Journal of Neuroscience Methods 316 (2019) 12–21

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Delay differential analysis for dynamical sleep spindle detection

Aaron L. Sampson[a,b,1,*], Claudia Lainscsek[a,c,1], Christopher E. Gonzalez[a,b], István Ulbert[d,e], Orrin Devinsky[f], Dániel Fabó[g], Joseph R. Madsen[k], Eric Halgren[h], Sydney S. Cash[i], Terrence J. Sejnowski[a,c,j]

[a] Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[b] Neurosciences Graduate Program, University of California San Diego, La Jolla, CA 92093, USA
[c] Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093, USA
[d] Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2, H-1117 Budapest, Hungary
[e] Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, H-1083 Budapest, Hungary
[f] New York University Comprehensive Epilepsy Center, New York, NY 10016, USA
[g] Epilepsy Centrum, National Institute of Clinical Neurosciences, Budapest, Hungary
[h] Departments of Radiology and Neurosciences, University of California San Diego, La Jolla, CA 92093, USA
[i] Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Harvard University, Boston, MA 02114, USA
[j] Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA
[k] Departments of Neurosurgery, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA

A B S T R A C T

Background: Sleep spindles are involved in memory consolidation and other cognitive functions. Numerous automated methods for detection of spindles have been proposed; most of these rely on spectral analysis in some form. However, none of these approaches are ideal, and novel approaches to the problem could provide additional insights.

New method: Here, we apply delay differential analysis (DDA), a time-domain technique based on nonlinear dynamics to detect sleep spindles in human intracranial sleep data, including laminar electrode, stereoelectroencephalogram (sEEG), and electrocorticogram (ECoG) recordings.

Results: We show that this approach is computationally fast, generalizable, requires minimal preprocessing, and provides excellent agreement with human scoring.

Comparison with existing methods: We compared the method with established methods on a set of intracranial recordings and this method provided the highest agreement with human expert scoring when evaluated with F_1 score while being the second-fastest to run. We also compared the results on the DREAMS surface EEG data, where the method produced a higher average F_1 score than all other tested methods except the automated detections published with the DREAMS data. Further, in addition to being a fast and reliable method for spindle detection, DDA also provides a novel characterization of spindle activity based on nonlinear dynamical content of the data.

Conclusions: This additional, non-frequency-based perspective could prove particularly useful for certain atypical spindles, or identifying spindles of different types.

1. Introduction

1.1. Sleep spindles

Sleep spindles are discrete events consisting of 11-16 Hz oscillations (the precise frequency range varies across subjects) recorded primarily in stage 2 non-REM sleep, and to a lesser extent in stage 3 non-REM sleep (Berry et al., 2012). Spindles display a characteristic waxing and waning pattern in amplitude, and generally last between 0.3 and 3 s, recurring every 5–15 s (Bonjean et al., 2012; Leresche et al., 1991). Sleep spindles arise from the activity of thalamocortical circuitry. They have become a subject of study for their potential roles in memory consolidation and other cognitive functions (Sejnowski and Destexhe, 2000; Schabus et al., 2004; Fogel et al., 2007), as well as in psychiatric and neurological disorders (Ferrarelli et al., 2007; Petit et al., 2004; Ktonas et al., 2007).

Numerous methods for automated spindle detection have been proposed, most of which rely on spectral analysis in some form (Warby et al., 2014; O'Reilly and Nielsen, 2015). Here, we propose an alternative approach using a nonlinear time-domain algorithm which is computationally fast and therefore capable of detecting spindles in real time.

* Corresponding author at: Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA.
E-mail addresses: asampson@ucsd.edu, asampson@salk.edu (A.L. Sampson).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jneumeth.2019.01.009
Received 16 May 2018; Received in revised form 4 January 2019; Accepted 20 January 2019
Available online 30 January 2019
0165-0270/

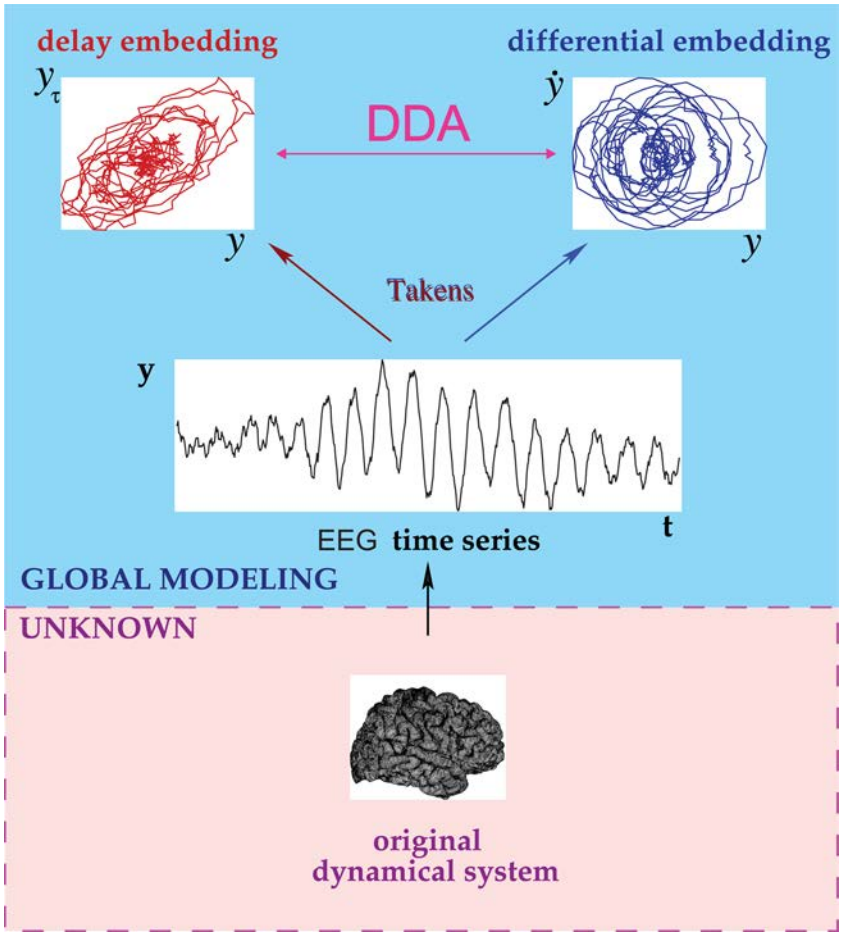

Fig. 1. Delay differential analysis (DDA). For an unknown dynamical system (such as the brain) from which we can record a single variable over time (such as ECoG data), embedding theory states that we can recover the nonlinear invariant properties of the original system. DDA combines delay and differential embeddings in a functional form which allows time-domain classification of the data. For a given polynomial model form, we estimate the coefficients and least-squares error, which form a low-dimensional feature space. This figure is adapted from Lainscsek and Sejnowski (2015).

1.2. Delay differential analysis

Delay differential analysis (DDA) is a time-domain classification framework based on embedding theory in nonlinear dynamics (Kremliovsky and Kadtke, 1997; Lainscsek et al., 2013). An embedding reveals the nonlinear invariant properties of an unknown dynamical system (here the brain) from a single time series (here intracranial recordings). The embedding in DDA serves then as a sparse nonlinear functional basis onto which the data are mapped (Fig. 1). Since the basis is built on the dynamical structure of the data, preprocessing (such as filtering) is not necessary. DDA yields a small number of features (around 4), far fewer than traditional spectral techniques, which provide a power at each frequency (often 100–200 frequencies). In either case, the size of the feature set might vary depending on the parameters used. Also, either set of features can be combined or collapsed to yield a measure that can be thresholded. However, working with a constrained feature space is often desirable. This approach greatly reduces the risk of overfitting, and therefore helps to ensure that a model that was selected using a single EEG channel from one subject can be applied to a wide range of data from different subjects, channels, and recording systems.

One can also view DDA models as sparse Volterra series (Volterra, 1887, 1959). A general nonlinear real-valued function can be expressed as a Taylor series expansion of functionals of increasing complexity around a fixed point. Rather than retain all low-order terms in the expansion, DDA imposes restricted complexity on the analysis by using a low-dimensional sparse delay differential equation (DDE) model. In a model of this type, linear and nonlinear components of the data are analyzed in an interconnected manner. This reduces the computational load, and further, by leaving some of the non-relevant dynamics unmodeled, it is possible to greatly reduce the effect of artifacts and other signals unrelated to the particular classification task of interest.

DDEs combine differential with delay embeddings as a functional embedding where (non-) linear polynomial functions of the delay terms are used (Lainscsek et al., 2017). The general form of the DDEs is

$$\dot{x}(t) = \sum_{i=1}^{I} a_i \prod_{n=1}^{N} x_{\tau_n}^{m_{n,i}} \quad \text{for} \quad \tau_n,\, m_{n,i} \in \mathbb{N}_0 \tag{1}$$

where I is the number of monomials in the model, N is the number of delays, $m_{n,i}$ is the order of the nth term in the ith monomial, and x_{τ_n} represents $x(t-\tau_n)$. The time derivative of the data, $\dot{x}(t)$, is estimated with a weighted center derivative (Miletics and Molnárka, 2005):

$$\dot{x}(t) = \frac{1}{2M} \sum_{m=1}^{M} \frac{x(t+m) - x(t-m)}{m} \tag{2}$$

where M is the number of points used.

For a given model, we compute a small set of features, which are the estimated coefficients a_i in Eq. (1) as well as the least-squares error. The error is defined as:

$$\rho = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\dot{x}_{t_k} - \sum_{i=1}^{I} a_i \prod_{n=1}^{N} x_{\tau_n,t_k}^{m_{n,i}} \right)^2} \tag{3}$$

where K is the number of time points, and x_{τ_n,t_k} represents $x(t_k - \tau_n)$.

2. Methods

2.1. Data

DDA was applied to laminar, stereoelectroencephalogram (sEEG),

Table 1
Human-marked spindle properties for the 15 recordings.

Subject	Channel[b]	Type	Scoring	f_s (Hz)	Number	Mean duration (s)	Mean peak freq. (Hz)
L1	1–23, left frontal	Laminar	I	2000	144	1[a]	15.0580
L2	1–23, right frontal	Laminar	I	2000	48	1[a]	11.8063
L3	1–23, right frontal	Laminar	I	2000	137	1[a]	12.8836
L4	1–23, right frontal	Laminar	I	2000	50	1[a]	12.4320
L5	1–23, right temporal	Laminar	I	2000	72	1[a]	13.2750
S1	1 (RCIN3)	sEEG	II	500	57	0.84	12.5395
S1	2 (LCIN4)	sEEG	II	500	135	0.91	12.8115
S1	3 (LSF6)	sEEG	II	500	47	0.72	12.6363
S2	1 (LCIN3)	sEEG	II	500	213	1.79	12.7073
S2	2 (LSF3)	sEEG	II	500	218	1.42	13.1963
S2	3 (RCIN5)	sEEG	II	500	146	1.25	12.9723
S2	4 (LFR1)	sEEG	II	500	227	1.57	12.3713
S3	1 (OF7)	sEEG	II	500	138	0.87	12.7769
S4	1 (RPF5)	sEEG	II	512	152	1.15	12.7569
S4	2 (ROF4)	sEEG	II	512	81	0.98	13.9615
S5	1 (RAF6)	sEEG	II	512	124	0.96	13.0326
E1	1 (GR28)	ECoG	II	512	82	1.05	12.4093
E1	2 (GR53)	ECoG	II	512	13	1.36	11.7415
E1	3 (GR38)	ECoG	II	512	92	1.18	13.2799
E2	1 (AGR52)	ECoG	II	1024	47	0.71	12.1440

[a] The mean duration cannot be determined from Type I scoring because only a single time point was marked across all channels (1–23). One second of data is designated as spindle data for structure selection.

[b] RCIN – right cingulate, LCIN – left cingulate, LSF – left subfrontal, LFR – left frontal, OF – orbitofrontal, RPF – right posterior frontal, ROF – right orbitofrontal, RAF – right anterior frontal, GR – grid (subject E1 grid channels 28, 38, and 53 were all located over posterior frontal cortex with 28 the most inferior and 53 the most superior), AGR – anterior grid (subject E2 anterior grid channel 52 was located over middle posterior frontal cortex).

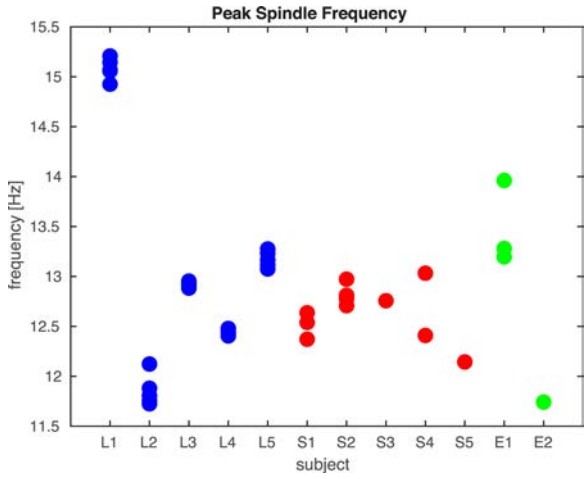

Fig. 2. Spindle frequencies. For each of the five laminar, five sEEG, and two ECoG subjects, the peak frequency (between 11 and 17 Hz) was computed for all human-marked spindles and the mean peak frequency across all spindles is plotted as one point for each channel. For laminar subjects, five of the channels are plotted – spindles were marked based on evaluation of all channels. For the sEEG and ECoG subjects, spindles were marked on an individual-channel basis, and all scored channels are plotted. Color indicates the type of recording. Note that laminar recordings were collected from cortex identified as probably epileptogenic.

and electrocorticogram (ECoG) recordings from patients with intractable epilepsy.

The laminar recordings studied here come from five patients, designated L1 to L5. Recordings and data were obtained under Institutional Review Board (IRB) approval with informed consent from participants in accordance with the Declaration of Helsinki.

The additional recordings used for this study consisted of sEEG (depth electrode) recordings from five patients, designated S1 to S5, and ECoG recordings from two patients, designated E1 and E2, with long-standing pharmaco-resistant complex partial seizures. These recordings used a standard clinical recording system (XLTEK, Natus Medical Inc., San Carlos, CA) with sampling rates of 500, 512, or 1024 Hz. The reference for the sEEG electrodes was an electrode placed over the C2 spinous process on the posterior neck. For the ECoG (cortical surface electrode) recordings, the reference channel was a strip of electrodes located outside the dura mater and facing the skull at a region remote from other grid and strip electrodes. Placement of the intraparenchymal (sEEG) electrodes and subdural electrode arrays was chosen to confirm the hypothesized seizure focus and locate epileptogenic tissue in relation to essential cortical areas, thus directing surgical treatment.

The decision to implant, as well as the electrode targets and the duration of implantation were entirely clinically based with no input from this research study. All data were handled following protocols approved by the IRB of the Massachusetts General Hospital according to National Institutes of Health guidelines.

sEEG data used for this study consist of three channels from subject S1, four channels from subject S2, one channel each from subjects S3 and S5, and two channels from subject S4. ECoG data used here consist of three channels from subject E1 and one channel from subject E2. All data selected for use in this study were exclusively from stage two sleep, during time periods when no seizures were occurring.

2.2. Spindle marking

Both the data used for developing the detector and those used for testing were drawn from human expert-scored intracranial recordings: 23-channel laminar electrodes in five subjects (L1–L5) and single-channel scored sEEG and ECoG recordings from subjects S1–S5 and E1–E2. In the laminar data set, the scorer marked a single time point for each identified spindle based on evaluation of all 23 channels (here designated type I scoring). In the sEEG and ECoG data, the beginning and end of all spindles were marked on the basis a single channel (type II scoring). In type II scoring, therefore, the beginnings of spindles are defined as the point where spindle oscillations become visually apparent to the scorer, and the end is defined as the point where these oscillations are no longer apparent. Also, in type II scoring, the scorer marked all potential spindles, regardless of clarity. By including both types of human scoring as well as a range of spindle quality, we aim to develop a robust detector that can function even with non-ideal data.

Since only a single time point was marked in type I scoring, a

window of one second around each marker was taken as the spindle (that is, the beginning of each spindle was defined as 0.5 s before the mark and the end was defined as 0.5 s after the mark), and a wider window of 1-3 s around each marker was excluded from classification as non-spindle data (only data at least 1.5 s before or after a mark were considered non-spindle data). Table 1 summarizes the properties of the marked spindles in both data sets: the recording type (laminar electrodes, sEEG, or ECoG), the scoring type (I or II), the sampling rate f_s, the number of marked spindles, the mean spindle duration, and the mean peak frequency (between 11 and 17 Hz) for all spindles in each recording. Since type I scoring involved marking spindles on the basis of multiple channels, the peak frequencies are computed as the mean of the peak frequency across the five channels in which spindles are most visually apparent. The peak frequencies for all channels for each subject are plotted in Fig. 2.

2.3. Supervised structure selection

Structure selection of the model ultimately relied on data from one channel from one subject. Since DDA uses specific time delays, adjustments need to be made for sampling rate, and to facilitate this, the model (polynomial form and delays) was selected using data with the lowest sampling rate in the available data set (this allows for easy adjustment to higher sampling rates). Here, we used an sEEG recordings sampled at 500 Hz. Data from these subjects and channels were divided into half-second epochs and marked as spindle or non-spindle based on how each epoch had been marked by a human expert in the manner described above. Among these 500 Hz recordings, the one for which spindle and non-spindle epochs proved most separable was used to select a model for use with new data.

In order to select the model from these training data, the set of models to be considered was first subjected to constraints based on model forms that had proven effective in previous applications of DDA, ensuring the sparsity of the model. The general form of the model shown in Eq. (1) was constrained to two delays ($N \leq 2$), three terms ($I = 3$), and up to third-order nonlinearities ($\Sigma_n m_{n,i} \leq 3$). This resulted in a total of 188 unique DDE model forms, upon which we performed an exhaustive search. The delays τ_1 and τ_2 were allowed to vary between approximately 1 and 80 ms at intervals of $1/f_s$.

We performed repeated random subsampling cross-validation (Kohavi et al., 1995) to evaluate the performance of each model. This method involves repeatedly dividing the data at random into training and testing sets. (Note that throughout we use the terms “training” and “testing” to refer to these repeated random splits of the data for cross-validation. New data, not used in the structure selection of a particular model, are referred to as “validation” data.) This prevents overfitting of the model and ensures generalizability. Here, the repeated random splits were carried out for the model selection data, assigning 70% of spindle and non-spindle epochs to the training set, and the remaining 30% to the testing set. Using the model coefficients $a_{k,i}$ and error ρ_k obtained from each epoch k of the training data, we used the human expert-scored labels l_k (i.e. 0 for non-spindle and 1 for spindle) to obtain a vector of weights W for the features by finding a least-squares solution to:

$$\begin{pmatrix} 1 & a_{1,1} & a_{1,2} & a_{1,3} & \rho_1 \\ 1 & a_{2,1} & a_{2,2} & a_{2,3} & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & a_{k,1} & a_{k,2} & a_{k,3} & \rho_k \end{pmatrix} W = \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_k \end{pmatrix}. \quad (4)$$

The additional constant term avoids constraining the separating hyperplane to pass through the origin in feature space. The weights W can be applied to the features computed from the testing data which provides a one-dimensional distance D from an optimal hyperplane of separation between spindle and non-spindle feature sets. We can evaluate how well this distance corresponds to the human expert-scored labels of the testing data by computing the area under the receiver

Table 2
DDA spindle detection performance on all recordings.

Subject	Channel	A'	F_1	False discovery rate	False negative rate
L1	11	0.6023	0.2685	0.5323	0.8117
L2	11	0.6934	0.2991	0.7107	0.6903
L3	11	0.7423	0.2892	0.4701	0.8011
L4	11	0.7784	0.4948	0.5590	0.4365
L5	11	0.7529	0.3679	0.6682	0.5872
Laminar mean		0.7139	0.3439	0.5881	0.6654
S1	1 (RCIN3)	0.8785	0.5404	0.5924	0.1983
S1	2 (LCIN4)	0.9066	0.7685	0.2340	0.2290
S1	3 (LSF6)	0.8716	0.4345	0.6953	0.2428
S2	1 (LCIN3)	0.9120	0.3464	0.0380	0.7887
S2	2 (LSF3)	0.9170	0.5410	0.0265	0.6254
S2	3 (RCIN5)	0.8514	0.5601	0.1723	0.5768
S2	4 (LFR1)	0.9262	0.3970	0.0386	0.7499
S3	1 (OF7)	0.9062	0.8211	0.1718	0.1858
S4[a]	1 (RPF5)	0.4886	0.0749	0.8372	0.9514
S4	2 (ROF4)	0.8421	0.7201	0.1541	0.3731
S5	1 (RAF6)	0.8186	0.6290	0.3222	0.4133
sEEG mean		0.8830	0.5758	0.2445	0.4383
E1	1 (GR28)	0.8385	0.6081	0.3954	0.3884
E1[a]	2 (GR53)	0.6254	0.0462	0.9722	0.8636
E1	3 (GR38)	0.7726	0.5128	0.4000	0.5522
E2	1 (AGR52)	0.8112	0.3478	0.7692	0.2941
ECoG mean		0.8074	0.4896	0.5215	0.4116

[a] These recordings are excluded from the means and further analysis due to poor quality.

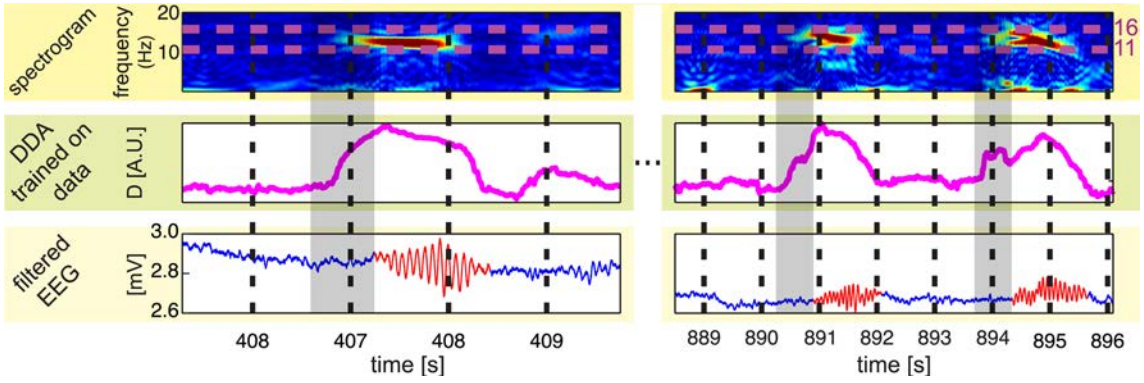

Fig. 3. Spindle detection. The lowest row in the plot shows the data with the spindles in red, as marked by a human expert. In the middle row, a DDA spindle detection output (trained on one channel from a different subject) is shown. We also show the spectrograms (in the top row) for reference. The gray-shaded regions indicate the width of the time windows used for computing both the DDA features and the spectrogram (650 ms). Since we plot the time points on the x-axis for the start points of the sliding windows, all points within a shaded region use windows that include some amount of spindle data.

operating characteristic (ROC) curve or F_1 score. The ROC is constructed by plotting the hit rate against the false alarm rate for various spindle detection thresholds for D. The area under the curve defined by the plotted points, A', should be equal to 0.5 for random chance detection, and 1 for perfect separation of the groups (Hand and Till, 2001). A' can be obtained by taking

$$A' = \frac{S_0 - n_0(n_0 + 1)}{2n_0 n_1} \tag{5}$$

where n_0 and n_1 represent the number of points in each of two classes labeled 0 and 1 (here, non-spindle and spindle epochs), and S_0 is obtained by first ranking all points by their probability of being classified as 0, then summing the ranks of the true class 0 points. In practice, once a specific model form has been selected, it is often sufficient to use a single feature for classification.

While A' is useful for structure selection of the DDA model, we evaluate final performance with another measure, the F_1 score, which is more widely used for evaluating spindle detection (Dice, 1945; Sørensen, 1948). F_1 scores are computed from the confusion matrix according to:

$$F_1 = \frac{2\mathrm{TP}}{\mathrm{FN} + \mathrm{FP} + 2\mathrm{TP}} \tag{6}$$

where TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives. For this purpose, the human scoring is considered the "ground truth". F_1 scores are used in Section 3.1 for comparison between the outputs of several spindle detection methods. As additional measures, we also compute the false discovery rate ($\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}$) and false negative rate ($\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}}$).

The cross-validation was repeated 100 times and the maximal A' was used to select the optimal model form and values of the delays. Using this procedure, for spindle detection in the laminar, sEEG, and ECoG data at all sampling rates, an effective DDE model is:

$$\dot{x} = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1}^2 \tag{7}$$

with $\tau_1 = 16\,\delta t = 32$ ms and $\tau_2 = 25\,\delta t = 50$ ms for 500 Hz data. For spindle detection, we find that the single feature a_2 provides sufficient information for good detection performance. In general, the threshold for spindle detection is set to 1.2 standard deviations above the mean of a_2. This threshold has been empirically determined to provide good agreement with human scoring and was fixed throughout.

Despite the fact that these data come from subjects with different types of electrodes and different sampling rates, it is possible to obtain spindle detection which agrees with human scoring across multiple recordings as well as multiple human scorers would tend to agree with each other (Basner et al., 2008). Because we use nonlinear models, all terms are connected and linear as well as nonlinear terms contain both linear and nonlinear information. For this reason the delays do not correspond to particular frequencies as one might expect (Lainscsek and Sejnowski, 2015). Adjustments need to be made for data with different sampling rates. In order to apply a selected DDA model to data with a higher sampling rate, we need to change the delays and derivatives in the following way: The delays can be just the approximate multiples (e.g. from 500 Hz to 1000 or 1024 Hz they would be doubled). For the derivatives we keep the number of total points constant but take for this example every second data point. For data with lower sampling rates (e.g. the DREAMS data in Section 3.1), results can only be obtained by upsampling the data to the minimum sampling frequency of 500 Hz before applying the model.

2.4. Application to full-time data

Having selected a model form and delay pair according to the above procedures, we compute the corresponding a_2 coefficient in sliding time windows across the full length of all recordings. We use windows of length around 650 ms, shifted by around 200 ms per step. Since the number of spindle and non-spindle epochs in the training data are not equal, the optimal threshold for spindle detection may vary slightly between recordings. Nevertheless, for the sake of testing a fully automated method, we maintained the aforementioned 1.2 standard deviation above mean a_2 threshold for all results shown here. The beginning of each detected spindle is therefore defined as the point at which the normalized a_2 value increases this threshold, and the end is defined as the point at which it subsequently decreases below the threshold. (Note that threshold-setting does not affect A', since this is a threshold-independent measure, but does determine the F_1 scores, which are computed from the confusion matrix for a particular threshold.) As a final step, any threshold crossings less than 300 ms in length are excluded and marked as non-spindle. The remaining threshold-crossings are the identified spindles. We evaluate detector performance by comparing these time points identified as spindle by the detector with those identified by the human expert.

3. Results

Applying the detector to laminar, sEEG, and ECoG data, we obtain a mean area under the ROC curve, A', of 0.82 and a mean F_1 score of 0.50. For the laminar data, we take just one central channel from each electrode array for evaluating all methods. Since these data were scored based on all channels, but some superior and inferior channels lacked clearly visible spindles, one of the channels (channel 11) with apparent spindles was chosen for evaluating spindle detection performance. All available (individually scored) sEEG and ECoG channels were used. For comparison, DDA frequency-band detectors (discussed in Appendix A) for 11–14 Hz and 11–17 Hz yield mean A' values of 0.72 and 0.77 and mean F_1 scores of 0.21 and 0.18 respectively. Such a difference in performance indicates that in addition to the frequency characteristics of spindles, nonlinear information might also be relevant. Fig. 3 shows the output the data-trained DDA spindle detector. Since the data-trained DDA detector shows higher agreement with human scoring than the frequency-based DDA detector, it is used exclusively for the remainder of the manuscript.

The A' values, F_1 scores, false discovery rates, and false negative rates for the DDA spindle detector on all subjects are listed in Table 2. Note that in Section 3.1, F_1 scores are used to compare methods. Where cross-recording averages are reported, two recordings are excluded since all automated detectors perform poorly, and these were originally selected as recordings that were difficult to score.

Table 3
Comparison of detection methods for all data.

Method	Mean percentage of human-scored spindles	Mean length (s)	Mean F_1	False discovery rate	False negative rate	CPU[a] time (s) per recording
Mölle	105.0457	0.4871	0.4871	0.2856	0.5994	30.5645
Martin	141.9600	0.4754	0.4754	0.3427	0.5441	2.5615
Andrillon	46.3362	0.4028	0.4028	0.2078	0.7022	0.3922
Hagler	116.2967	0.4591	0.4591	0.2963	0.6225	1.8177
DDA	89.8979	0.4970	0.4970	0.3861	0.4969	1.6389

[a] All methods were implemented in MATLAB 9.4 (R2018a) and tested on the same 12-core (Intel Xeon X5690 @ 3.47 GHz) system. The DDA detector calls an executable written in C for a key step in the procedure.

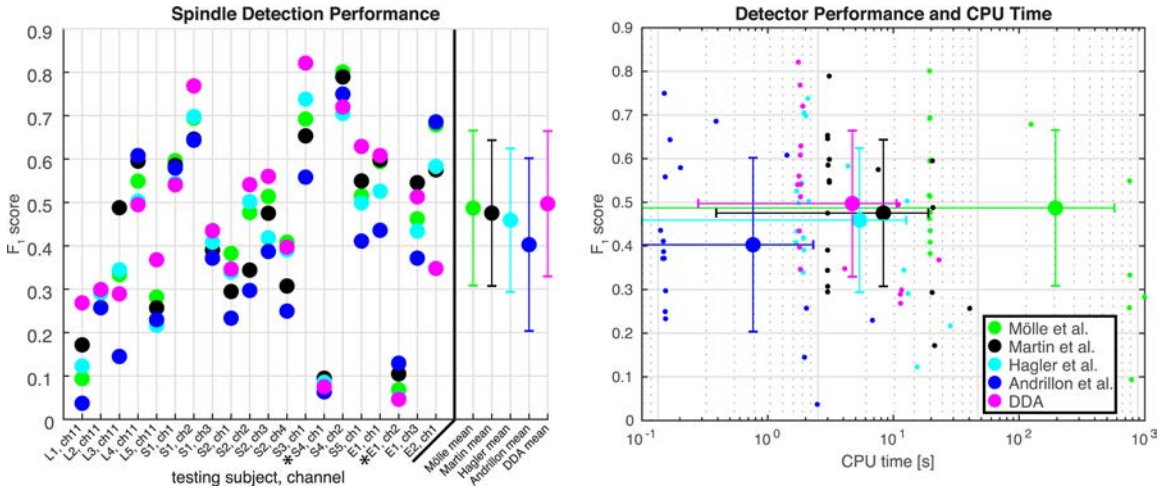

Fig. 4. Detection methods comparison. In the left panel, F_1 score is plotted for a set of automated spindle detection methods and DDA for the various laminar, sEEG, and ECoG recordings. The means (points) and standard deviations (bars) across all recordings for each detector are plotted at the far right – these exclude two recordings (denoted by *) of poor quality for which all methods yield low performance. These recordings are also omitted from the right panel. At right, the F_1 score for all recordings is plotted against CPU time for each detection method. Each detector was run on 20 intracranial recordings, the mean across all recordings (except the two noted exclusions) is plotted with a larger marker, standard deviations across all recordings are plotted as bars in both CPU time and F_1 score. Note that not all recordings are of equal length, so some variation in the CPU time is to be expected.

3.1. Comparison with established methods

Warby et al. (2014) presented a comparison of several automated methods for spindle detection with scoring by human experts and non-experts. Here, we compare the DDA spindle detector to two of the automated methods considered there (Mölle et al., 2002; Martin et al., 2013) and a modified version (Andrillon et al., 2011) of a third (Ferrarelli et al., 2007), as well as an additional method designed for intracranial data (Hagler et al., 2018). Warby et al. used two additional detectors (Bódizs et al., 2009; Wendt et al., 2012) which are excluded here due to their reliance on the comparison of specific channels from a standard EEG montage, making them unsuitable for use with intracranial recordings from disparate locations.

It is important to note that for all of these methods, spindle detection performance may be lower here than with some other data, since no preprocessing or artifact removal steps have been applied here prior to the core processing steps for spindle detection intrinsic to each method. Further, these data present a mix of recordings of different quality and spindle clarity, as evaluated by human expert scoring.

Mölle et al. used a 12-15 Hz bandpass finite impulse response (FIR) filter and subsequently computed a root mean square (RMS) signal with 50 ms time resolution and a 100 ms time window from the filtered data. Spindles were then detected using a thresholding procedure, with beginning and end threshold crossings between 0.4 and 1.3 s required for spindle detection. This threshold was set automatically by the algorithm for each subject as originally published, but was always greater than 5 μV (Mölle et al., 2002).

The approach of Martin et al. was similar: data were first bandpass filtered from 11 to 15 Hz using an FIR filter applied both forward and reverse. The RMS of the signal was then computed using 0.25 s windows. The threshold for spindle detection was set at the 95th percentile and required two consecutive RMS time points (corresponding to 0.5 s) for a spindle (Martin et al., 2013).

We also use a slightly modified version of the detector of Andrillon

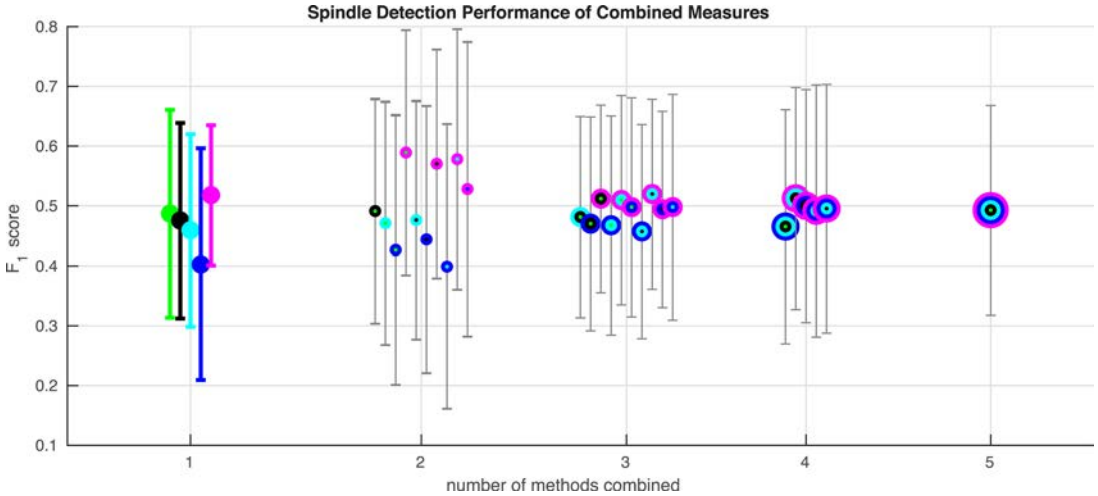

Fig. 5. Combining features from the various methods. Spindle detection measures from the various tested methods were combined by taking a mean at each time point, and the agreement of these averaged measures with human scoring was evaluated via F_1 score. Two recordings with poor detector performance for all methods were omitted here. Colors correspond to the different methods, when methods are combined, concentric circles corresponding to the combined measures are plotted at one point. For all methods and combinations of methods, the mean across all recordings is shown. Error bars represent the standard deviation across recordings. Mean F_1 scores for these combinations of detectors are also shown in Table 4. It is noteworthy that there is a significant boost in detection performance only when combining DDA with any one of the spectral methods. No other combination of methods provides such a boost.

Table 4
Combining detection measures from the various methods. The highest-performing combinations of detectors are marked in bold.

# combined	Mölle et al.	Martin et al.	Hagler et al.	Andrillon et al.	DDA	F_1 score
1	0.4871	0.4754	0.4591	0.4028	**0.5179**	
	X	X				0.4912
	X		X			0.4709
	X			X		0.4264
	X				**X**	**0.5892**
2		X	X			0.4761
		X		X		0.4439
		X			X	0.5704
			X	X		0.3991
			X		X	0.5781
				X	X	0.5280
	X	X	X			0.4813
	X	X		X		0.4701
	X	X			X	0.5119
	X		X	X		0.4674
3	X		X		X	0.5098
	X			X	X	0.4978
		X	X	X		0.4571
		X	X		X	**0.5197**
		X		X	X	0.4943
			X	X	X	0.4979
	X	X	X	X		0.4653
	X	X	X		X	**0.5125**
4	X	X		X	X	0.5000
	X		X	X	X	0.4917
		X	X	X	X	0.4954
5	X	X	X	X	X	0.4927

et al., itself a modified version of the method of Ferrarelli et al. (2007). Putative spindles were identified by applying a zero-phase fourth-order Butterworth bandpass filter for 9–16 Hz. Instantaneous amplitude was computed using a Hilbert transform, and the threshold for detection was set at three standard deviations from the mean, with a threshold for the beginning and end of spindles set at one standard deviation. Only events with durations between 0.5 and 2 s were marked as spindles, and spindles separated by less than 1 s were merged.

Finally, we also apply a method developed for and previously applied to intracranial recordings of the type we consider here, which was developed by Hagler et al. This technique relies on an initial detection based on instantaneous power in the spindle band (11–17 Hz) using a smoothed wavelet convolution. Any initially identified spindles under 0.5 s in duration are excluded. Further, the ratio of Fourier power in the spindle band relative to power in the 4–9 Hz range is used to remove artifacts and weak spindles. (Hagler et al., 2016).

In order to compare these various techniques with differing methodologies, we convert the raw outputs of each technique to a binary index of spindle or non-spindle for each time point. These binary detection indices are then compared by computing the F_1 score of each method against the human expert-marked spindles. The mean across subjects of the number of spindles detected (expressed as a percentage of the number of spindles marked by the human expert), spindle length, F_1 score, and false positive and negative rates (relative to human expert scoring) for each of these methods are shown in Table 3. The F_1 scores as well as CPU time for all methods and recordings are shown in Fig. 4. DDA provides the highest average F_1 score and the second lowest average CPU time.

Notably, as shown in Fig. 2, one of the recordings (L1) had a higher mean peak spindle frequency than all others. That recording has a low F_1 score (see Fig. 4) for all comparison methods. DDA, in contrast, detected those spindles relatively well since the goal was to detect dynamical patterns in the data.

To assess the advantage provided by using DDA features in addition to spectral features, Fig. 5 and Table 4 show the mean F_1 scores for various combinations of the different detection methods. Of note is the fact that combining the DDA measure of spindle activity with other measures generally provides a better measure than combining two or more spectral methods, since it provides different information. Note that the F_1 scores for the DDA detector alone in Fig. 5 and Table 4 do not match exactly the scores in the earlier figures and tables. This is due to an additional step of averaging the DDA features across the overlapping windows at each time point. This provides a measure with time resolution equal to original data which can then be combined with other measures on a point-by-point basis.

Finally, for comparison, DDA and the other detection methods were applied to the DREAMS dataset, collected and made available by

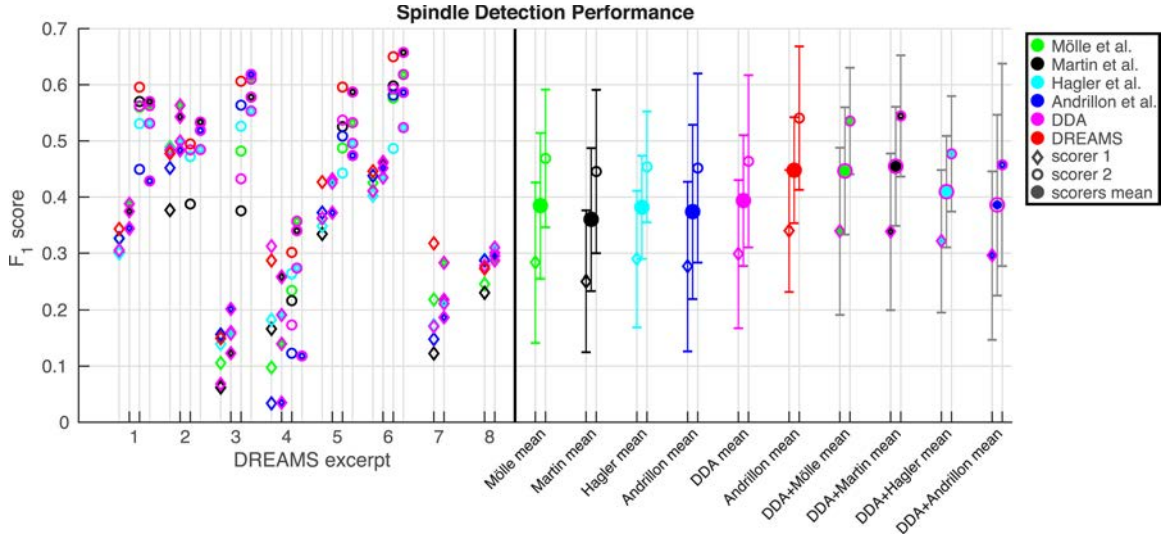

Fig. 6. Spindle detection on DREAMS data. F_1 score is plotted for a set of automated spindle detection methods and DDA for the eight surface EEG excerpts included in the DREAMS data set. For six of the eight excerpts, two human experts scored the data. For these six recordings, F_1 scores based on the first expert's markings are plotted as diamonds at left, and the scores based on the second expert's markings are plotted as open circles at right. The means (diamonds and open circles) and standard deviations (bars) across all recordings for each detector's agreement with both experts are plotted at the far right with the same markers, and the means of each method's agreement with both scorers are plotted as larger filled circles. Combinations of the other measures with DDA, as shown in Fig. 5 are shown with the colors for each of the methods combined. In addition to the six methods shown previously, we also show here the F_1 scores of the automated spindle detections included with the DREAMS data with both human experts in red.

Université de Mons, TCTS Laboratory (Stéphanie Devuyst, Thierry Dutoit) and Université Libre de Bruxelles, CHU de Charleroi Sleep Laboratory (Myriam Kerkhofs) (Devuyst et al., 2011). The DREAMS data consist of surface EEG with spindles marked by two human experts. Using these data allow the above detection methods to be compared on surface EEG data, as well as compared to automated spindle detections from a method implemented by the original authors and made available with the data. This technique is based on bandpass filtering and applying a recording-specific threshold. While the DREAMS automated detections provide better agreement with the human scorers than the intracranial data-trained DDA detector or any of the other tested methods (Devuyst et al., 2011). We cannot compare directly with this method since only the data and automated detections are available, and not the code. We therefore cannot test the DREAMS method on our dataset. Further, as can been seen in Fig. 6, there is also a large discrepancy between the two human scorers, with one scorer also only having scored six of the eight subjects. Issues with the scoring of these data were previously noted by O'Reilly and Nielsen (2015). Further, it is noteworthy that DDA still provides reasonable spindle detection after structure selection based solely on intracranial data. Most significantly, however, we also show the combinations of two detectors (as shown in Fig. 5). For these data, combining our DDA measure with the measure produced by the method of Martin et al. provides the highest average agreement with the two human scorers among all tested methods and combinations of methods.

4. Discussion and conclusions

DDA is a powerful novel tool for detecting sleep spindles in EEG and intracranial recordings. DDA requires minimal pre-processing of signals and can be rapidly applied to large datasets. When compared with several well-established and reliable frequency-based methods, DDA provides the highest level of agreement with human scoring (evaluated here with F_1 score). Further, DDA is the second fastest of the tested methods, where the only faster method produces the lowest F_1 scores. DDA therefore holds great promise for real-time applications. We also tested all methods on the publicly available DREAMS data, consisting of surface EEG recordings scored by two expert scorers. Again, DDA provides the highest F_1 score of the previously tested methods when taking the average across both scorers. The automated detections made available with the DREAMS data however, do provide better agreement with the human scorers. It should be noted that the DREAMS data is a small and heterogeneous data set, and therefore somewhat limited for evaluation purposes (O'Reilly and Nielsen, 2015).

An important caveat for the results from intracranial data presented here is that they are based on comparison with the spindle markings by a single human expert. Despite this, the fact that several automated methods produce similar detections indicates that the markings are reasonable. Further, similar results are achieved using the same approaches on an EEG data set scored by two experts. It is also important to note the classic bias that our implementation of other previously published detectors may not be as fully perfected as the novel method developed for this paper. Other implementations on other data and comparing to other human scoring might not produce the same relative performance numbers. However, this is only a concern when looking at each method separately. As shown in Figs. 5 and 6, combining our nonlinear time-domain method with any of the tested spectral-based methods, the performance is increased dramatically, beyond the relatively differences between individual methods. This indicates that spectral and nonlinear methods account for different information in the original signal: DDA looks for dynamical differences while spectral methods look for content in a specific spindle frequency band.

Combining two spectral measures does not provide the same advantage as combining linear and nonlinear features. Additionally, we have demonstrated that DDA models built on the data show superior performance to those built to detect specific frequencies, which indicates that using the nonlinear signature of the spindle provides access to additional information. Accessing this type of information could prove especially useful in future studies focused on spindles of different types, or occurring in patients with neuropsychiatric disorders. Finally, it is worth emphasizing again the robustness of DDA measures in general to noise and artifacts due to the sparsity of the feature space. This is a significant advantage for many data sets.

A version of the DDA spindle detector for use on Linux systems using MATLAB has been made available at http://snl.salk.edu/~asampson/SPINDLES/index.html.

Acknowledgements

The authors would like to thank Dr. Werner Doyle for his role in the collection of these data. This work was supported by the Howard Hughes Medical Institute and the Crick-Jacobs Center for Theoretical and Computational Biology, the U.S. Office of Naval Research under Grants N00014-10-1-0072 and N00014-12-1-0299, and the NIH grants R01-EB009282 and R01-NS104368.

Appendix A. Frequency-based spindle detection

All spindle detection techniques DDA is compared to are based on decomposing the signal into oscillatory components, and therefore have very different assumptions: while DDA assumes nonlinearity of the (unknown) underlying dynamical system, spectral methods assume linear superposition of stationary sinusoids. To interpret the differences in detector performance we need to answer the question of what is gained by using nonlinear instead of linear analysis.

In Lainscsek and Sejnowski (2015) a connection between DDA and spectral analysis was made: a one term linear DDE can be used for frequency detection while a one term nonlinear DDE can detect frequency/phase couplings in the time domain. A DDE with linear and nonlinear terms will have vanishing nonlinear coefficients for purely harmonic signals. For data that contain nonlinear couplings between frequencies or other nonlinear signal components, linear as well as nonlinear terms contain both linear and nonlinear information. Superposition does not work due to nonlinearities in the model. Therefore no connection between frequencies and delays can be made for real-world signals that are generally nonlinear.

Applying the same three-term, nonlinear DDE used for the spindle data to simulated data (noise-diluted sinusoids) can serve as a test of what can be gained by adding nonlinear information, and as a bridge between this technique and traditional wavelet or other spectral methods. The effectiveness of the frequency detector at detecting spindles is also informative as to how much of the relevant dynamical information is related to the dominant frequencies, which is of interest since many spindle detection techniques rely on spectral analysis (Warby et al., 2014).

The DDA frequency detector relies on the same structure selection framework as the data-trained spindle detector, but the DDE model form is fixed to match the model selected using the real data, and only the values of the delays are selected based on the simulated data. For the purposes of comparison with the data-trained detector, we select for frequency bands in the simulated data that correspond to sleep spindles in the EEG sigma band, defined alternately as 11–14 Hz or 11–17 Hz. By comparing the delays which are most successful at detecting these frequencies with those that are selected for the task of sleep spindle detection, we can gain insight into the information added by nonlinear analysis.

The simulated data is generated according to:

$$S_i = A_i \cos(\omega_i t + \varphi_i) + \epsilon \tag{8}$$

Table 5
Selected delays (τ_1, τ_2) for specified bands, units of $\delta t = 1/f_s$.

f_s	delays [δt]			
	11–14 Hz		11–17 Hz	
	> 11 Hz	< 14 Hz	> 11 Hz	< 17 Hz
2000	(8,105)	(8,105)	(8,69)	(7,39)
1024	(1,44)	(19,4)	(4,37)	(4,20)
512	(23,43)	(8,2)	(17,19)	(10,2)
500	(39,18)	(10,2)	(2,17)	(2,9)

with $\omega_i = 2\pi f_i$ for 9991 equally-spaced frequencies f_i between 0.1 and 100 Hz, equal amplitudes $A_i = 1$, random phases $0 < \varphi_i \leq 2\pi$, and added white noise ϵ with a signal-to-noise ratio of 5 dB. Starting from the full set of frequencies, we divide into nearly-equal groups for training and testing, with training data consisting of frequencies f_i from 0.1 to 100 Hz, and the testing data consisting of frequencies f_i from 0.11 to 99.99 Hz, both sets with 0.02 Hz frequency intervals. This ensures that we validate on slightly different frequencies from the training data, still in the desired range. For our simulated training data, we select data with frequencies f_i in the sigma band. As was the case for the data-driven detector, we train separately for each sampling rate, generating simulated data to match each of the sampling rates in the laminar, sEEG, and ECoG data. We then choose delays for each sampling rate f_s.

Selecting a model to provide sensitivity to specific frequency bands requires an additional step, in that we first select "high-pass delays" which are sensitive to frequencies above the lower bound we wish to set (here, 11 Hz), and then additional "low-pass delays" which are sensitive to frequencies below the upper bound (here, 14 or 17 Hz).

The delays chosen for each sampling rate for each definition of the sigma band (11–14 Hz or 11–17 Hz) are shown in Table 5. Note that in some cases, the same delays can be used in both the "high-pass DDE" and "low-pass DDE", since different weights can be applied to the features to select for different frequency ranges.

As with the data-driven detector, we apply a vector of weights to the features for both the lower and upper bounds, in this case obtaining two values of D, which we call D_1 and D_2. We combine them by summing their absolute values and applying the sign of the lesser of d_1 and d_2:

$$D = \frac{\min(D_1, D_2)}{|\min(D_1, D_2)|}(|D_1| + |D_2|). \tag{9}$$

We will therefore obtain positive values only in the region where both are positive, which should correspond to the "DDA pass band".

Fig. 7 shows the frequency response of the detector on simulated data. Given its strong selectivity for frequencies in the desired range, it was

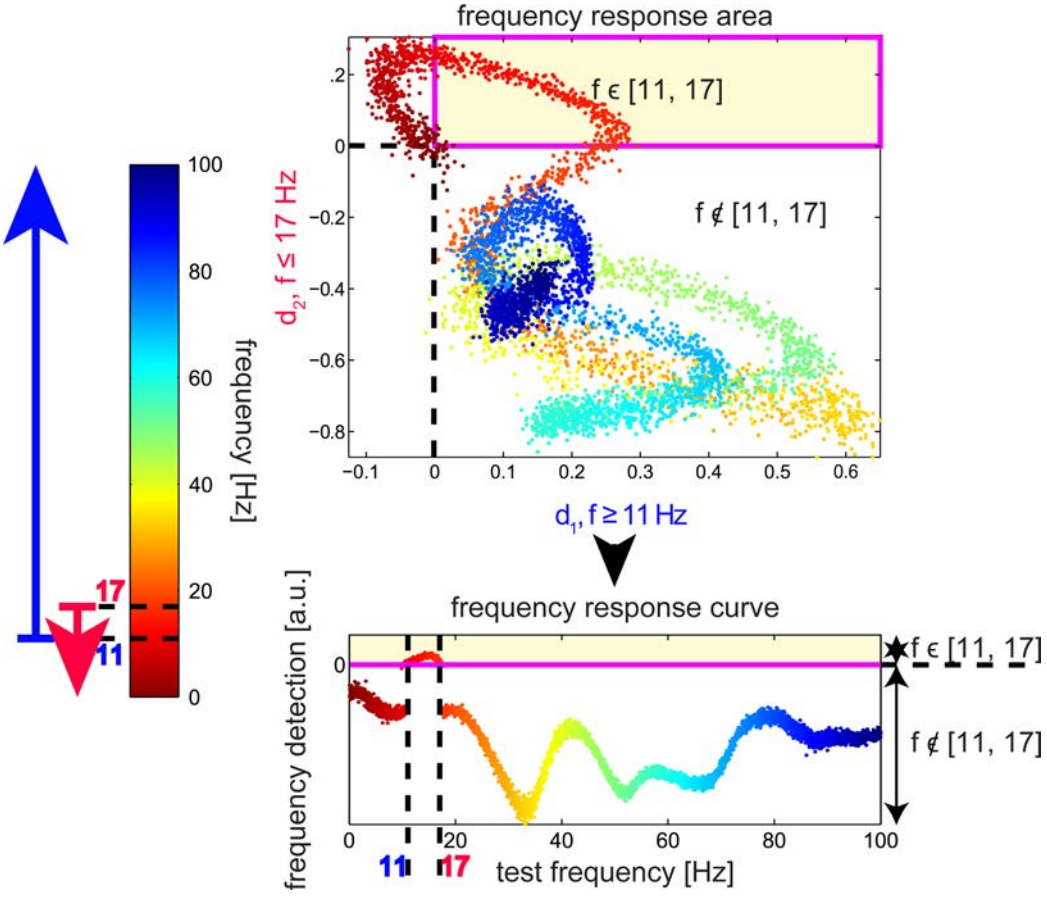

Fig. 7. Frequency band detection. Applying the DDE model with two different delay pairs, one sensitive to frequencies above 11 Hz and one sensitive to frequencies below 17 Hz, we can obtain an output which is positive only in the desired band. In the top panel, the distance from the hyperplane values computed from both DDEs (d_1 and d_2) are plotted for test frequencies ranging from 0.1 to 100 Hz. The frequency of the test data is color-coded according to the color bar at left, from 0.1 (red) to 100 Hz (blue). Points falling into the upper right quadrant (shaded yellow) have positive values for both d_1 and d_2, and we select delays such that only frequencies in the desired range (11–17 Hz) fall into this area. In the lower plot, d_1 and d_2 are combined according to Eq. (9) to obtain a one-dimensional index that is positive only for frequencies in the desired range. This procedure was also used to obtain delays and corresponding weights for frequency ranges 11–14 Hz and 12–15 Hz.

applied to the sleep spindle data as a means of detecting frequency content in the spindle band which uses the same methodology as the data-based DDA spindle detector. This allows for direct comparison between the frequency-based and data-based DDA approaches.

References

Andrillon, T., Nir, Y., Staba, R.J., Ferrarelli, F., Cirelli, C., Tononi, G., Fried, I., 2011. Sleep spindles in humans: insights from intracranial EEG and unit recordings. J. Neurosci. 31 (49), 17821–17834.

Basner, M., Griefahn, B., Penzel, T., 2008. Inter-rater agreement in sleep stage classification between centers with different backgrounds. Somnol. Schlafforschung Schlafmedizin 12 (1), 75–84.

Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Marcus, C.L., Vaughn, B.V., et al., 2012. The AASM Manual for the Scoring of Sleep and Associated Events. Rules, Terminology and Technical Specifications. American Academy of Sleep Medicine, Darien, Illinois.

Bódizs, R., Körmendi, J., Rigó, P., Lázár, A.S., 2009. The individual adjustment method of sleep spindle analysis: methodological improvements and roots in the fingerprint paradigm. J. Neurosci. Methods 178 (1), 205–213.

Bonjean, M., Baker, T., Bazhenov, M., Cash, S., Halgren, E., Sejnowski, T., 2012. Interactions between core and matrix thalamocortical projections in human sleep spindle synchronization. J. Neurosci. 32 (15), 5250–5263.

Devuyst, S., Dutoit, T., Stenuit, P., Kerkhofs, M., 2011. Automatic sleep spindles detection-overview and development of a standard proposal assessment method. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC. IEEE. pp. 1713–1716.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Ferrarelli, F., Huber, R., Peterson, M.J., Massimini, M., Murphy, M., Riedner, B.A., Watson, A., Bria, P., Tononi, G., 2007. Reduced sleep spindle activity in schizophrenia patients. Am. J. Psychiatry 164 (3), 483–492.

Fogel, S.M., Nader, R., Cote, K.A., Smith, C.T., 2007. Sleep spindles and learning potential. Behav. Neurosci. 121 (1), 1–10.

Hagler, D.J., Cash, S.S., Halgren, Eric, 2016. Heterogeneous Origins of Human Sleep Spindles in Different Cortical Layers.

Hagler, D.J., Ulbert, I., Wittner, L., Erőss, L., Madsen, J.R., Devinsky, O., Doyle, W., Fabo, D., Cash, S.S., Halgren, E., 2018. Heterogeneous origins of human sleep spindles in different cortical layers. J. Neurosci. 38 (12), 3013–3025.

Hand, D.J., Till, R.J., 2001. A simple generation of the area under the ROC curve for multiple class classification problems. Mach. Learn. (45), 171–186.

Kohavi, R., et al., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai, vol. 14 1137–1145.

Kremliovsky, M.N., Kadtke, J.B., 1997. Using delay differential equations as dynamical classifiers. In: Applied Nonlinear Dynamics and Stochastic Systems near the Millennium. vol. 411, AIP Publishing. pp. 57–62.

Ktonas, P.Y., Spyretta Golemati, H., Tsekou, T., Paparrigopoulos, C.R., Soldatos, P., Xanthopoulos, V., Zervakis, Sakkalis M., Ortigueira, M.D., 2007. Potential dementia biomarkers based on the time-varying microstructure of sleep EEG spindles. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007. IEEE. pp. 2464–2467.

Lainscsek, C., Sejnowski, T.J., 2015. Delay differential analysis of time series. Neural Comput.

Lainscsek, C., Hernandez, M.E., Weyhenmeyer, J., Sejnowski, T.J., Poizner, H., 2013. Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals. Front. Neurol. 4.

Lainscsek, C., Weyhenmeyer, J., Cash, S.S., Sejnowski, T.J., 2017. Delay differential analysis of seizures in multichannel electrocorticography data. Neural computation 29 (12), 3181–3218.

Leresche, N., Lightowler, S., Soltesz, I., Jassik-Gerschenfeld, D., Crunelli, V., 1991. Low-frequency oscillatory activities intrinsic to rat and cat thalamocortical cells. J. Physiol. 441 (1), 155–174.

Martin, N., Lafortune, M., Godbout, J., Barakat, M., Robillard, R., Poirier, G., Bastien, C., Carrier, J., 2013. Topography of age-related changes in sleep spindles. Neurobiol. Aging 34 (2), 468–476.

Miletics, E., Molnárka, G., 2005. Implicit extension of Taylor series method with numerical derivatives for initial value problems. Comput. Math. Appl. 50 (7), 1167–1177.

Mölle, M., Marshall, L., Gais, S., Born, J., 2002. Grouping of spindle activity during slow oscillations in human non-rapid eye movement sleep. J. Neurosci. 22 (24), 10941–10947.

O'Reilly, C., Nielsen, T., 2015. Automatic sleep spindle detection: benchmarking with fine temporal resolution using open science tools. Front. Hum. Neurosci. 9, 353.

Petit, D., Gagnon, J.-F., Fantini, M.L., Ferini-Strambi, L., Montplaisir, J., 2004. Sleep and quantitative EEG in neurodegenerative disorders. J. Psychosom. Res. 56 (5), 487–496.

Schabus, M., Gruber, G., Parapatics, S., Sauter, C., Klosch, G., Anderer, P., Klimesch, W., Saletu, B., Zeitlhofer, J., 2004. Sleep spindles and their significance for declarative memory consolidation. Sleep 27 (8), 1479–1485.

Sejnowski, Terrence J., Destexhe, Alain, 2000. Why do we sleep? Brain Res. 886 (1), 208–223.

Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5, 1–34.

Volterra, V., 1887. Sopra le funzioni che dipendono da altre funzioni. Atti della Reale Accademia dei Lincei 3, 97–105.

Volterra, V., 1959. Theory of Functionals of Integral and Integro-Differential Equations. Dover Publ.

Warby, S.C., Wendt, S.L., Welinder, P., Munk, E.G.S., Carrillo, O., Sorensen, H.B.D., Jennum, P., Peppard, P.E., Perona, P., Mignot, E., 2014. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat. Methods 11 (4), 385–392.

Wendt, S.L., Christensen, J.A.E., Kempfner, J., Leonthin, H.L., Jennum, P., Sorensen, H.B.D., 2012. Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE 4250–4253.