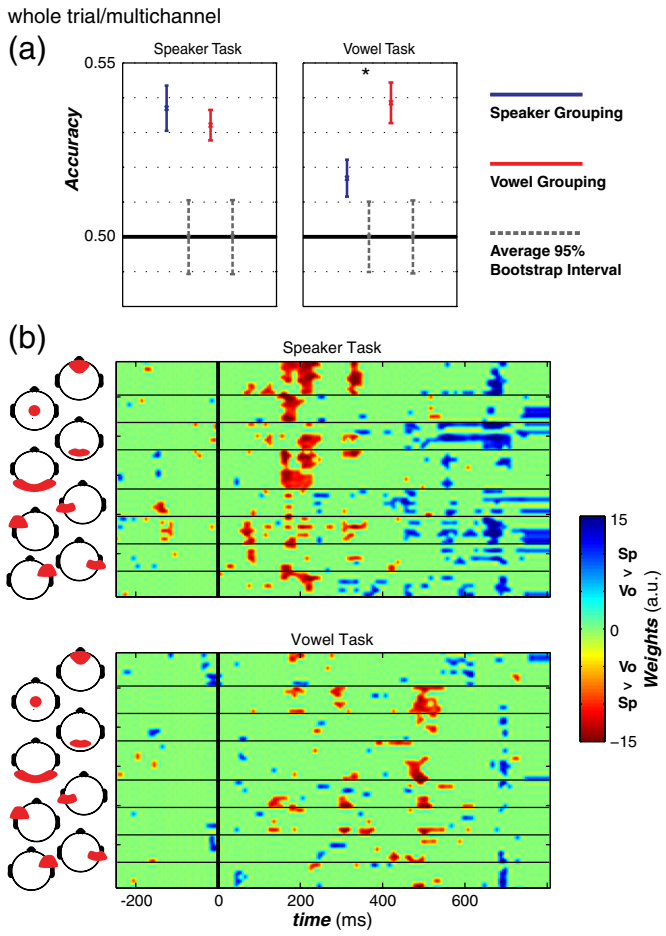


Fig. 6. Results of the whole trial/multichannel analysis. (a) Classification performances using all sampling points and channels are presented for the speaker (left) and vowel task (right panel). Bars denote the average and SEM of classification performance across subjects. (b) Weight differences for all channels are presented by blue and red colors indicating larger weights for the speaker and vowel grouping, respectively.

distinguish experimental conditions above chance level, at the fine-grained level of speaker and vowel identity. Although low, our single-trial classification accuracies were significant even at a single subject level, which indicate that – although noisy – EEG single trial responses carry information on the neural processing of individual speech sounds. Significance was assessed with a resampling approach (permutation testing) that detects systematic classification biases and provides an empirical estimation of the chance level. In particular, our classification performances were lower than those typically reported in EEG-based BCI experiments (e.g. ~70–90% for motor imagery based BCIs Lotte et al., 2007). There may be several reasons for this. First, our experimental paradigm included stimuli and conditions that result in largely similar EEG signals (low CNR) compared to BCI paradigms that are constructed to maximize response differences between classes. Another reason may be the data processing scheme prior to classification. Compared to our approach, analyses in BCI experiments often employ more sophisticated preprocessing and feature selection techniques (e.g. Laplacian filtering, common spatial patterns, genetic algorithms Bashashati et al., 2007; van Gerven et al., 2009). In this study, preprocessing of data included standard filtering and ICA but no further feature selection and enhancement. Univariate or multivariate feature selection algorithms – which may lead to considerable increases of accuracy values especially in high-dimensional cases – were not applied. Especially in cases of classifications with many features *(whole-trial/multichannel)*, wrapper methods such as Recursive Feature Elimination (De Martino et al., 2008; Guyon et al., 2002) should be beneficial for both to obtain higher accuracies and select informative features.

With regard to the classification algorithm, we selected a GNB classifier that may be seen as a 'pseudo' multivariate approach, which has the advantage of providing interpretable weights (similar to *t* statistics) as it assumes independency among features (i.e. diagonal covariance matrix). Other machine-learning techniques such as linear discriminant analysis (e.g. Duda et al., 2001) and support vector machines (Vapnik, 1995), which take feature correlations into account, have been previously applied to EEG datasets in the context of BCI (e.g. Bashashati et al., 2007; Lotte et al., 2007). The use of these or other classifiers may lead to higher accuracies compared to GNB-based classification, but this may come at the cost of the interpretability of the results and increase of computation time.

In this study, trials were classified based on EEG time courses. However, representing trials by means of event-related (de)synchronization (Pfurtscheller and Lopes da Silva, 1999) or measures of coherence and synchrony (e.g. Besserve et al., 2007; Bonte et al., 2009; Varela et al., 2001) and classifying those may provide complementary and more detailed information (e.g. phase estimates, band-pass filtered signals and wavelet coefficients have been employed by Luo and Poeppel (2007), Kerlin et al. (2010), and Rieger et al. (2008), respectively). Finally, our analyses could only detect effects that were strictly time-locked to the stimulus. Thus, single trials of the same condition that differed in terms of latencies could not be accurately classified. As this may be a relevant aspect in EEG/MEG, it is desirable – for future extensions of the proposed method – to include classification schemes that account for possible latency differences across trials.

Types of classification

The focus of the present study was on examining how grouping of features in the temporal and spatial (channel) domain influences the results of EEG classification analyses (see Table 1). With regard to the temporal domain, a choice needs to be made between a hypothesis-driven analysis limited to a few temporal windows of interest (i.e. *predefined windows approach* which is closest to conventional ERP analyses) and a data-driven analysis with feature sets consisting of signal amplitude at all time points within a trial. This *whole-trial* approach promises to be more sensitive as several ERP components may contribute to distinguishing between two experimental conditions (Blankertz

Table 1
Qualitative comparison of classification approaches.

	Predefined windows		Shifting windows		Whole trial	
	Single channel	Multiple channels	Single channel	Multiple channels	Single channel	Multiple channels
A-priori selection	Yes	Yes	No	No	No	No
Time course	–	–	Accuracies	Accuracies	Weights	Weights
Topographies	Accuracies	Weights	Accuracies	Weights	Accuracies	Weights
Amount of tests	#Channels by #windows (305)	#Windows (5)	#Channels by #windows (6161)	#Windows (101)	#Channels (61)	1 (1)

The numbers in brackets denote the amount of statistical tests required in this particular study with 61 channels, 5 predefined time windows, and 101 shifting windows.

et al., 2011). As a possible alternative, we also examined a shifting-window approach with multiple sequential classifications that – compared to the whole trial approach – possesses the advantage of assessing information content over time. Regarding the spatial domain, the choice is between performing multiple classifications channel-by-channel and a single classification using all channels simultaneously.

In general, a multichannel and whole-trial approach seems desirable as it does not rely on previous assumptions and enables the pattern recognition algorithm to fully exploit information contained in both the topographic and temporal distribution of signal amplitudes. Furthermore, such a data-driven approach relies on a single classification, thus avoiding the problem of multiple comparisons, which applies – at different extents – to all other combinations (see Table 1: *Amount of tests*).

Results of our analyses, however, highlighted several aspects that need to be considered when using this type of approach. When using all available features in a single classification (*whole-trial/multi-channel*), detection of both informative time windows and topographies is based on feature weights whereas a single accuracy value describes the overall information content. As illustrated in Fig. 6, this approach detected the general effect on accuracy of the *vowel task* for the classification of vowels compared to speakers (Fig. 6.a, right panel) but failed to detect the expected opposite modulation for the speaker task (Fig. 6.a, left panel). Furthermore, the interpretation (and statistical testing) of weights to derive topographical and temporal information is not straightforward (Fig. 6.b). Especially when the number of features is very large, estimates of weights may be noisy. In case of SVM or LDA — based classification, additional issues may arise. For instance, Blankertz et al. (2011) describe a hypothetical case where high weights (as determined by a LDA classifier) are associated with one channel that does not contribute any class-related information. At the cost of increasing the number of classifications, the number of features can also be reduced by using a *whole trial/single channel* approach (Fig. 5). The analysis resulted in neurophysiologically plausible accuracy-based topographic maps that clearly highlight the task dependence of the informative neuronal sources but could only roughly indicate which intervals are relevant (Fig. 5.b).

Our results for the *shifting window* approaches (Fig. 4) indicate that these are the most appropriate for tracing information content over time. In fact, both single and multi-channel analyses were able to detect – without prior hypotheses on the temporal windows – the early and task-independent processing of vowels, which becomes maximal at ~200 ms (corresponding to P2). This is in accordance with the idea that an early stimulus-driven analysis processes – by default – acoustic features which are informative of speech content, like first or second formant frequencies (e.g. Bonte et al., 2009; Obleser et al., 2004). Although less significantly, our results additionally show that speaker identity information is present at similar latencies indicating bottom-up processing of speaker-relevant acoustic features, like fundamental frequency and timbre (Belin et al., 2004; Bonte et al., 2009; Charest et al., 2009).

Later task dependent processing (~280 ms), expressed by enhanced classification performance (*single channel* analysis) or higher weights (*multichannel* approach) for the task-relevant dimension of stimuli, was found to occur mostly at right (speakers) or left lateralized (vowels) channels, which is in accordance with earlier studies (Belin and Zatorre, 2003; Formisano et al., 2008; Hickok and Poeppel, 2007; van Kriegstein and Giraud, 2004). Additionally, a late task-dependent effect between 450 and 700 ms after sound onset was detected that is most likely related to the memory maintenance of the relevant information for performing correctly the one-back task.

In the case of the *single channel/shifting window* analysis the amount of multiple testing is highest and statistical testing would require a proper correction. Using the Bonferroni approach is known to result in over-conservative corrections. A proper correction requires an empirical estimate of the likelihood that k consecutive windows are significant by chance, which in turn requires permutation testing for each channel and time window. The computational load for this is very high, however it is becoming tractable thanks to the increasing availability of parallel processing. The number of classifications is greatly reduced with a multiple channel approach, which thus seems the most viable choice for tracing the temporal profile of information content also because the number of features considered in each classification is not excessively large (corresponds to the number of channels). As a consequence of the reduced number of tests, early and late effects on speaker and vowel grouping could be still detected after correcting for multiple testing (FDR) for the *multichannel* but not the *single channel* approach (Fig. 4).

Conclusions

We have illustrated different ways of analyzing EEG data by means of a pattern classification algorithm. Outcomes of the analyses show that grouping or separating available features (channels, time windows) helps highlighting different aspects of information content in the data. Because of the high temporal resolution of EEG (and MEG) a shifting window approach with sequential multi-channel classifications proved to be the most valuable as it allows tracing the temporal evolution of stimulus and task-related neural information processing.

Acknowledgments

Financial support by the Netherlands Organization for Scientific Research, Innovative Research Incentives Scheme VENI Grant 451-07-002 (MB) and VIDI Grant 452-04-330 (EF) is gratefully acknowledged. We thank Giancarlo Valente for comments and discussions.

References

Åberg, M.C., Wessberg, J., 2007. Evolutionary optimization of classifiers and features for single-trial EEG Discrimination. Biomed. Eng. Online 6, 32. doi:10.1186/1475-925X-6-32.
Bashashati, A., Fatourechi, M., Ward, R.K., Birch, G.E., 2007. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. J. Neural Eng. 4 (2), R32–R57. doi:10.1088/1741-2560/4/2/R03.
Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14 (16), 2105–2109. doi:10.1097/01.wnr.0000091689.94870.85.
Belin, P., Fecteau, S., Bédard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn. Sci. 8 (3), 129–135. doi:10.1016/j.tics.2004.01.008.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57 (1), 289–300.
Besserve, M., Jerbi, K., Laurent, F., Baillet, S., Martinerie, J., Garnero, L., 2007. Classification methods for ongoing EEG and MEG signals. Biol. Res. 40. doi:10.4067/S0716-97602007000500005.
Birbaumer, N., 2006. Breaking the silence: brain–computer interfaces (BCI) for communication and motor control. Psychophysiology 43 (6), 517–532. doi:10.1111/j.1469-8986.2006.00456.x.
Bishop, C.M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics), 2nd ed. Springer, New York.
Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.-R., 2011. Single-trial analysis and classification of ERP components — a tutorial. NeuroImage 56 (2), 814–825. doi:10.1016/j.neuroimage.2010.06.048.
Bonte, M., Valente, G., Formisano, E., 2009. Dynamic and task-dependent encoding of speech and voice by phase reorganization of cortical oscillations. J. Neurosci. 29 (6), 1699–1706. doi:10.1523/JNEUROSCI.3694-08.2009.
Charest, I., Pernet, C.R., Rousselet, G.A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., et al., 2009. Electrophysiological evidence for an early processing of human voices. BMC Neurosci. 10, 127. doi:10.1186/1471-2202-10-127.
Darlington, R.B., Hayes, A.F., 2000. Combining independent p values: extensions of the Stouffer and binomial methods. Psychol. Methods 5 (4), 496–515. doi:10.1037/1082-989X.5.4.496.
Delorme, A., Makeig, S., 2004. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Meth. 134 (1), 9–21. doi:10.1016/j.jneumeth.2003.10.009.
De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43 (1), 44–58. doi:10.1016/j.neuroimage.2008.06.037.
Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification: Pattern Classification, 2nd ed. John Wiley & Sons, New York.
Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322 (5903), 970–973. doi:10.1126/science.1164318.
Golland, P., Fischl, B., 2003. Permutation tests for classification: towards statistical significance in image-based studies. In: Taylor, C., Noble, J. (Eds.), Information Processing in Medical

Imaging, Vol. 2732. Springer, Berlin/Heidelberg, pp. 330–341. doi:10.1007/978-3-540-45087-0_28.
Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422. doi:10.1023/A:1012487302797.
Haynes, J., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7 (7), 523–534. doi:10.1038/nrn1931.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8 (5), 393–402. doi:10.1038/nrn2113.
Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. 30 (2), 620–628. doi:10.1523/JNEUROSCI.3631-09.2010.
Kriegstein, K.V., Giraud, A., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage 22 (2), 948–955. doi:10.1016/j.neuroimage.2004.02.020.
Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. J. Neural Eng. 4 (2), R1–R13. doi:10.1088/1741-2560/4/2/R01.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54 (6), 1001–1010. doi:10.1016/j.neuron.2007.06.004.
Makeig, S., Westerfield, M., Jung, T., Enghoff, S., Townsend, J., Courchesne, E., Sejnowski, T.J., 2002. Dynamic brain sources of visual evoked responses. Science 295 (5555), 690–694. doi:10.1126/science.1066168.
Mitchell, T., 1997. Machine Learning, 1st ed. Mcgraw-Hill.
Obleser, J., Elbert, T., Eulitz, C., 2004. Attentional influences on functional mapping of speech sounds in human auditory cortex. BMC Neurosci. 5 (1), 24. doi:10.1186/1471-2202-5-24.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45 (1, Supplement 1), S199–S209. doi:10.1016/j.neuroimage.2008.11.007.
Pfurtscheller, G., Lopes da Silva, F.H., 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. 110 (11), 1842–1857. doi:10.1016/S1388-2457(99)00141-8.
Rieger, J.W., Reichert, C., Gegenfurtner, K.R., Noesselt, T., Braun, C., Heinze, H., Kruse, R., et al., 2008. Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. NeuroImage 42 (3), 1056–1068. doi:10.1016/j.neuroimage.2008.06.014.
Schad, A., Schindler, K., Schelter, B., Maiwald, T., Brandt, A., Timmer, J., Schulze-Bonhage, A., 2008. Application of a multivariate seizure detection and prediction method to non-invasive and intracranial long-term EEG recordings. Clin. Neurophysiol. 119 (1), 197–211. doi:10.1016/j.clinph.2007.09.130.
van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., et al., 2009. The brain–computer interface cycle. J. Neural Eng. 6 (4), 041001. doi:10.1088/1741-2560/6/4/041001.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory, 1st ed. Springer.
Varela, F., Lachaux, J., Rodriguez, E., Martinerie, J., 2001. The brainweb: phase synchronization and large-scale integration. Nat. Rev. Neurosci. 2 (4), 229–239. doi:10.1038/35067550.
Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M., 2002. Brain–computer interfaces for communication and control. Clin. Neurophysiol. 113 (6), 767–791. doi:10.1016/S1388-2457(02)00057-3.