International Journal of Neural Systems, Vol. 21, No. 1 (2011) 65–78

DOI: 10.1142/S0129065711002651

World Scientific
www.worldscientific.com

RECURRENCE-BASED ESTIMATION OF TIME-DISTORTION FUNCTIONS FOR ERP WAVEFORM RECONSTRUCTION

MATTHIAS IHRKE[*,†,§], HECKE SCHROBSDORFF[*,†,¶]
and J. MICHAEL HERRMANN[*,‡,‖]
[*]*Bernstein Center for Computational Neuroscience Göttingen*
[†]*MPI for Dynamics and Self-Organization*
Bunsenstraße 10, 37073, Göttingen, Germany
[‡]*Institute for Perception, Action and Behaviour*
University of Edinburgh, School of Informatics
10 Crichton Street, Edinburgh, EH8 9AB, UK
[§]*ihrke@nld.ds.mpg.de*
[¶]*hecke@nld.ds.mpg.de*
[‖]*michael.herrmann@ed.ac.uk*

We introduce an approach to compensate for temporal distortions of repeated measurements in event-related potential research. The algorithm uses a combination of methods from nonlinear time-series analysis and is based on the construction of pairwise registration functions from cross-recurrence plots of the phase-space representations of ERP signals. The globally optimal multiple-alignment path is approximated by hierarchical cluster analysis, i.e. by iteratively combining pairs of trials according to similarity. By the inclusion of context information in form of externally acquired time markers (e.g. reaction time) into a regularization scheme, the extracted warping functions can be guided near paths that are implied by the experimental procedure. All parameters occurring in the algorithm can be optimized based on the properties of the data and there is a broad regime of parameter configurations where the algorithm produces good results. Simulations on artificial data and the analysis of ERPs from a psychophysical study demonstrate the robustness and applicability of the algorithm.

Keywords: Event-related potential; dynamic time warping; cross-recurrence plot; line of synchrony; hierarchical clustering.

1. Introduction

Variability in processing speed and the ensuing variability in the timing of signal features is an important modulating effect in the analysis of EEG data.[7] This is particularly true in the study of *event-related potentials* (ERPs), where the timing and amplitude of signal features is compared over experimental conditions. ERPs are segments of electroencephalographic data that are locked to an external event. Usually, signals from multiple trials of the same experimental procedure are averaged in order to improve the poor signal-to-noise ratio of single trial data. There is a large body of research that relies exclusively on the interpretation of modulations of signal features in averaged ERPs and it is therefore crucial that the averaged potentials include all meaningful features. In the following, we will discuss shortcomings of the widely used pointwise average and propose an alternative estimator of the underlying signal.

Averaging procedures are based on statistical models of the data. We will start with a simple model that implies pointwise averaging and proceed to more complex models that follow from the phenomenology of ERPs and provide a more realistic description of the data. Assume we have N signals $s_i(t)$, $i = 1, \ldots, N$, measured in a time interval $\mathcal{T}$ (without loss of generality $\mathcal{T} = [0, 1]$). Assuming that the data s has been generated by a noisy and stationary process, the individual trials can be described by a

pointwise sum of the stationary signal and noise

$$s_i(t) = s_0(t) + \eta, \tag{1}$$

where the noise term η is independent across trials and of zero mean. It follows that the pointwise average

$$\hat{s}_0(t) = \langle s_i(t) \rangle_i := \frac{1}{N} \sum_{i=1}^{N} s_i(t)$$

of the individual trials results in an unbiased and optimal estimate in the mean-square error sense, since

$$\hat{s}_0(t) = \langle s_i(t) \rangle_i = \langle s_0(t) + \eta \rangle_i = s_0 + \langle \eta \rangle,$$

if the noise has zero mean. Therefore, averaging over a sufficient number of trials eliminates the noise leaving the constant signal intact.

Differences of processing speed and hence in the time scale of single trials are obviously ignored by this scheme. In case of trial-dependent temporal variations, the assumption of an underlying isochronous signal $s_0(t)$ is questionable and the pointwise average results in a sub-optimal estimate of the signal (i.e. it does not converge against the true signal s_0). Previous research suggests that trial-dependant signal modulation is prominent in ERP data[31]: statistical coherence measures calculated on signal residuals $s_i - \langle s_i \rangle_i$ show significant deviations from the assumption of zero-mean and independant and identically distributed noise that is inherent to model (1).

Several techniques have been proposed to deal with the issue of temporal variation across trials. In Ref. 31 the model $s_i(t) = s_0(t - \tau_i) + \eta$ was investigated. It includes a trial-dependent time shift τ_i that can be compensated for easily (e.g. by a cross-correlation based technique[35]). Another attempt to compensate for temporal distortion prior to averaging used parametric functions[7]: The authors proposed to stretch or compress the single signals in order to match the average reaction time, by moving the sampling points in time according to a linear, quadratic, cubic or quartic function. However, the choice of the adequate warping function remains an open question.

The assumption of a specific class of distortion functions is a strong restriction that has been generalized by considering order-preserving time warping of the data in the generating model

$$s_i(t) = \alpha_i s_0(\phi_i^{-1}(t)) + \eta, \tag{2}$$

where $\phi_i : \mathcal{T} \to \mathcal{T}$ are monotonous time-distortion functions (also called alignment, warping or registration functions) and $\alpha_i > 0$ scaling constants.[10,11,32] Following this model, an adequate average would compensate for the trial-by-trial temporal variation

$$\hat{s}_0(t) = \frac{1}{N} \sum_{i=1}^{N} s_i(\phi_i(t)), \tag{3}$$

where the ϕ_i functions are the solutions to the variational problems

$$\phi_i = \arg\min_{\phi} \int_{\mathcal{T}} d(s_0(t), s_i(\phi(t)))\, dt, \tag{4}$$

d being an adequate distance function (e.g. Euclidean). Obviously, the problem from Eq. (4) cannot be solved directly because the true signal s_0 is unknown and the time-distortion functions used to form the improved average from Eq. (3) must therefore be estimated from mutual relations among the trials. Mathematically, the problem 4 is ill-posed because the condition of monotonicity and possibly further constraints such as on the smoothness of ϕ are not included in its formulation.

Previous approaches to the approximation of the ϕ_i functions in absence of the true signal used e.g. an iterative algorithm to solve the Euler-Lagrange equations corresponding to the variational problem 4. These have to be satisfied for any minimal solution of the integral over a cost-functional

$$F(\phi_1, \ldots, \phi_N)(t) = \sum_{i=1}^{N} d\left(s_i \circ \phi_i, \langle s_j \circ \phi_j \rangle_j\right)$$

that minimizes the deviation of the warped-average from the individually warped curves.[34] Other approaches calculated pairwise registration functions and iteratively combined pairs of trials to form an improved average (e.g. Refs. 26 and 33).

While the latter approach does not yield an explicit estimate of the alignment functions ϕ_i (only the average curve is constructed), it is attractive from a computational perspective because the globally optimal alignment of two curves can be efficiently calculated using a dynamic programming strategy called *dynamic time warping* (DTW).[22] Dynamic

time warping has already been used in an EEG context for classification[16] and averaging.[26] For example, Picton[26,27] proposed an unrestricted dynamic time-warping strategy on raw EEG data to estimate the warping function between single trials. However, this method lacks the robustness required in the application to noisy ERP data. Previously, we have extended this approach by a guided regularization scheme, making it applicable also to noisy data.[10,11]

Here, we will introduce a novel method for the estimation of the structural average in Eq. (3). Our approach is based on (cross-)recurrence plots[4] which emerge in nonlinear dynamical system analysis when comparing time-series from the same dynamical system in terms of trajectories through a reconstruction of their phase-space. The line-of-synchrony extracted from these plots can be used to efficiently extract a warping function for two time-series in a nonparametric fashion. The generality of this approach requires regularization of the solution in order to reduce the effects of overfitting. By introducing context-information into the regularization scheme, we construct a meaningful regularization function that reduces the ambivalence of the distortion function and increases the specificity of the reconstructed signal. For the construction of the structural average, we implement a hierarchical pairwise averaging scheme that establishes a number of candidate solutions within clusters of similar trials which then can be used with increased confidence to identify warping functions also for across-cluster comparisons.

The remainder of the paper is organized as follows: First we will discuss the general approach to the construction of the improved average; Next, we present the theoretical background for the building blocks of the algorithm, namely phase-space analysis and cross-recurrence plots as well as dynamic time-warping which is used for line-of-synchrony extraction. Furthermore, we discuss our regularization scheme that allows to incorporate meaningful context-information into the estimation of the warping functions. Finally, we present results on simulated ERP data and data acquired from a psychophysiological experiment. The source code of the described algorithms is freely available in the software library `libeegtools`.[12]

2. Methods

As discussed above, when estimating the structural average curve from Eq. (3), we need to rely on mutual relations among trials to estimate the optimal distortion function that aligns a single trial with the true signal s_0. We therefore consider the minimum of the functionals

$$\mathcal{J}_{ij}(\phi) = \int_{\mathcal{T}} \varrho d(s_i \circ \phi, s_j) + (1-\varrho) G(\phi)\, dt, \quad (5)$$

where s_i and s_j are two different trials ($i \neq j$), $d : \mathbb{R} \times \mathbb{R} \to [0,1]$ is a normalized distance function and $G : \mathcal{V} \to [0,1]$ is a regularization term that penalizes irregular paths. The parameter $\varrho \in \mathbb{R}$ specifies the mutual impact of the two terms. The global minimum of Eq. (5) in the space of monotonic, differentiable functions $\mathcal{V} \subset \mathcal{C}^1$ can be calculated efficiently by dynamic time-warping.

In the following, we present our choice of the distance d which is based on cross-recurrence plots and a method to include context-information into the regularization term G. Furthermore, we will address the issue of moving from Eq. (5) to the case of aligning N trials in a structural average (see Fig. 1 for an outline of the complete algorithm).

2.1. *Phase space reconstruction*

In empirical studies, the different variables of an observed dynamical system are often not directly measurable. EEG potentials, e.g. are produced by the activity of large numbers of neurons which can neither be measured simultaneously nor are all individual neurons of equal importance in a given task. Nevertheless, it is possible, at least for stationary data, to reconstruct the essential dynamics from the observable signals.[25] This is reflected in a growing body of research in the EEG community that analyzes EEG data in a reconstructed phase space rather than using classical signal processing tools.[23,29]

While there are several options for constructing an appropriate embedding (mathematically this is a diffeomorphism of the original phase space) e.g. with multi-channel recordings,[3,14] using a time-delay embedding[24,30] is an attractive approach. This is due to the fact, that a one-dimensional time series $s(t)$ is sufficient for the reconstruction of the possibly

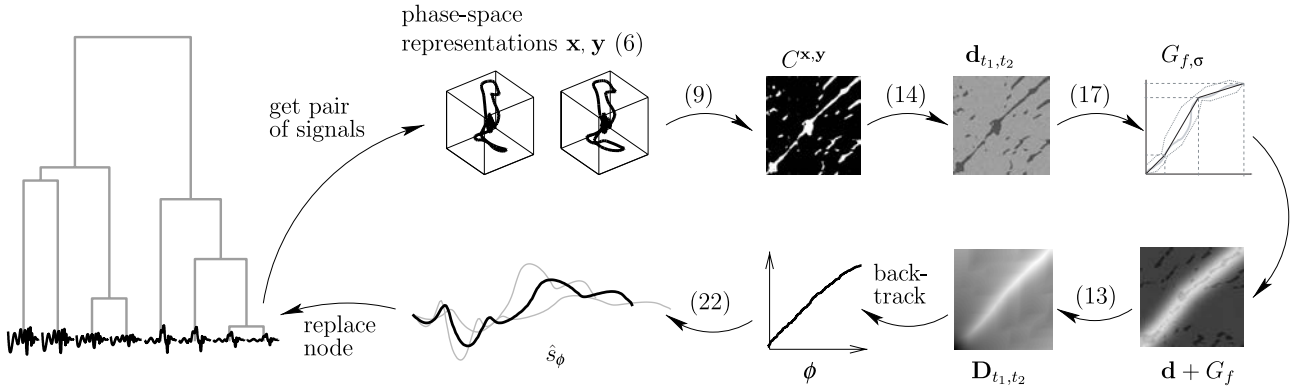

Fig. 1. Summary for the proposed algorithm. A pair of signals is embedded in a reconstructed phase-space (6) and the time-distortion function ϕ is reconstructed from their CRP. The trials are then combined and the parent node of the signals is replaced with their average. Numbers in parentheses refer to equation numbers.

high-dimensional phase-space $\mathbf{x}(t) \in \mathbb{R}^m$

$$\mathbf{x}(t) = \sum_{j=0}^{m} s(t - j\tau)\mathbf{e}_j \tag{6}$$

with $\mathbf{e}_j$ being the unit vector along the jth coordinate of the embedding space and m and τ parameters of the embedding. In the remainder of this paper, we will use time-delay embeddings because often only few electrodes are of interest for a specific research question. The methods easily generalize to multi-channel or a combination of multi-channel and time-delay reconstruction.

The time-delay embedding (6) depends, however, strongly on the choice of the *embedding dimension* m and the *time lag* τ (usually a multiple of the sampling time Δt). Finding optimal values for the two parameters is not trivial and has been subject to intense research.[6,15] Optimality in the present problem is characterized, on the one hand, by the representation of the dynamical properties of the data that might not be captured by low-dimensional embeddings. On the other hand, high embedding dimensions might compromise the locality of information entering the warping procedure and decrease performance of the algorithm.

The minimal embedding dimension m that still covers the pertinent information in the data can be determined by the *false nearest neighbors* method.[15] "False" neighbors (due to a too small embedding dimension) are detected using the fact that they disappear when increasing the embedding dimension by 1. The number m is the smallest dimension for which an increase of the embedding dimension does not lead to a further decrease of the number of nearest neighbors. Furthermore, we use an information-theoretic approach[6] to estimate the time lag τ by the first local minimum of the mutual information. The mutual information is considered to be a better criterion than the autocorrelation function, because it measures the probabilistic dependence rather than only a global linear dependence.[6] For results of these parameter estimation procedures, see Sec. 3.2.

2.2. *Recurrence plots*

Recurrences, i.e. points in state space that are visited more than once, are fundamental characteristics of many dynamical systems, ranging from the financial market to epidemics and brain dynamics.[20] In order to visualize and quantify such recurrences, Eckmann[4] introduced the notion of recurrence plots (RP). Considering a general dynamical system

$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}(t)),$$

where $\mathbf{x}(t) \in \mathbb{R}^m$ and $t \in \mathcal{T}$, we define its recurrence plot as

$$R^{\mathbf{x}}_{\varepsilon}(t_1, t_2) = \Theta(\varepsilon - ||\mathbf{x}(t_1) - \mathbf{x}(t_2)||), \tag{7}$$

where $t_1, t_2 \in \mathcal{T}$ and Θ the Heaviside step function, i.e. $R^x_\varepsilon : \mathcal{T}^2 \to \{0, 1\}$ is a binary function that is equal to one whenever the trajectory in phase space falls within an ε-ball of a previously visited point and zero otherwise. A recurrence plot provides a visualization of the dynamics of the underlying system including information about stationarity, cyclicism, laminar states etc. (for a review see Ref. 20).

The determination of a suitable threshold ε from Eq. (7) for signal-analysis with recurrence plots remains an open research problem.[18] In fact, the "optimality" of any chosen threshold depends on the task to which the RP is going to be applied. For an example of the impact of different threshold choices, see Fig. 2. In order to determine the neighborhood criterion we apply a fixed amount of neighbors (FAN) criterion in order to determine ε on a per-point basis for each point t on the trajectory.[20] That means, that $\varepsilon(t)$ is determined such, that each point has the same amount of neighbors. The advantage of this procedure is the elimination of the impact of any scale difference in the applied curves such that no amplitude normalization is necessary. In our simulations, we use a constant fraction κ of the number of samples n, i.e. $\varepsilon(t)$ is determined, such that

$$\sum_{t'} R^{\mathbf{x}}_{\varepsilon}(t,t') = \lfloor \kappa n \rfloor, \tag{8}$$

where $\lfloor\cdot\rfloor$ rounds its argument to the smaller of the two nearest integers. For the estimation of κ, we apply a criterion that measures the visibility of the structures in the RP that are critical for our method and penalizes an excessive number of recurrences in the plot (see Sec. 3.2).

Recurrence plots have been applied to ERP data[13,19,20] for analyzing properties of the brain as a dynamical system. Here we are proposing a method based on recurrence plots for the synchronization of individual trials prior to averaging their ERPs. In order to reduce distortions of the average due to timing variability of the curve we employ a detection of the line of synchrony in the cross-recurrence plot of the ERP of two trials.

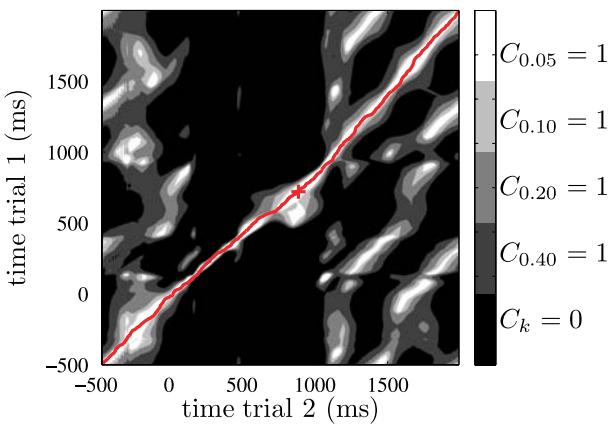

Fig. 2. A CRP calculated for two trials from an artificial data set, see Sec. 3.3 and various values of κ, Eq. (8). Since the data is generated from the same underlying model, the plot shows a line-of-synchrony in red that goes through the simulated reaction times (cross). Different shades of gray convey recurrences for different values of κ while black denotes non-recurrences for any of the plotted κ-values (see legend). The LOS is already well visible for small values of κ (e.g. $\kappa = 0.05$).

2.3. *Line-of-synchrony detection in cross-recurrence plots*

A generalization of recurrence plots are the so-called *cross-recurrence plots* (CRPs) where two signals are simultaneously embedded into the same phase space.[20] For this embedding to be meaningful, both signals must be measurements from the same underlying dynamical system. The cross-recurrence plot is given by

$$C^{\mathbf{x},\mathbf{y}}_{\varepsilon}(t_1,t_2) = \Theta\left(\varepsilon - ||\mathbf{x}(t_1) - \mathbf{y}(t_2)||\right), \tag{9}$$

where $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^m$ are the phase space trajectories of the two signals. The parameter ε is determined using Eq. (8) and will be henceforth omitted.

It has been shown, that CRPs of similar signals with varying time scale, show a "distorted main diagonal", the *line of synchrony* (LOS), that can be extracted and used for resynchronization[21,36] (see Fig. 2). This approach has been applied to geological data[21] and was found to yield results comparable to manual tuning. The algorithm from Ref. 21 applies a recursive growing-window strategy where a window is placed around the current LOS end point. The window grows into x or y direction until no additional recurrence sites are covered by enlarging its size. The geometric mean of the current LOS end point and the last detected recurrence site is the next point of the LOS. The method is parametrized by d_x and d_y, the maximum size of the window in x and y direction, respectively. It has been shown, that this algorithm works quite well for small deviations from the main diagonal. However, applying the method to EEG data is a somewhat more delicate undertaking, because (i) the initial signal-to-noise ratio is extremely low, (ii) the temporal variability can be quite large[10] and (iii) the underlying signals can be different. In behavioral experiments, large variations in reaction times are observed[28] that are not even one order of magnitude smaller than the mean reaction time. Assuming the reaction times to indicate equivalent states of the brain dynamics in two

trials, we have to accept that a significant temporal distortion of the signal must have occurred. We were therefore led to propose a nonparametric and more flexible algorithm for determining the line of synchrony that is based on a dynamic time-warping strategy.

2.4. *An algorithm for line-of-synchrony extraction*

We aim to optimize the number of recurrences on the LOS while simultaneously minimizing its length. Let ϕ' be the first derivative of ϕ, then our goal is to maximize the line integral over the cross-recurrence plot (i.e. the number of recurrences along the path)

$$\int_{\mathcal{T}} C^{\mathbf{x},\mathbf{y}}(t,\phi(t))\sqrt{1+\phi'(t)^2}\,dt \tag{10}$$

and to minimize the length of the path

$$\int_{\mathcal{T}} \sqrt{1+\phi'(t)^2}\,dt \tag{11}$$

for $\phi \in \mathcal{V}$ as before and boundary conditions $\phi(0)=0$ and $\phi(1)=1$.

When dealing with binary functions $C^{\mathbf{x},\mathbf{y}}$, there are many possible solutions for minimizing Eq. (10) because in a neighborhood containing zeros, it is arbitrary which path is taken. We therefore restrict our search space to all paths satisfying Eq. (10) and minimize Eq. (11) in this restricted space.

In practice, we perform the simultaneous optimization of Eqs. (10) and (11) by a single application of the *dynamic time warping* algorithm (DTW)[22] to a modified recurrence matrix. DTW finds the monotonous path through a discrete distance matrix $\mathbf{d}_{i,j}$ that minimizes the sum of the matrix entries along that path. In other words, DTW finds the optimal alignment $\hat{\mathbf{p}} = ((l_1,k_1),\ldots,(l_K,k_K))$ between two discrete signals $u_i = s_1(i\Delta t)$ and $v_j = s_2(j\Delta t)$ for $i,j=1,\ldots,n$ and $i\Delta t \in \mathcal{T}$ of the form

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \sum_{(l,k)\in\mathbf{p}} d(u_l, v_k) \tag{12}$$

for a cost function d. Equation (12) is a discretized version of the variational problem from Eq. (5). DTW applies a dynamic programming strategy that can be efficiently calculated in $\mathcal{O}(n^2)$ (but there are optimized algorithms that run in $\mathcal{O}(n\log n)$[2]). In summary, the algorithm operates by cumulating $\mathbf{d}_{i,j}$

$$\mathbf{D}_{i,j} = \mathbf{d}_{i,j} + \min\{\mathbf{D}_{i,j-1}, \mathbf{D}_{i-1,j}, \mathbf{D}_{i-1,j-1}\} \tag{13}$$

and subsequently backtracking the minimal adjacent element through $\mathbf{D}$. The details of the DTW algorithm are discussed elsewhere[22] as well as how it can be applied to EEG data analysis.[10,11,26]

In our case, application of DTW to the binary matrix $1 - C^{\mathbf{x},\mathbf{y}}(t_1,t_2)$, where $t_1,t_2 \in \{i\Delta t \,|\, i\Delta t \in \mathcal{T};\ i=1,\ldots,n\}$ results in the warping function minimizing the number of non-recurrences along the way, i.e. maximizing Eq. (10). As stated above, there is more than one solution to this minimization since the cost is not influenced by additional zeros along the path. To impose the path length criterion from Eq. (11), we modify the recurrence matrix that enters the DTW algorithm such that

$$\mathbf{d}_{t_1,t_2} = (1 - C^{\mathbf{x},\mathbf{y}}(t_1,t_2)) + \eta, \tag{14}$$

where η is uniformly distributed uncorrelated positive noise of an amplitude much smaller than 1. The amplitude of the noise must be chosen such that the expectation of the sum of all noise components is less than 1

$$\mathrm{E}\left(\sum_{t_1,t_2}\eta\right) = \mathrm{E}(n^2\eta) = n^2\mathrm{E}(\eta) < 1. \tag{15}$$

This choice ensures that no recurrence is neglected due to the impact of the cumulating noise component. There are two effects of the added noise component: First, since $\eta > 0$, it serves as a soft penalty of the path length. Second, the noise breaks the tie if the backtracking stage of the algorithm faces identical values.

We combine the two signals using the solution to Eq. (5) thus acquired by simultaneously averaging in time and amplitude

$$\begin{pmatrix}\hat{t}\\ \hat{s}\end{pmatrix} = \frac{1}{2}\begin{pmatrix} t_1 + \phi(t_1)\\ s_1(t_1) + s_2(\phi(t_1))\end{pmatrix} \tag{16}$$

and find the sampling points of the original time scale by interpolation.

2.5. *Regularization*

For real EEG data, it is not always clear that the data generated during identical trials are realizations of the same cognitive processes. It is therefore

well possible that the embedding in the same phase-space will in practice not yield satisfactory results (i.e. the CRP does not show a distorted main diagonal). Because the DTW strategy from Sec. 2.4 finds the globally optimal minimum of the warping path, it can yield pathological warping functions for such a case. It is therefore crucial to apply a regularization rule that avoids such cases.

Obviously, the quality of the result can be improved by using prior knowledge. Relevant criteria can be smoothness or the deviation from the non-warped case but also context information from the experimental protocol. In the following, we will propose a strategy that makes use of timing information available from the experimental context to avoid too irregular paths.

2.5.1. *Cross-recurrences with external events*

In behavioral experiments, subjects usually perform some task while the EEG data is recorded. From the subject's behavior, context information can be extracted that bears the potential to improve the results in the analysis of EEG data. In simple tasks requiring a single action of the subject, the reaction time is a valuable event and has already been used to improve averaging.[7] However, in many cases not only the reaction time is available, but the experimental protocol allows also for the recording of other distinct temporal events that can be matched to the time course of the recorded EEG data. We refer to these points in time as *time markers* and label the ith marker ($i = 1, \ldots, M$) in trial j by $\tau_{i,j}$.

In fact, the stimulus onset is the most obvious time marker. It provides the basis of ERP because it is used for segmenting the continuous stream of EEG data into ERP segments. If the experimental task includes e.g. button-presses, utterances or specific movements it is possible to record more than just this fundamental time marker. Given that these markers occur similarly in all trials but with a temporal variation, one could argue for using this behavioral information to guide the distortion function through corresponding points in time as measured by these markers. Reference 7 already used such an approach using a specific time-marker, the reaction time, as the defining property of the warping function. The method discussed here is a generalization of this approach and includes it as a special case.

In order to use the time markers to bias the search for the distortion function in favor of those functions that map $\tau_{i,j}$ to $\tau_{i,k}$ (where $j \neq k$ are two trials), we propose the following regularization scheme: Considering the series of time markers $\tau_{1,j}, \ldots, \tau_{M,j}$, we construct a function $f : \mathcal{T} \to \mathcal{T}$ as a piecewise interpolation between $(\tau_{i,j}, \tau_{i,k})$ and $(\tau_{i+1,j}, \tau_{i+1,k})$ for all i. We realize a regularization function as a two-dimensional Gaussian

$$G_f(x, y; \sigma) = \frac{1}{2\pi\sigma} \exp\left(-\frac{\Delta\{f\}(x, y)}{2\sigma^2}\right) \qquad (17)$$

calculated on the distance transform $\Delta\{f\}$ of f which computes the minimum Euclidean distance from the embedding of f in $\mathbb{R}^2$ and thus forms a "Gaussian corridor". The distance transform $\Delta\{f\}$ of an $(n-1)$-dimensional function f is a scalar function of dimension n and defined as the minimal Euclidean distance from each $\mathbf{x} \in \mathbb{R}^n$ to the manifold $F = \{x_n = f((x_1, \ldots, x_{n-1})^T)\} \subset \mathbb{R}^n$ that is

$$\Delta\{f\}(x) = \min_{\xi \in F} \|\xi - \mathbf{x}\|^2.$$

We implement an efficient algorithm for calculating the distance transform of a two-dimensional function, the dead-reckoning algorithm from Ref. 8.

Using the distance transform realizes a flexible strategy where a variety of different prior information can easily be integrated to find more suitable solutions. Reference 7, e.g. argued that a close-to-quadratic warping function between time-markers is realistic because brain activity from cognitive processes should be similar close to a stimulus and begin to diverge later. Including such prior information is easily done by simply connecting the time-markers with functions that feature parabolic deviations from the connecting line instead of straight lines when constructing f from Eq. (17).

If there is high confidence in the recorded time-markers, a stricter regularization that more explicitly guides the distortion function through corresponding time markers between trials can be used. Define σ from Eq. (17) to be dependent on the distance of the orthogonal projection of any point (x, y) on f to the closest time marker along f

$$d_\perp(x, y) = \min_i \left\{ \sqrt{(x - \tau_{i,j})^2 + (y - \tau_{i,k})^2 - \Delta\{f\}(x, y)} \right\} \qquad (18)$$

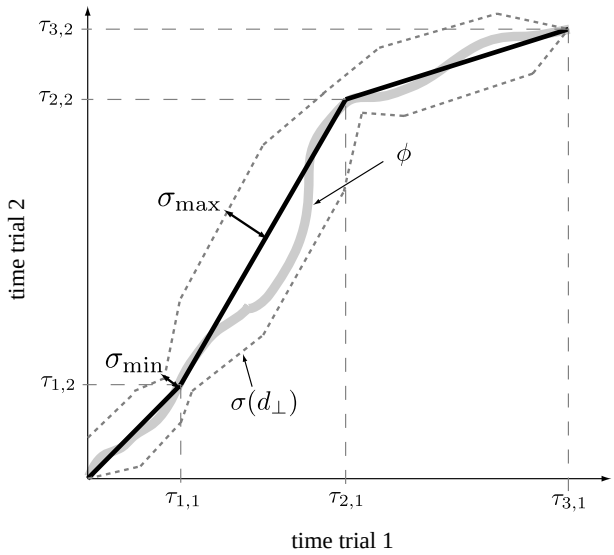

Fig. 3. Sketch of the regularization matrix. The standard deviation of the Gaussian σ varies with distance to the closest time marker $d_\perp$ and narrows down to guide the path through corresponding markers.

by choosing

$$\sigma(d_\perp) = \begin{cases} \sigma_{\max} & \text{if } d_\perp > \hat{d}_\perp \\ (\sigma_{\max} - \sigma_{\min}) d_\perp \hat{d}_\perp^{-1} & \\ \quad + \sigma_{\min} & \text{otherwise.} \end{cases} \tag{19}$$

Here, $\sigma_{\min}$ and $\sigma_{\max}$ are parameters that specify the tolerated deviation from a linear displacement as in Ref. 7 and $\hat{d}_\perp$ is the distance at which the Gaussian corridor begins to narrow down to the closest time marker (see Fig. 3). In the discrete case, $\sigma_{\max}$ should be normalized by $\sqrt{2}N - 1$ before evaluating Eq. (19). The parameter $\sigma_{\min}$ reflects the confidence in the externally recorded time-markers and is necessary for numerical reasons (we chose $\sigma_{\min} = 0.05$).

Finally, the regularization term is added to the distance function 14 and the variational problem from Eq. (5) is solved by dynamic time-warping. The choice of the two main regularization parameters ϱ and $\sigma_{\max}$ is discussed in Sec. 3.2.

2.6. *Hierarchical averaging*

The regularization helps to avoid pathological paths that can e.g. occur when the two signals are emitted by different dynamical systems due to diverging cognitive processes involved. An even better approach would filter such cases and only combine trials that can be assumed to have been generated from a similar system. Technically, one of the many classification techniques such as support-vector machines (SVM) or neural networks can be used. These methods have successfully been applied to EEG-data to classify mental tasks[17] or to identify epileptic seizures.[5] However these methods group the data into a given number of discrete classes. For our averaging procedure, a hierarchical method seems to be more suitable since it allows to iteratively combine trials of growing diversity.

Previously, averaging of N signals was achieved by subsequent application of Eq. (16) to trials drawn randomly without repetitions from $\{1, \ldots, N\}$.[26] However, a more suitable method seems to be the application of a hierarchical approach, i.e. similar trials are averaged prior to dissimilar ones. This is due to the fact that warping is more easily accomplished and yields numerically more stable results, the better the match of the two curves. We therefore argue for a hierarchical averaging procedure that combines the most similar trials before integrating the resulting average with the other trials in a tree-like fashion. In fact, this corresponds to a hierarchical cluster analysis and can be done by using the complete-linkage distance between clusters I and J

$$\Delta(\{s_i \mid i \in I\}, \{s_j | j \in J\}) = \max_{i \in I, j \in J} D(s_i, s_j) \tag{20}$$

and building a dendrogram according to Δ. Here, D is a distance metric between two trials given e.g. by the warp distance

$$D(s_i, s_j) = \min_\phi \int_{\mathcal{T}} d(s_i(t), s_j(\phi(t))) \sqrt{1 + \phi'(t)^2}\, dt \tag{21}$$

or its regularized variant. Using the warp-distance has the advantage of comparing the trials on the basis of their general shape rather than pointwise discrepancies. The averaging is done by adding corresponding points in both signals by modifying Eq. (16) as

$$\begin{pmatrix} \hat{t} \\ \hat{s} \end{pmatrix} = \frac{1}{\omega_1 + \omega_2} \begin{pmatrix} \omega_1 t_1 + \omega_2 \phi(t_1) \\ \omega_1 s_1(t_1) + \omega_2 s_2(\phi(t_1)) \end{pmatrix}, \tag{22}$$

where ω_1 and ω_2 are the numbers of trials already averaged in each of the current nodes. This procedure achieves averaging both in time and amplitude and results in a proper weighting of the impact of the single-trial curves in the final average.

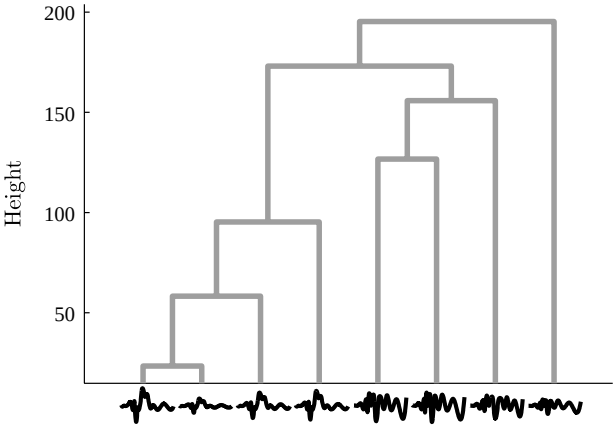

Fig. 4. Using a hierarchical cluster-analysis based on signal-shape similarity (Eq. (20)), trials are progressively combined (Eq. (22)) to form a final average: each node that is the parent of exactly two leaves is replaced by the warped average of the two children. This procedure is repeated until a single average is obtained at the root node.

Figure 4 illustrates the process of hierarchical averaging. Notewhorthy is that at this stage, possible outliers in the data can be be identified from the tree structure. Branches that contain single trials but join the tree only near the root may be pruned (such as the right-most branch in Fig. 4) if this can be justified within the context of the experiment.

3. Results

3.1. *Construction of Artificial Data*

In order to test the consistency of the algorithm, we construct artificial data to closely resemble our experimental data (see Sec. 3.4). Correspondingly, we generate single trial traces on the interval $\mathcal{T} = [-500\,\text{ms}, 2000\,\text{ms}]$ from a template s_0 that is of a typical ERP shape (see Fig. 5). Using the model from Eq. (2) we construct distortion functions ϕ_i based on a biased random walk where the probability to move away from the linear interpolation through corresponding time-makers decays with the distance to this function. In addition to the constant stimulus-onset marker $\tau_{1,0} = \tau_{1,i} = 0\,\text{ms}$, we generate realistic reaction-times $\tau_{2,i} \sim \mathcal{N}(\tau_{2,0}, \sigma_{\text{rt}})$ where the template's reaction time was chosen to be $\tau_{2,0} = 800\,\text{ms}$ and $\sigma_{\text{rt}} = 100\,\text{ms}$ which is of a similar magnitude to our real EEG data (see Sec. 3.4). Here $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ. The signal is multiplied by a constant

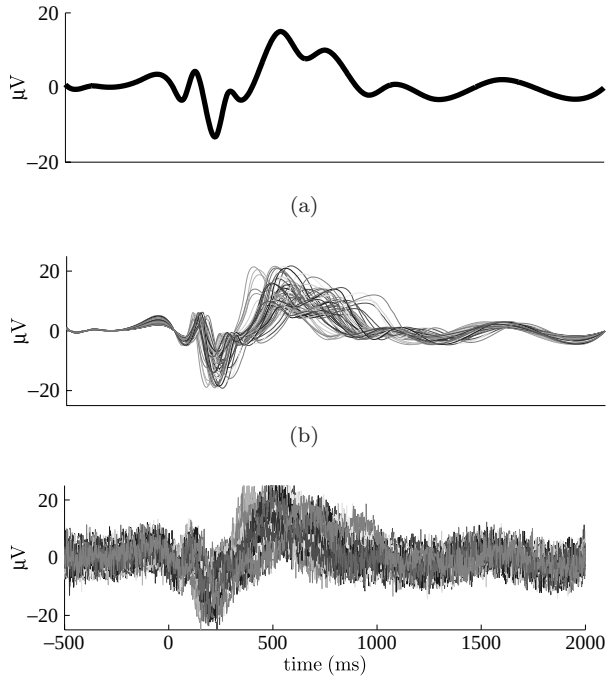

Fig. 5. Construction of the artificial data. A template is constructed manually (a) from which temporally distorted and rescaled versions are derived (b). Finally, noise is added to the single-trial data (c).

term α and subject to a significant amount of white gaussian noise

$$\eta \sim \mathcal{N}(0, \beta^2),$$

where β controls the amplitude of the added noise.

3.2. *Parameter estimation*

There are five parameters to be chosen when applying the discussed algorithm: besides choosing the embedding parameters m and τ, there is the recurrence-plot paramter κ controlling the number of recurrences as well as the regularization parameters ϱ and $\sigma_{\max}$. We discuss our choice of the embedding parameters that is based on state-of-the-art methods from nonlinear time-series analysis. Our determination of the regularization parameters is based on cross-validation.

3.2.1. *Embedding arameters*

The embedding parameters m_{opt} and τ_{opt} must be estimated from the data. Applying the false-nearest-neighbors method[15] to our ERP-data (see Sec. 3.4)

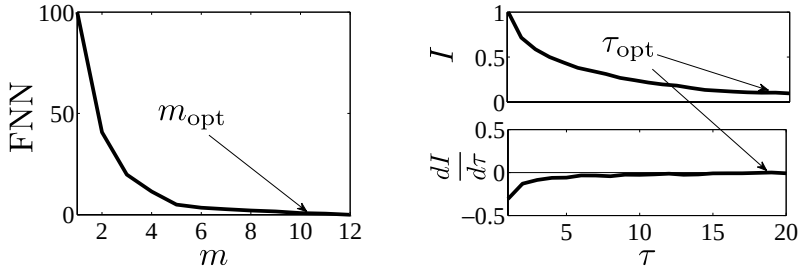

Fig. 6. Percentage of false nearest neighbors (FNN) as a function of embedding dimension m (left plot). After $m_{\text{opt}} = 10$, no false nearest neighbors are detected. Right plot: the first local minimum of the mutual information I estimates the optimal time lag τ_{opt} (first zero-crossing of the first derivative of I).

we obtain an estimate of $m_{\text{opt}} = 10$, see Fig. 6. This is a similar value as found in earlier studies that estimated time-delay embedding parameters for ERP data.[13] The first local minimum of the mutual information[6] provides an estimate for the time-lag of $\tau_{\text{opt}} = 18$ sampling units which seems to be feasible for data sampled with 1000 Hz (Fig. 6).

An appropriate value for the parameter κ for the construction of the cross-recurrence plots ensures that the line-of-synchrony is well visible but not too thick. To determine such a value, we calculated the criterion

$$\text{crit}(\kappa) = \frac{2}{N(N-1)} \sum_{i \neq j} \frac{\Phi(\kappa, C_\kappa^{\mathbf{x}_i,\mathbf{x}_j}, \phi)}{\mathcal{L}(\phi)\kappa} \qquad (23)$$

with $\mathcal{L}$ being the arc-length of ϕ (Eq. (11)), Φ the line-integral of the LOS ϕ (calculated for an initial guess for κ) over the CRP as in Eq. (10). Because $\kappa \in [0, 1]$ is proportional to the number of recurrences in the CRP

$$\kappa \sim \iint_{t_1,t_2} C_\kappa^{\mathbf{x}_i,\mathbf{x}_j}(t_1, t_2)\, dt_1 dt_2$$

(see Eq. (8)), measure 23 penalizes an increasing number of recurrences while requiring a large number of recurrence points on the LOS. From Fig. 7, we can deduce that values in $[0.05, 0.13]$ are suitable. We choose $\kappa = 0.1$ for all following simulations.

3.2.2. *Regularization parameters*

For the artificial data, we can directly calculate how closely the final average resembles the original data curve from which it was generated. Figure 8 shows the warp-distance between original and reconstruction depending on the regularization parameters.

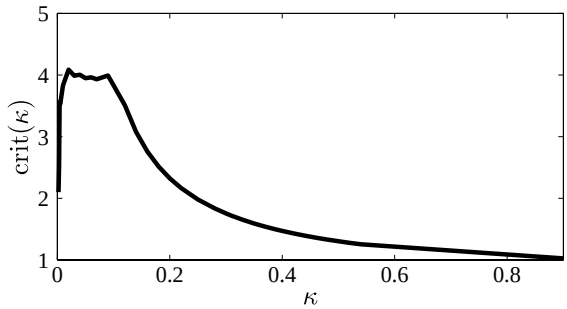

Fig. 7. Parameter scan for the CRP-parameter κ, see Eqs. (8) and (23). The criterion is maximized in the range $[0.05, 0.13]$ and we choose the value $\kappa = 0.10$ for our simulations.

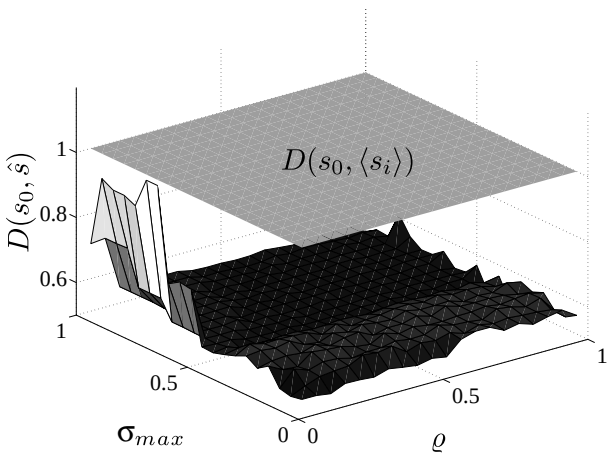

Fig. 8. Reconstruction results (normalized warp-distance to the underlying signal) of our method for various values of ϱ and $\sigma_{\max}$ for $N = 50$ trials of artificial ERP-data using the stimulus-onset and reaction-time markers $\tau_{1,i}, \tau_{2,i}$. The algorithm is very robust with respect to the choice of the parameters and produces better results than the conventional average (gray plane) for all settings of the regularization parameters.

We chose the warping distance D from Eq. (21) because we are interested in recovering the shape of the signal, rather than minimize the pointwise deviations. Obviously, the method is very robust concerning the choice of these parameters and it performs significantly better than the conventional, pointwise average. For an example of reconstructing the generating signal for our artificial data, see Fig. 11. These simulations used the stimulus-onset and reaction-time markers $\tau_{1,i}, \tau_{2,i}$ that were used to generate the single-trial artificial data (Sec. 3.1).

In the case of real EEG-data, the regularization parameters are conveniently determined by Leave-One-Out crossvalidation (cf. Ref. 9). The following expression is minimized

$$CV(\varrho, \sigma_{\max}) = \frac{1}{N} \sum_{i=1}^{N} D(s_i(t), \langle s_j(t) \rangle^{-i}_{\varrho,\sigma_{\max}}), \quad (24)$$

where $\langle s_j(t) \rangle^{-i}_{\varrho,\sigma_{\max}}$ is the regularized average leaving out trial i, i.e. it is calculated for $j = 1, \ldots, i-1, i+1, \ldots, N$. Figure 9 shows the result for our EEG-data set. Apparently, the distance and the regularization terms in Eq. (5) should receive about equal weight ($\varrho \approx 0.5$) while the $\sigma_{\max}$ parameter should be set quite narrow $\sigma_{\max} \approx 0.1$. This means that for our data, the time-markers do indeed add significant information to the formation of the average which should be closely followed by the registration functions. We used the stimulus-onset and

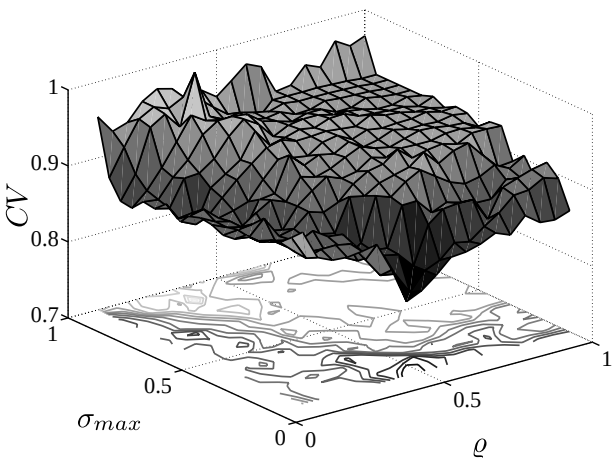

Fig. 9. Normalized crossvalidation score for reconstruction based on the DTW metric as a function of the parameters ϱ and $\sigma_{\max}$ for real ERP data, $N = 52$ (see Sec. 3.4). There is a unique minimum at $\varrho \approx 0.5$ and $\sigma_{\max} \approx 0.1$.

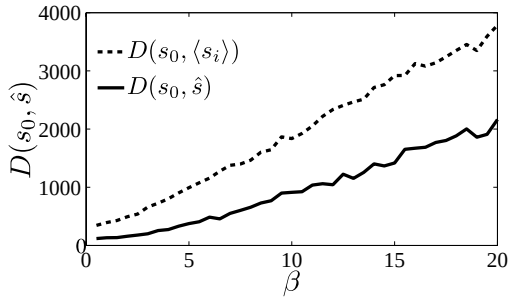

Fig. 10. Reconstruction error as a function of noise amplitude for artificial data. While the error for our reconstruction method (black) increases with growing noise amplitude β, its slope is less steep than that of the conventional average (dashed).

the response-onsets of the individual trials as time-markers for our approximation of the average curve.

3.3. *Reconstruction results for artificial data*

For artificial data (see Sec. 3.1), we can directly calculate the reconstruction error because the "true" signal s_0 is known. Again using the warping distance 21, we can see that the reconstructed signal using our method more closely resembles the generating signal than the pointwise average, see Figs. 10 and 11. The slope of the reconstruction error with growing noise intensity is less steep than for the pointwise average. Thus, the method can be confidently applied also to data that is contaminated with strong noise. Because we encounter high noise levels in realistic settings, stability against the noise is an important feature of any algorithm for the analysis of ERP data.

Finally, we show examples for the temporal averaging based on the extracted LOS. In Fig. 11, the shape of the original signal (white) can be recovered even from very noisy data (gray curves are single trials). In contrast, the conventional pointwise average (dashed) washes out some of the peaks.

3.4. *EEG data from a psychophysiological study*

The data presented here was recorded in a psychophysical priming experiment involving the presentation of overlapping pictograms. The experimental protocol demanded a vocalized response by naming one of the pictograms with a matching label (for details of the experimental setup and the

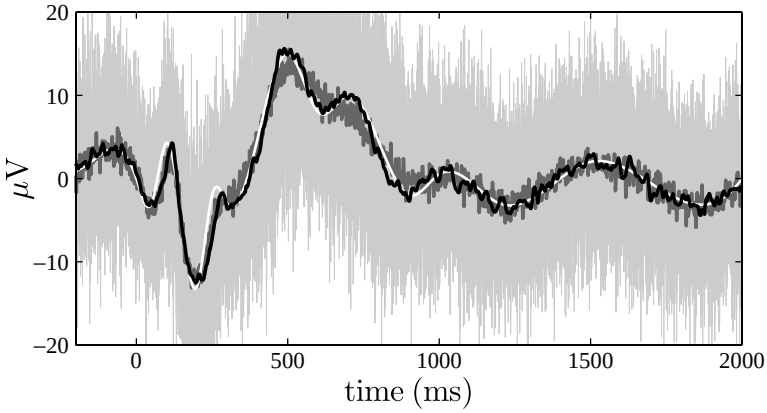

Fig. 11. Reconstruction of the signal s_0 (white) from very noisy single-trial data (gray, $\beta = 6, N = 30$). Our algorithm (black) is able to reconstruct all major features of the waveform, while the conventional average (dark gray) washes out some of the peaks/throughs.

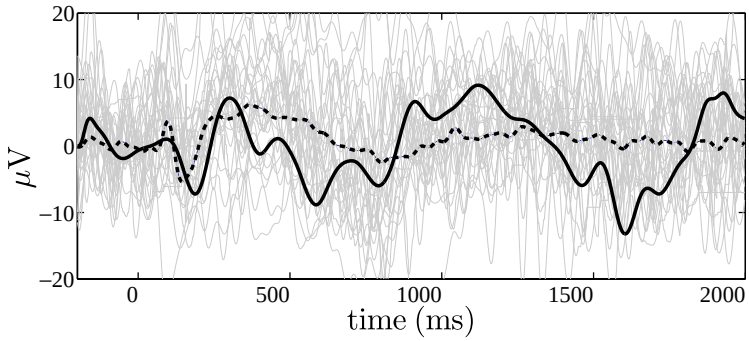

Fig. 12. Real ERP data from 134 trials at electrode location CPz. Our resynchronization method (solid) shows more pronounced and better localized peaks (e.g. P300) than the conventional average (dashed). The single trials (grey) were transformed for plotting according to the described algorithm (i.e. the plot shows $s_i(\phi(t))$).

recording equipment, refer to Ref. 1). We could therefore use two time markers for our analysis, stimulus and response onset.

The EEG was recorded in 50 trials from 63 head electrodes arranged in an extended 10–20 system, using a 64-channel BrainAmp MR amplifier and an electrode cap (Brain Products Inc., Germany) with sintered Ag/AgCl electrodes. The sampling rate was 1,000 Hz and band-pass was set to 0.1–70 Hz. Vertical electro-oculogram (EOG) was monitored from an electrode positioned 1 cm below the right eye. The FCz electrode served as active reference. All impedances were kept below 5 kΩ. Before data analysis with our method, EEG segmentation was performed, resulting in $[-200, 2000]$ ms epochs, with zero indicating probe display onset. Epochs were then baseline-corrected with respect to the $[-200, 0]$ ms interval, and an initial artifact rejection was performed to identify epochs with technical artifacts (i.e. amplitudes exceeding 1 mV). Finally, we applied a bandpass filter with cutoff frequencies 0.5 and 20 Hz.

The results of the application of our method to the data are exemplified in Fig. 12. Our average produces pronounced peaks in the expected time window (e.g. P300) that are better visible and more tightly localized in time than when using conventional averaging. In addition, the shape of the average shows more variations and several peaks and throughs become visible that are not detected by the pointwise average. The hypothesis underlying our averaging procedure was that important signal modulations were washed out by cross-sectional averaging and it is therefore not surprising that additional peaks and throughs are visible in our approach. Obviously, however, it remains to be tested whether

these features are actually reproducible in real ERP-studies, a question that should be the objective of future investigations.

4. Conclusion

We presented an algorithm for the estimation and compensation of temporal distortion in ERP-segments. Our method is based on line-of-synchrony detection in cross-recurrence plots and a pairwise combination based on the estimated warping function. The robustness of the algorithm is improved by introducing a regularization scheme into the estimation of the distortion function and appropriate settings for the regularization parameters can efficiently be found by cross-validation. Pairwise averaging based on signal-shape similarity was used in order to avoid numerical problems with too irregular warping paths and local minima of the regularization procedure. Furthermore, an additional improvement is possible by the use of context information that is often available in complex experimental paradigms. Recent experiments have introduced additional time markers by design in order to increase the explanatory power of the data. In this way more specific information on the underlying cognitive processes can be obtained provided the data analysis allows for this information to be extracted. We have demonstrated that our method can be successfully applied to artificial and real ERP data and that it can provide the user with improved averages. Further details, source code and updates of the presented algorithms are publicly available.[12]

Acknowledgment

This work was supported by BMBF grant number 01GQ0432. Fruitful discussions with Marcus Hasselhorn, Ulrich Parlitz, Jörg Behrendt and Theo Geisel are gratefully acknowledged.

References

1. J. Behrendt, H. Gibbons, H. Schrobsdorff, M. Ihrke, J. M. Herrmann and M. Hasselhorn, Event-related brain potential correlates of idendity negative priming from overlapping pictures, *Psychophysiology* **47** (2010) 921–930.
2. S. S. P. Chan, Fast dtw: Toward accurate dynamic time warping in linear time and space, *Intelligent Data Analysis* **11**(5) (2007) 561–580.
3. I. Dvorák, Takens versus multichannel reconstruction in EEG correlation exponent estimates, *Physics Letters A* **151**(5) (1990) 225–233.
4. J. P. Eckmann, S. O. Kamphorst and D. Ruelle, Recurrence plots of dynamical systems, *Europhysics Letters* **5** (1987) 973–977.
5. O. Faust, U. R. Acharya, L. C. Min and B. H. Sputh, Automatic identification of epileptic and background EEG signals using frequency domain parameters, *International Journal of Neural Systems* **20**(2) (2010) 159.
6. A. Fraser and H. Swinney, Independent coordinates for strange attractors from mutual information, *Physical Review A* **33**(2) (1986) 1134–1140.
7. H. Gibbons and J. Stahl, Response-time corrected averaging of event-related potentials, *Clinical Neurophysiology* **118**(1) (2007) 197–208.
8. G. J. Grevera, The "dead reckoning" signed distance transform, *Computer Vision and Image Understanding* **95**(3) (2004) 317–333.
9. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer (2001).
10. M. Ihrke, H. Schrobsdorff and J. M. Herrmann, Compensation for speed-of-processing effects in EEG-data analysis, in *Lecture Notes on Computer Science*, Springer **5326** (2008) 354–361.
11. M. Ihrke, H Schrobsdorff and J. M. Herrmann, Denoising and averaging techniques for electrophysiological data, in R. Wennberg and J. L. Perez-Velazquez (eds.), *Coordinated Activity in the Brain: Measurements and Relevance to Brain Function and Behaviour*, Springer (2009).
12. M. Ihrke, Libeegtools, http://libeegtools.sf.net.
13. J. Jeong, S. Kim and S. Han, Non-linear dynamical analysis of the EEG in Alzheimer's disease with optimal embedding dimension, *Electroencephalography and Clinicial Neurophysiology* **106** (1998) 220–228.
14. H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press (2004).
15. M. Kennel, R. Brown and H. Abarbanel, Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A* **45**(6) (1992) 3403–3411.
16. H. Lee, Y. D. Kim, A. Cichocki and S. Choi, Nonnegative tensor factorization for continuous EEG classification, *International Journal of Neural Systems* **17**(4) (2007) 305–318.
17. N. Y. Liang, P. Saratchandran, G. B. Huang and N. Sundararajan, Classification of mental tasks from EEG signals using extreme learning machine, *International Journal of Neural Systems* **16**(1) (2006) 29–38.
18. N. Marwan, How to avoid potential pitfalls in recurrence plot based data analysis, *International Journal of Bifurcation and Chaos* (2010), in press.

19. N. Marwan and A. Meinke, Extended recurrence plot analysis and its application to ERP data, *International Journal of Bifurcation and Chaos* **14**(2) (2004) 761–771.
20. N. Marwan, M. C. Romano, M. Thiel and J. Kurths, Recurrence plots for the analysis of complex systems, *Physics Reports* **438** (2007) 237–329.
21. N. Marwan, M. Thiel and N. R. Nowaczyk, Cross recurrence plot based synchronization of time series, *Nonlinear Processes in Geophysics* **9**(3–4) (2002) 325–331.
22. C. Myers and L. Rabiner, A level building dynamic time warping algorithm for connected word recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **29**(2) (1981) 284–297.
23. H. Osterhage, F. Mormann, T. Wagner and K. Lehnertz, Measuring the directionality of coupling: Phase versus state space dynamics and application to EEG time series, *International Journal of Neural Systems* **17**(3) (2007) 139–148.
24. N. H. Packard, J. P. Crutchfield, J. D. Farmer and R. S. Shaw, Geometry from a time-series, *Physical Review Letters* **45**(9) (1980) 712–716.
25. U. Parlitz, Nonlinear time-series analysis, in J. A. K. Suykens and J. Vandewalle (eds.), *Nonlinear Modeling: Advanced Black-Box Techniques*, Kluwer Academic Publishers, Boston (1998), pp. 209–239.
26. T. W. Picton, M. Hunt, R. Mowrey, R. Rodriguez and J. Maru, Evaluation of brain-stem auditory evoked potentials using dynamic time warping, *Electroencephalography and Clinical Neurophysiology* **71**(3) (1988) 212–225.
27. T. W. Picton, O. G. Lins and M. Scherg, The recording and analysis of event-related potentials, in F. Boller and J. Grafman (eds.), *Handbook of Neuropsychology*, Elsevier Science B.V. (1995), pp. 3–73.
28. H. Schrobsdorff, M. Ihrke, B. Kabisch, J. Behrendt, M. Hasselhorn and J. M. Herrmann, A computational approach to negative priming, *Connection Science* **19**(3) (2007) 203–221.
29. C. Stam, Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field, *Clinical Neurophysiology* **116**(10) (2005) 2266–2301.
30. F. Takens, Detecting strange attractors in turbulence, *Lecture Notes in Mathematics* **898** (1981) 366.
31. W. A. Truccolo, M. Ding, K. H. Knuth, R. Nakamura and S. L. Bressler, Trial-to-trial variability of cortical evoked responses: Implications for the analysis of functional connectivity, *Clinical Neurophysiology* **113**(2) (2002) 206–226.
32. K. Wang, H. Begleiter and B. Porjesz, Warp-averaging event-related potentials, *Clinical Neurophysiology* **112**(10) (2001) 1917–1924.
33. K. M. Wang and T. Gasser, Alignment of curves by dynamic time warping, *The Annals of Statistics* **25**(3) (1997) 1251–1276.
34. K. M. Wang and T. Gasser, Synchronizing sample curves nonparametrically, *The Annals of Statistics* **27**(2) (1999) 439–460.
35. C. D. Woody, Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals, *Medical and Biological Engineering and Computing* **5** (1967) 539–553.
36. N. V. Zolotova, D. I. Ponyavin, N. Marwan and J. Kurths, Long-term asymmetry in the wings of the butterfly diagram, *Astronomy and Astrophysics* **503**(1) (2009) 197–201.