Journal of Neuroscience Methods 313 (2019) 13–23

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Ground truth construction and parameter tuning for the detection of sleep spindle timing in rodents

Blaine Harper[a], Jean-Marc Fellous[a,b,c,*]

[a] Psychology Department, University of Arizona, Tucson, AZ, United States
[b] Biomedical Engineering Department, University of Arizona, Tucson, AZ, United States
[c] Program in Applied Mathematics, University of Arizona, Tucson, AZ, United States

ARTICLE INFO

Keywords:
Sleep spindle
Event detection
Rat
Slow wave sleep
Consolidation

ABSTRACT

Background: The precise detection of cortical sleep spindles is critical to basic research on memory consolidation in rodents. Previous research using automatic spindle detection algorithms often lacks systematic parameter variations and validations.

New Method: We present a method to systematically tune and validate algorithm parameters in automatic spindle detection algorithms using a moderate number of human raters.

Results: Comparing a Hilbert transform-based algorithm to a ground truth constructed by six human raters, this method produced a parameter set yielding an F1 score of 0.82 at 10 ms resolution. The algorithm performance fell within the range of human agreement with the ground truth. Both human and algorithm failures arose largely from disagreement in spindle boundaries rather than spindle occurrence. With no additional tuning, the algorithm performed similarly in recordings from different days or rats.

Comparison with existing methods: Most spindle detection algorithms do not perform systematic parameter variations and validation using a ground truth. To our knowledge, our study is the first in which rodent spindle data is scored by humans, and in which an automatic spindle detection algorithm is evaluated with respect to this ground truth. The rodent data from this study make it possible to compare our algorithm with others previously tested on human data.

Conclusions: We present a general ground truth based approach for the tuning and validation of spindle extraction algorithms and suggest that algorithms aimed at extracting precise spindle timing in rats should use a systematic approach for parameter tuning.

1. Introduction

Different types of neural oscillations occur during waking and sleeping states in the mammalian brain. These oscillations include complex interactions between inhibition and excitation and support many cognitive processes by regulating the activity of cells in the regions in which they occur (Buzsáki, 2006). Sleep spindles—bouts of oscillatory activity of a few hundred milliseconds duration at approximately 10–15 Hz during non-rapid eye movement sleep—are of particular interest in the field of memory consolidation (Luthi, 2014; Niknazar et al., 2015; Jiang et al., 2017). In humans, spindle density increases during sleep after learning in paired-associates tasks (Gais et al., 2002; Schabus et al., 2004) and in procedural memory tasks (Fogel and Smith, 2006). Spindles are thought to support the integration of new information with prior knowledge (Tamminen et al., 2010). Changes in memory performance across the lifetime may be related to changes in spindle physiology observed in older adult humans (Mander et al., 2014; Helfrich et al., 2018).

In rats, spindle density increases during sleep after olfactory paired-associates learning (Eschenko et al., 2006), supporting the validity of rodent models of sleep-dependent memory consolidation. Rodent studies allow for the collection of large data sets that include simultaneously occurring oscillatory population activity and single cell firing from multiple recording sites. Such recordings have identified the timing relationship between sharp wave-ripple oscillations in the hippocampus and spindle onset in the cortex as a potential mechanism for memory consolidation (Siapas and Wilson, 1998). In conjunction with slow wave activity preceding spindle onset, this interaction has since been shown to support, at least in part, recall performance (Latchoumane et al., 2017). Cell activity across regions coordinates

* Corresponding author at: University of Arizona, Department of Psychology, 1503 E University Blvd, room 312, Tucson, AZ, 85721, United States.
E-mail address: fellous@email.arizona.edu (J.-M. Fellous).

https://doi.org/10.1016/j.jneumeth.2018.11.023
Received 2 November 2018; Accepted 28 November 2018
Available online 07 December 2018
0165-0270/

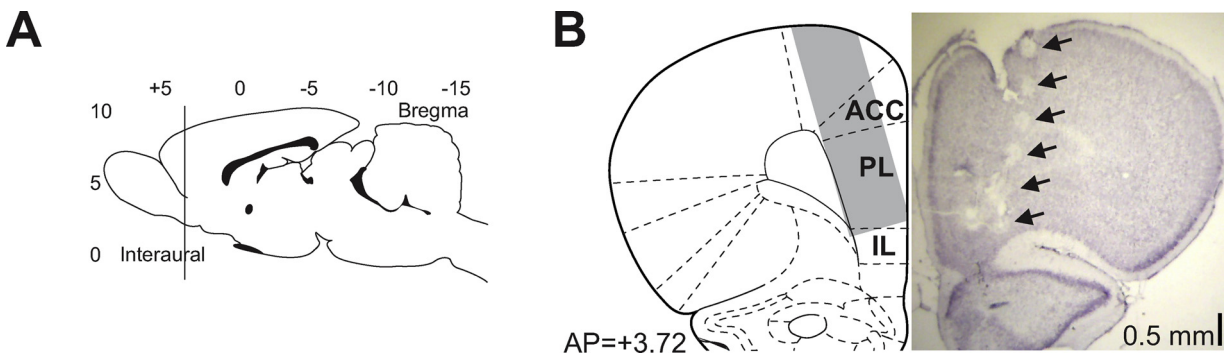

Fig. 1. Methods for spindle recordings. A: Stereotaxic coordinates for hyperdrive implantation. B: Left, cartoon showing approximate extent of tetrode tracks in deep mPFC layers (grey area). Right, photograph of Nissl staining. Arrows indicate electrolytic lesions.

with consolidation-related oscillations as well, producing correlated firing activity in the hippocampus and cortex in conjunction with sharp wave-ripples (Sirota et al., 2003; Wierzynski et al., 2009). However, the distinct functional role of spindles is not yet fully understood.

Defining the contributions of spindles requires a reliable method for their detection (Harper et al., 2016). Spindles are usually detected either as events with delineated boundaries or as increases in spectral power in a specific frequency band. Because spindles with identified start and end times are well suited for the study of timing relationships between spindles and other oscillations or single cells, many studies use event detection algorithms. These algorithms typically use a combination of bandpass filtering, thresholding, and frequency-domain transformations (Gais et al., 2002; Molle et al., 2002; Clemens et al., 2007; Muller et al., 2016). However, such studies often use different parameters and transforms with little or no reference to a procedure for parameter selection or metric assessing detection quality. This makes it difficult to compare detection quality and overall results from different studies. A procedure for detection evaluation is necessary to fully understand the magnitude and types of error in the results of a detection.

As an alternative to examining detection quality, the results from experimental recordings can be normalized using a control recording under the assumption that error rates are stationary and scale linearly with the number of spindles detected. This may suffice in studies of spindle density in short experiments, but could obscure fine temporal relationships in any statistics that are averaged across events. One possible way to overcome this difficulty is to assess the reliability of the detection against an empirically constructed ground truth. For this reason, many questions in the memory consolidation literature stand to gain from a unified evaluation procedure.

A number of studies have compared automatic spindle detection algorithms to manual scoring on data recorded from humans, usually following guidelines established by the American Academy of Sleep Medicine (Iber et al., 2007; Devuyst et al., 2011; Warby et al., 2014; Wallant et al., 2016). The performance of automatic detectors often does not match well that of human spindle identification, even in studies that compare algorithms with the goal of distinguishing which ones perform better than others on a given data set. In some cases, the algorithm parameters chosen for testing may contribute to this problem. Studies of detection algorithms often report on a chosen parameter set intended to be applied to any data (Warby et al., 2014; Lachner-Piza et al., 2018). However, tuning to individual subjects may yield different parameters, indicating that the performance of a single parameter set may vary across recordings (Lajnef et al., 2017). We propose a rigorous framework for spindle detection evaluation in the context of parameter variations that generalize to new datasets.

In addition, prior studies have not evaluated the reliability of automatic detection or inter-rater variability in the context of rodent sleep spindles. Previous studies in humans have recommended agreement between a minimum number of experts or non-experts to ensure a reliable ground truth, or gold standard, for comparison with automatic detection (Wendt et al., 2015; Zhao et al., 2017). Unlike with human sleep, for which a sleep technologist certification exists, expert qualifications are unclear and difficult to assess in the context of rodent sleep. Similarly, multiple databases exist containing expert-scored spindles from human sleep, but no database is yet available for rodent data. In this work, we create such a database using manual scoring by trained non-expert raters and develop a method to validate and adjust rodent spindle detection algorithms.

2. Methods

2.1. Animals

All methods were approved by the Institutional Animal Care and Use Committee of the University of Arizona. Experimental protocols followed all relevant NIH guidelines. Recordings from four adult male Brown Norway rats (7–8 months old) were used. Rats were housed on a reversed 12 h/12 h light cycle in a temperature- and humidity-controlled room and weighed a minimum of 85% ad libitum body weight. All animals were given at least 3 days of acclimation after delivery. Experiments were conducted in a familiar low-light (< 0.5 lx) environment during the first half of their dark phase.

2.2. Surgical procedures

Surgery was conducted using methods previously used (Valdes et al., 2015; Contreras et al., 2018). Rats were implanted under 2–3.0% isoflurane anesthesia with a 14-tetrode hyperdrive (Gothard et al., 1996) targeted to the right medial prefrontal cortex (mPFC, Fig. 1A, B. AP: +3.1, L: +1.1, angle: 9.0°). Two stainless steel electrodes were also implanted in the left or right dorsal CA1 of the hippocampus (AP: -4.5, L: +3.0, DV: 2.2). Another stainless steel electrode was implanted in the neck muscles for electromyographic characterization of sleep and wake states. Rats received two carprofen injections (Rymadil, 3 mg/kg subcutaneous at the end of surgery and 24 h later) and recovered for a minimum of 72 h after surgery. Electrodes were lowered by 300–350 μm per day until the target area was reached.

After recordings were completed, animals were injected with ketamine-xylazine (112 mg/kg ketamine and 14 mg/kg xylazine) under deep isoflurane anesthesia and transcardially perfused with phosphate buffered saline solution followed by 4% paraformaldehyde (PFA). Brains were extracted and stored in fresh PFA. They were transferred to a solution of 30% sucrose and 0.02% sodium azide 72 h after extraction and allowed to sink before sectioning.

2.3. Electrophysiology

Recordings used Teflon-coated nichrome wires (17 μm, Sandvik) gold-plated to an impedance of 500–1000 kΩ. Local field potentials were sampled at 2.4 kHz with 0.1 Hz (high pass) and 500 Hz (low pass)

filtering by a digital Lynx SX amplifier controlled by the Cheetah data acquisition software (Neuralynx, Inc., Bozeman, MT). The electromyogram (EMG) signal was sampled at the same rate with 200 Hz (high pass) and 2 kHz (low pass) digital filtering.

Spindles were recorded at depths of 1400–4500 μm to include the anterior cingulate and prelimbic regions of mPFC.

2.4. Histology

Electrode tracks were verified using electrolytic lesions (5–20 μA; 24 h and 30 min before perfusion). Fifty micrometer frozen sections were obtained using a Cryostat (Leica Biosystems Inc., Buffalo Grove, IL) and Nissl stained using a 0.5% cresyl violet solution (Fig. 1B, right). In some experiments, a series of lesions 634 μm apart was performed as an electrode was slowly retracted from its final depth. These lesions were used to assess tissue shrinkage after Nissl processing (Fig. 1B, right arrows).

2.5. Manual spindle detection

In order to assess the performance of our algorithm, a 'ground truth' (GT) was constructed. Raters scored spindles manually using a graphical user interface developed in the laboratory using MATLAB (Fig. 2A). This interface allows raters to scroll through a continuous recording, skip wake periods, adjust the visualization time window (default 2.0 s), and mark events with boxes and a quality rating (Fig. 2B–F). Quality rating included a category for spindles without a putative K-complex (Fig. 2E). Because spindle frequencies may coincide with late K-complex activity, raters were instructed to include K-complexes when co-occurring with spindles (Fig. 2B–D, F). Panel 1G shows an example of a spindle-free bout of sleep for comparison. The present study combines all spindle categories.

The raters contributing to this study had different amounts of experience relevant to spindle identification. Two raters had prior exposure to rodent sleep electrophysiology and spindle scoring on rodent data, three raters had prior exposure to sleep electrophysiology only, and one rater had no prior exposure to either electrophysiology or spindle scoring. A scoring manual was developed by the authors in order to standardize training (available on our website). All raters attended an in-person meeting to review the scoring manual and learn how to use the graphical user interface. The manual outlined inclusion criteria of spindle frequency (approximately 11–15 Hz), duration (at least 3 cycles, to ensure that detected events were oscillatory and to approximate the 200 ms minimum of the automatic detection), and amplitude, along with categorization based on the presence of a putative K-complex. No maximum spindle duration was imposed due to the presence of dense bouts of spindle activity that may obfuscate spindle start and end times; the study of these sequences is left for further work. The raters did not have access to the power spectrum of the trace and extracted and scored spindles on the basis of the raw voltage trace only.

2.6. Automatic spindle detection

All data processing and analyses were performed in MATLAB (Mathworks, Inc., Natick, MA) using custom-written code freely available from the laboratory website. The automatic detection algorithm used three simultaneously sampled spindle-rich channels. Results using a single channel were similar to results with three channels. Each channel was subsampled by a factor of 8 and filtered using a Butterworth bandpass filter of order 8 (MATLAB: butter(4,…); this was excluded from parameter variations because filter orders of 6 and 10 produced extraction results within 0.25% of order 8 results based on respective F1 scores relative to ground truth). The amplitude of the resulting signal was averaged across the 3 channels and a Hilbert transform was performed and smoothed using a Gaussian kernel. The peaks of the transform were detected above a threshold. Periods above threshold of at least 300 ms duration constituted candidate spindles and were ranked by peak height above baseline. No maximum spindle duration was imposed.

The parameter sweep tested five parameters. For the bandpass filter frequencies: 1 Hz increments were used to vary the low frequency cut-off from 7 Hz to 12 Hz, and the high frequency cut-off from 15 Hz to 20 Hz. This resulted in 36 bandpass filter combinations encompassing the ranges reported in most publications. The same bandpass filter was applied to all 3 voltage traces. A Hilbert transform was applied to the average of the three filtered traces. The width of the Gaussian smoothing applied to the Hilbert transform was varied from 200 to 500 ms in increments of 100 ms. A detection threshold was then applied to identify spindles from peaks in the resulting transform. The threshold was tested at values from 1 through 3.5 standard deviations above baseline in increments of 0.1.

An additional parameter was used to reject a percentage of the lowest detected spindle peaks ranked by their height above baseline. This parameter is useful for comparison between its effects and the impact of raising the detection threshold, as the extraction threshold and rejection criterion originate in the same measure (Hilbert peak height above baseline). There should exist for every increment in extraction threshold some decrease in rejection criterion that yields the same spindle detection output. However, because of the non-linearity of these variations, the precise pairing between these two quantities is difficult to establish analytically. This led us to include it in our results, even though it may not have a unique effect on the quality of the automatic spindle detection.

In total, the sweep included 29,952 different parameter sets in order to fully characterize the interaction between parameters. To manage memory, we divided the work into two chunks including all variations of bandpass filter, smoothing window, and rejection criterion, but separately varying the detection threshold from 1 to 2.3 and from 2.4 to 3.5. Each of these took approximately 5 h to run on a 64-bit Windows machine. The results of the two sweeps were combined into a single stored file for analysis. The potential use of limited sweeps for faster tuning is discussed later in the text.

2.7. Agreement measures

In spindle scoring, time occupied by spindles constitutes a minority of the sleep recording. This means that, as previous studies have pointed out, measures that account for the correct detection rate of true negatives (specificity) and/or chance rater agreement provide little benefit in indicating the extent of error in a detection (O'Reilly and Nielsen, 2015; Tsanas and Clifford, 2015). We verified that agreement between individual human raters and ground truth was similar between F1 score (mean 0.78 ± 0.04) and Cohen's kappa coefficient values (mean 0.75 ± 0.05). Because the F1 score facilitates the examination of sources of error in a detection, we report only the F1 score in our study.

For each rater, spindle start and stop times (in μs) were used to populate an array marking spindle presence in 10 ms bins. Ground truth included bins for which at least three of the six raters agreed. The agreement between ground truth and individual raters was assessed by comparing the bins occupied by spindles using the F1 score calculated as:

$$F_1 = 2 \times \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}} \quad (1)$$

where *recall* is the proportion of true positives out of the sum of true positives and false negatives, and *precision* is the proportion of true positives out of the sum of true positives and false positives. All F1 scores are based on agreement at 10 ms resolution.

As a secondary measure assessing sources of error in the spindle detection, individual spindle events were also categorized as true

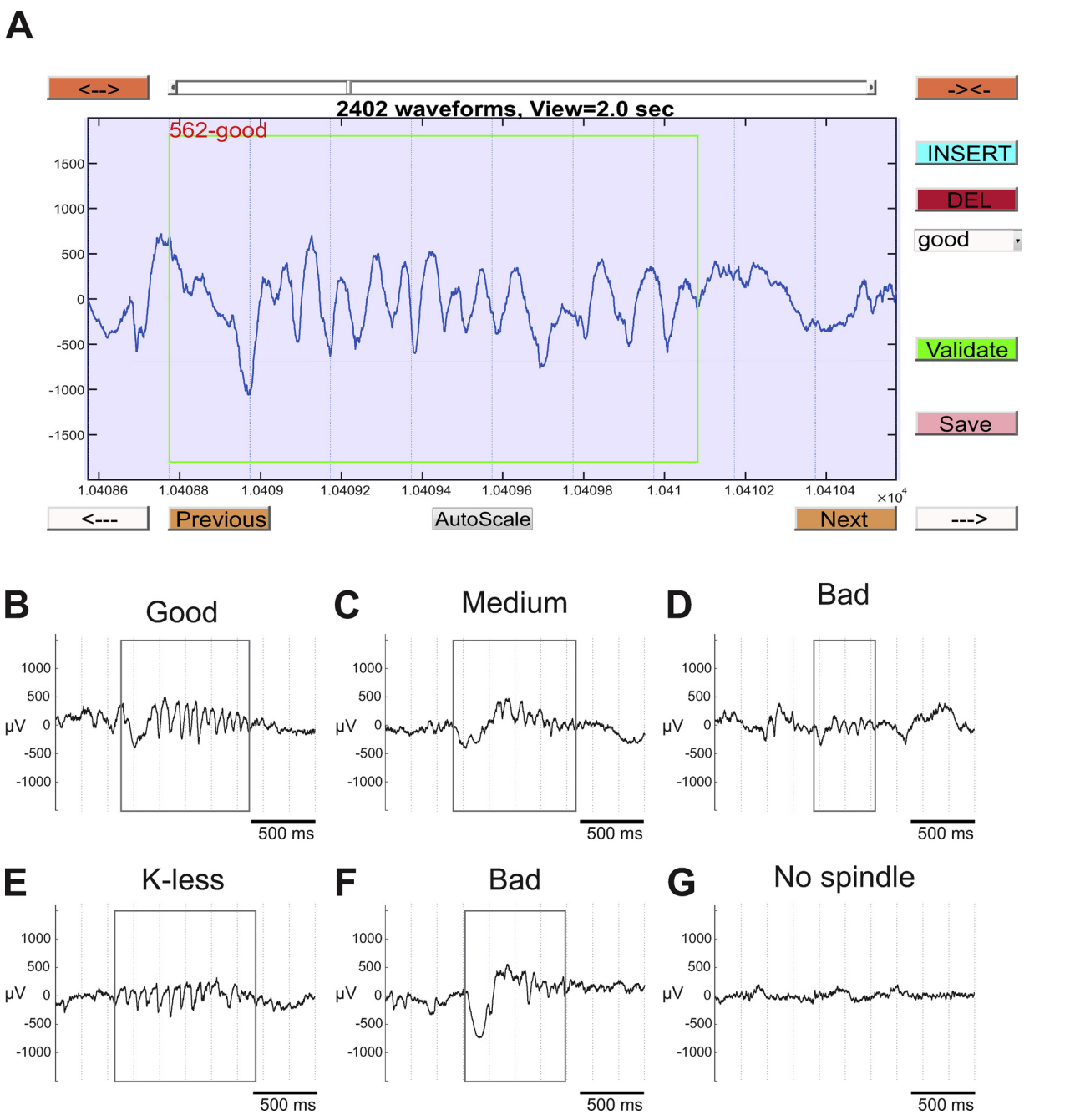

Fig. 2. Methods for spindle assessment by human raters. A: Graphic user interface used for spindle scoring. Vertical dotted lines mark 200 ms increments. The drop-down menu on the right allows spindle quality/category rating. B: Spindle with well-defined K-complex, frequency, and amplitude. C: Spindle with imperfect K-complex and amplitude. D: Brief, low-amplitude spindle. E: Spindle without visible K-complex. F: Irregular spindle with K-complex. G: Portion of sleep without spindles. All images show a 2000 ms window representative of the view used for spindle scoring. Dotted lines represent 200 ms intervals.

positives, false positives, and false negatives. True positives were detected spindles with at least one 10 ms bin of overlap with the ground truth. False positives were detected spindles that did not meet this criterion. False negatives were spindles that existed in the ground truth but had no overlap with any detected spindles. These categories made it possible to assess the relative contribution of soft failures and hard failures to F1 scores. Soft failures, or failures contiguous with ground truth spindles (jitter in true spindle start/end timestamps) were distinguished from hard failures, which are discrete events (incorrect spindles). Although soft failures and hard failures are identified using a criterion of event overlap, both are evaluated for their contributions to overall false positives and negatives in terms of 10 ms bins.

3. Data sharing

Whole datasets (raw EEG voltage recordings) including the tuning dataset and some testing datasets, corresponding rater scoring, and automatic detection results will be posted on the laboratory website and the CRCNS data sharing repository (CRCNS.org). The graphic user interface and training manual used for spindle scoring, as well as custom-written code used for figures and results, will be made freely available from the laboratory website (amygdala.psychdept.arizona.edu/lab.html).

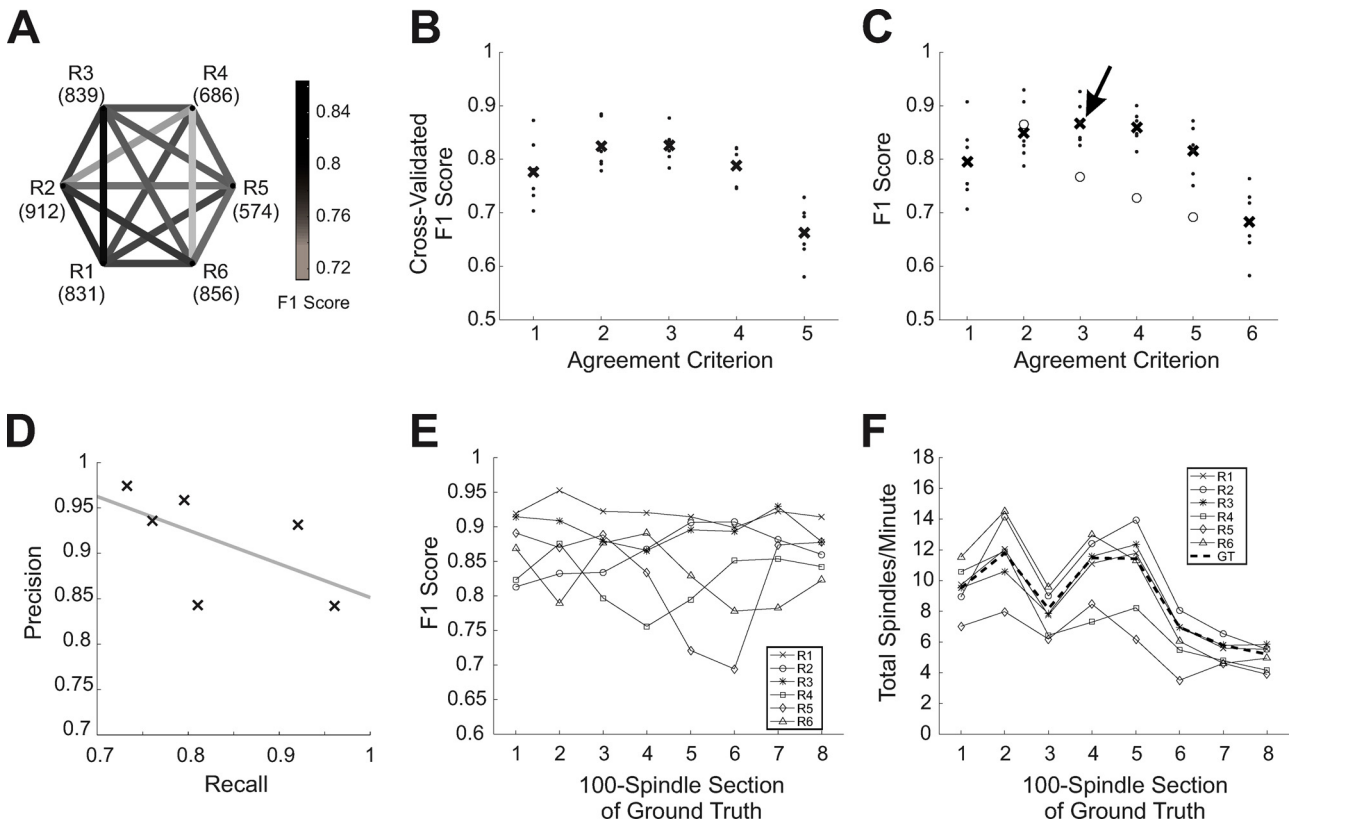

Fig. 3. Inter-rater variability and ground truth construction. A: Pairwise agreement between 6 human raters (R1-6). Each rater's detected spindle count is listed in parentheses. B: F1 score relative to ground truths constructed independently of comparison rater (Xs represent means and dots represent individual rater scores at a given agreement criterion). C: F1 score relative to ground truth. Circles show representative F1 scores for ground truths constructed using a number of raters equal to the agreement criterion. D: Relationship between recall (x axis) and precision (y axis) for each rater relative to ground truth. Line shows least-squares fit ($R2 = 0.341$). E: Individual rater F1 scores relative to ground truth. F: Spindle density computed for each rater and for ground truth (dashed line).

4. Results

4.1. Construction of a ground truth

Spindle scoring from six raters for one 150-minute recording formed the basis for our multi-rater scoring tests. Agreement across raters was compared at 10 ms resolution using the F1 score. Pairwise rater agreement was strong overall, with a mean F1 score of 0.78 ± 0.04 (Fig. 3A). To optimize the number of raters contributing to ground truth, we compared the mean F1 score of each rater to ground truths constructed using different thresholds of rater agreement. We refer to this threshold as the agreement criterion, which varies from 1 through 6 for ground truths constructed using all six raters. Previous publications have noted the limitations of ground truth construction from human raters, especially in the context of scoring from only two raters (O'Reilly and Nielsen, 2015; Tsanas and Clifford, 2015; Lachner-Piza et al., 2018). In a two-rater condition, only two ground truth construction rules are possible: combining all marks from both raters regardless of inter-rater agreement, or including only marks that both raters agreed on. These correspond to an agreement criterion of 1 and an agreement criterion of 6, respectively in Fig. 3B and C. Agreement criteria between these values implement a combination rule in which a minimum number of raters must agree, but these raters can be different for each spindle (e.g., one bin identified by an agreement criterion of two out of six could be validated by raters 1 and 2, while another is validated by raters 1, 4, and 6).

To evaluate how different construction rules influenced ground truth reliability with respect to individual raters, we first compared each rater to different levels of agreement between the other five raters using a leave-one-out cross-validation method. The highest mean F1 score occurred at an agreement criterion of three raters out of six (means shown using Xs, Fig. 3B). We then compared each rater to a ground truth using scores including their own, which produced a similar result (Xs, Fig. 3C).

In both cases, the range in rater F1 scores narrowed at moderate agreement criteria relative to the most relaxed (one-rater) and strictest (all-rater) agreements. This inverted U-shaped curve suggests that, past a certain level of agreement, further increases to the agreement criterion are detrimental, giving individual raters who detected fewer spindles disproportionate influence over the ground truth. Requiring that all 6 raters agree gave the lowest score of all combinations. In fact, setting the threshold equal to the number of contributing raters ('X raters out of X') led to worse individual F1 scores relative to those ground truths as the number of contributors increased (circles, Fig. 3C). On this basis, we chose agreement between any three out of six raters as our criterion for inclusion in the ground truth (arrow).

The F1 scores of individual raters relative to the ground truth had a mean of 0.87 ± 0.04 for the full recording. Each F1 score can be decomposed into its component scores of recall (1-False Negatives) and precision (1-False Positives) (Fig. 3D). Near-perfect F1 scores would have points clustering close to the maximum of 1 on both axes. However, false positives and negatives may occur at different rates and can have unequal influence on the F1 score. Three raters were noticeably biased toward false negatives relative to the ground truth, whereas only one was biased toward false positives. Interestingly, the relationship between recall and precision for our sample of six raters was quasi-linear ($R^2 = 0.341$, $p = 0.223$).

We next asked whether raters are consistent within the recording, from its start to its end. The ground truth contained 827 discrete spindle events, and the recording was divided into non-overlapping sections of

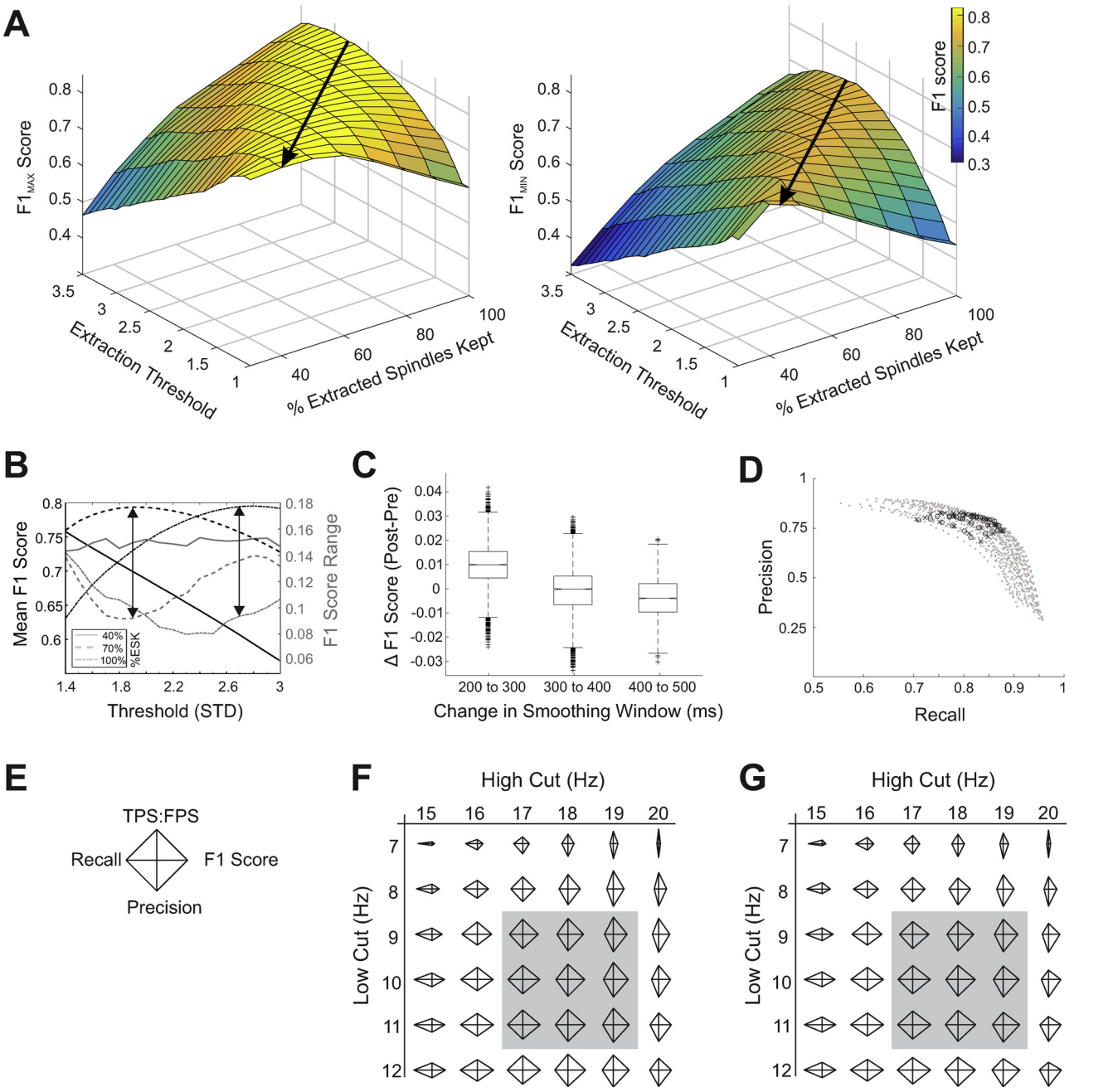

Fig. 4. Impact of parameter choice on automatic detection. A: Interaction between effect of extraction and rejection thresholds on maximum (left) and minimum (right) F1 scores. B: Mean and range of F1 scores as extraction and rejection thresholds vary. Arrows indicate parameter sets of interest. ESK = extracted spindles kept. C: F1 score range at different smoothing factor values. D: Relationship between recall and precision for a smoothing window of 300 ms (Xs: left arrow in 4B; Os: right arrow in 4B). E: Legend for panels F and G. TPS: true positive whole spindles; FPS: false positive whole spindles. F: Effect of bandpass filter on lower-threshold, higher-rejection condition (4B, left arrow). G: Effect of bandpass filter on higher-threshold, zero-rejection conditions (4B, right arrow).

100 spindles for higher F1 score resolution. Performance varied between raters, with a tendency for raters with lower overall F1 scores to demonstrate greater variability (Fig. 3E). No trends in F1 score were apparent across raters for any sections of the recording, which suggests that fluctuations were due to raters rather than changes in the recording quality. Most raters also detected similar spindle density relative to the ground truth throughout the recording (Fig. 3F).

4.2. Parameter optimization

Using the ground truth from this recording, we conducted a parameter sensitivity study testing the automatic detection algorithm on a range of bandpass filters, smoothing windows, extraction thresholds, and rejection criteria (details in Methods). Fig. 4A shows the interaction between extraction threshold and rejection criterion using maximum (left) and minimum (right) F1 scores from all bandpass filters and smoothing window combinations. The asymmetry in the sharp drop off from the central ridges (arrows) showed that the detriment of removing spindles above a high threshold (> 2.5 STD) was greater than the benefit of removing a percentage of spindles above a low threshold. (< 1.5 STD). The ridge in the minimum curve was narrower than the one in the maximum curve (arrows), indicating the need to monitor the range in F1 scores for a given extraction threshold and rejection criterion.

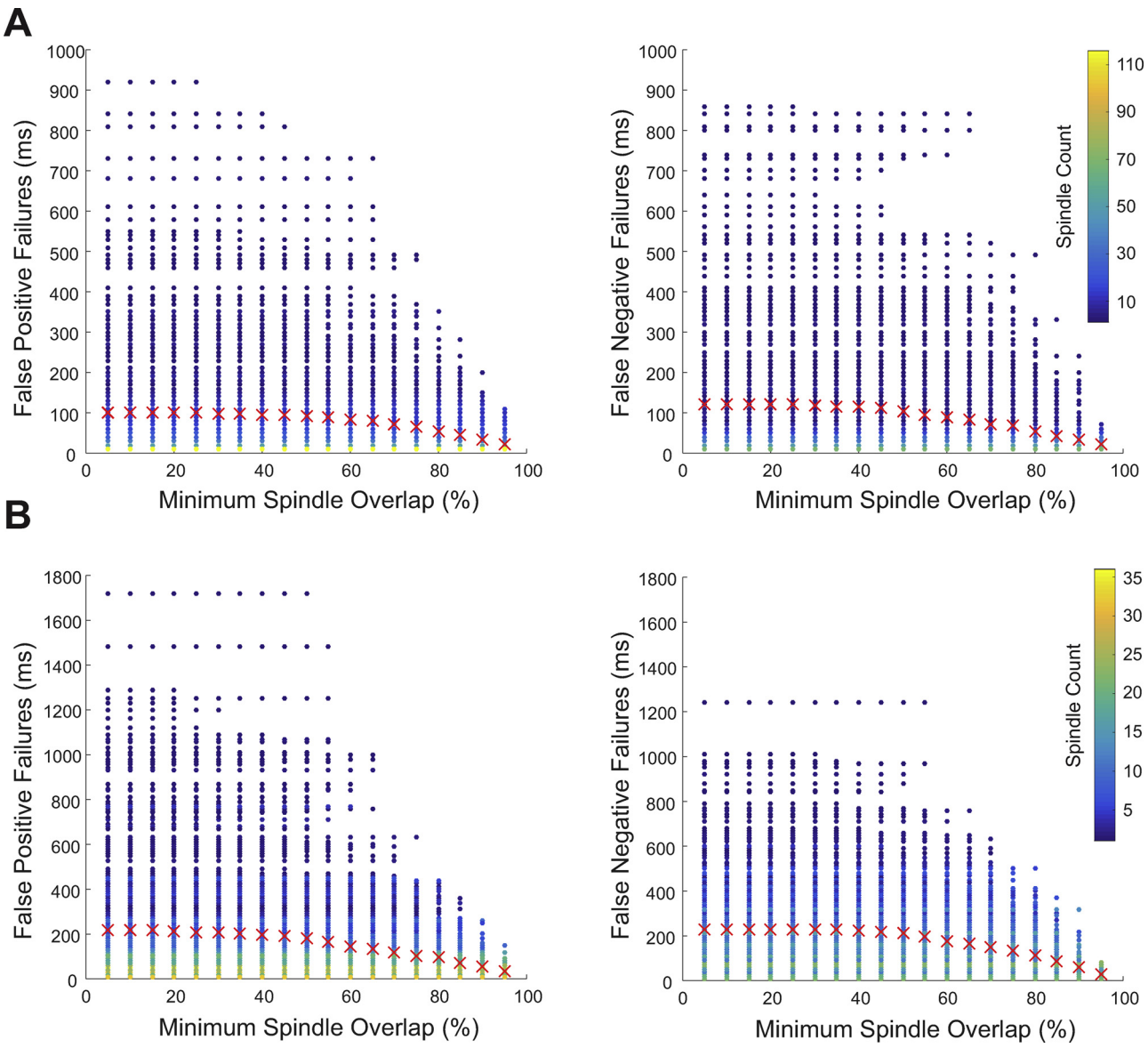

Fig. 5. Soft failures of the algorithm with respect to human raters failures. A: Incidence of false positive (left) and false negative (right) failures of the chosen parameter set relative to ground truth. B: Incidence of false positive (left) and false negative (right) failures of a representative human rater relative to ground truth. Counts are represented by the colors of the dots. Red Xs denote the mean for each overlap (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

This relationship can be optimized using cross-sections of the data. Ideally, a parameter set will have a high mean F1 score and a small range in F1 scores across all variations in bandpass filter and smoothing window. We compared these quantities for different extraction thresholds at a given rejection criterion (Fig. 4B). This method identified two parameter combinations of interest: one detecting the strongest 70% of spindles at least 1.9 standard deviations above baseline (left double-arrow), and another detecting all spindles at least 2.7 standard deviations above baseline (right double arrow).

We also studied the effect of the Hilbert transform smoothing window on the F1 score. A window of 300 ms had a significantly greater effect on F1 score than any other tested value as determined by one-way ANOVA ($F(2,22461) = 5076.77$, $p < 0.001$, Fig. 4C). It should be noted, however, that this effect only resulted in small changes in absolute F1 score, suggesting that the smoothing window has a relatively minor influence on spindle detection outcomes when compared to the other parameters studied here.

Observation of all parameter combinations with a 300 ms smoothing window showed a boomerang-shaped relationship between false positives and false negatives (Fig. 4D). This curve suggests a compromise between the two types of failures, with optimal parameter sets found in the area of curvature in the distribution (upper right corner of the graph). Indeed, most parameter sets using the extraction thresholds and rejection criteria of interest identified in Fig. 4B are located in this region (crosses and circles in Fig. 4D). Interestingly, there is a larger range in false positives (precision on y-axis, min: 0.2885, max: 0.8981) than false negatives (recall on x-axis, min: 0.5545, max: 0.9569) across all parameter sets, but the opposite is true within the parameter sets of interest (crosses and circles, Fig. 4D). This highlights the importance of conducting a large-scale parameter sweep in the optimization procedure, since local variations on an individual parameter setting may not yield a result space that reflects these relationships.

The two extraction threshold and rejection criterion combinations of interest show similar results across variations in bandpass filter frequency limits (Fig. 4E–G). We plot the ratio of true positive to false

positive spindle events (TPS:FPS), Recall, Precision and F1 score simultaneously using a lozenge diagram (Fig. 4E). In this display, larger values are further from the center and lozenges with larger areas reflect better bandpass filters. The two parameter sets identified with the arrows in Fig. 4B produce maximal results in the same range of bandpass filter boundaries (shaded area, Fig. 4F–G). To select a parameter set for closer examination, we first chose to proceed with only the higher-threshold, no-rejection combination (Fig. 4B, right arrow) because it required one fewer parameter at no additional cost to the results (Fig. 4G). We then chose the bandpass filter with the highest F1 score at that threshold, ensuring that the difference between recall and precision did not exceed 0.1 in order to prevent bias toward either false positive or false negative errors. This parameter set had recall of 0.8446 and precision of 0.7965, producing an F1 score of 0.8226 and a ratio of 13 true positive spindles to each false positive spindle (7.7% false positive event rate). Notably, these results were comparable to those produced by parameter sets tuned using half of the ground truth spindle events and testing on the other half (tuning with first half/testing on second half: F1 score 0.7983; tuning with second half/testing on first half: F1 score 0.7986). In summary, tuning using the full 150-minute recording identified a parameter set with an 11–17 Hz bandpass filter, 2.7 standard deviation extraction threshold, smoothing window of 300 ms, and no rejection parameter.

4.3. Sources of rater disagreement

The chosen parameter set detected 760 spindles occupying 795 s of the recording (mean spindle duration 1.05 ± 0.57 s). False positives and negatives at 10 ms resolution were divided into two categories. Soft failures represented disagreement in the start and/or end times of true spindles, and hard failures occurred as part of spindles that were detected by either the algorithm/human or ground truth but not both.

Fig. 5 shows the distribution of soft false negative spindles (individual points) for different amounts of minimum spindle overlap between the spindles detected by the human (Fig. 5A) and algorithm (Fig. 5B) and the ground truth. Spindles with no failures for the tested category were excluded from plots and statistics. For example, the error time of each spindle overlapping the ground truth by at least 20% appears in the fourth column of dots on each panel, with the dot color representing the number of spindles at a given level of false positives. Red crosses represent the average number of false positives at a given overlap level.

The distributions of soft false positive and false negative failures in human raters were skewed relative to the average, with most cases involving disagreement of less than 100 ms (median disagreement of 70 ms across all raters at minimum 5% spindle overlap; data from Rater 1 shown in Fig. 5A). False positives were more numerous overall (133.14 ± 73.85 s) than false negatives (70.55 ± 54.95 s). On average across all raters, false positives consisted of 49.85 ± 40.39 s of soft failure and 20.33 ± 17.05 s of hard failure. False negatives consisted of 64.59 ± 55.82 s of contiguous failure and 68.45 ± 54.16 s of discrete failure. As evidenced by the large standard deviations, individual raters did not exhibit any particular category of failure.

For the automatic detection algorithm, false positives consisted of 110.56 s of soft failures and 35.10 s of hard failures and false negatives included 122.67 s of soft failures and 37.36 s of hard failures. Strong skew was also present in the distributions of soft failures in the automatic detection algorithm, as can be seen by the asymmetry around the averages at all overlap levels (Fig. 5B). Fewer spindles had soft false positives (511 spindles, 67%) than soft false negatives (532 spindles, 70%). Soft false positives occupied 110.56 s, whereas soft false negatives occupied 122.67 s. This suggests that the chosen parameters optimized the balance between positive and negative failures.

Fig. 6 shows an analysis of the hard failures of the automatic detection algorithm. We used Welch's method to estimate the power spectral density over the time range of a given spindle identified by our detector, assigning a dominant frequency to the spindle. Hard failures did not differ in frequency from true spindle events (Fig. 6A). Both false positive and false negative hard failures had low spindle power when compared to true positives (Fig. 6B), suggesting that threshold adjustments (see methods) would likely not be effective in rescuing these failures. Most hard failures were of relatively short duration (less than 1 s, 6C middle and right panels). False negatives (Fig. 6C, right) showed a lower minimum duration than true positives or false positives, consistent with the 300 ms minimum duration for automatic detection; raters were trained to score events featuring a minimum of 3 cycles rather than a minimum duration, allowing them to detect events briefer than 200 ms that the automatic detection would not extract.

In sum, hard failures of the automatic detection algorithm accounted for 70.46 s (52 false positive spindles and 71 false negative spindles) of failure relative to ground truth, and were characterized by short durations and low power. Rescuing them could involve the use of complementary algorithms with additional detection criteria.

4.4. Applicability to other datasets

To test whether optimization using a single recording could generalize across sessions and rats, we compared automatic detections using the parameter set chosen above to additional 150-minute recordings scored by a single rater (Rater 1 in Fig. 3A, E, F). These recordings were not used for parameter tuning and were introduced exclusively to test the tuned parameter set. In total, the rat from the session used for parameter tuning (Figs. 3–5) contributed three recordings and two other rats contributed two recordings each. Five recordings were taken in anterior cingulate cortex and two were obtained from prelimbic cortex.

The F1 scores for 100-spindle sections of the rater's scoring were comparable across recordings, although the results favored the original recording used to tune the parameters (Fig. 7A, 'Tuning session'). In our sample, differences between sessions did not appear to vary systematically by recording depth or rat, suggesting that the automatic detection algorithm was not sensitive to particular recording depths, or spindle properties between rats. Further work with a larger number of sessions at different depths and from different rats should be collected to confirm this finding.

One recording session exhibited a poor overall F1 score (arrow, Fig. 7A). We plotted the standard deviation of the amplitude of the Hilbert transform over the course of the entire recording ('Full recording') and during detected spindle times ('Spindle Only') for each of the sessions used. All 8 recordings had comparable ratios of transform variability during spindles to overall variability across the recording (1.45 ± 0.04 to 1). In our samples, the variability of the transform during detected spindles was always higher than when computed using the entire recording ($H(2) = 1.8$, $p = 0.18$, Fig. 7B). Interestingly, the recording that had the lowest F1 score was also the one with the lowest Hilbert transform variability (lowest line, arrow, Fig. 7B). In this case, the standard parameter set chosen may not have been optimal for this session. Additional data and procedures may be required to find a parameter set (e.g. threshold) adapted to the overall properties of recordings with low signal variability.

5. Discussion

Many studies of cortical activity during sleep focus on spindles at coarse resolution, estimating spindle density based on changes in power spectra. Relatively few studies test the precise timing of spindle generation because of the difficulty to extract them automatically in a reliable manner. Our results outline a general method to validate automatic individual spindle detection algorithms with a moderate number of human raters, tune algorithm parameters, and generalize these parameters to other recordings with limited or no rater validation.

We show that a reliable ground truth may be constructed based on

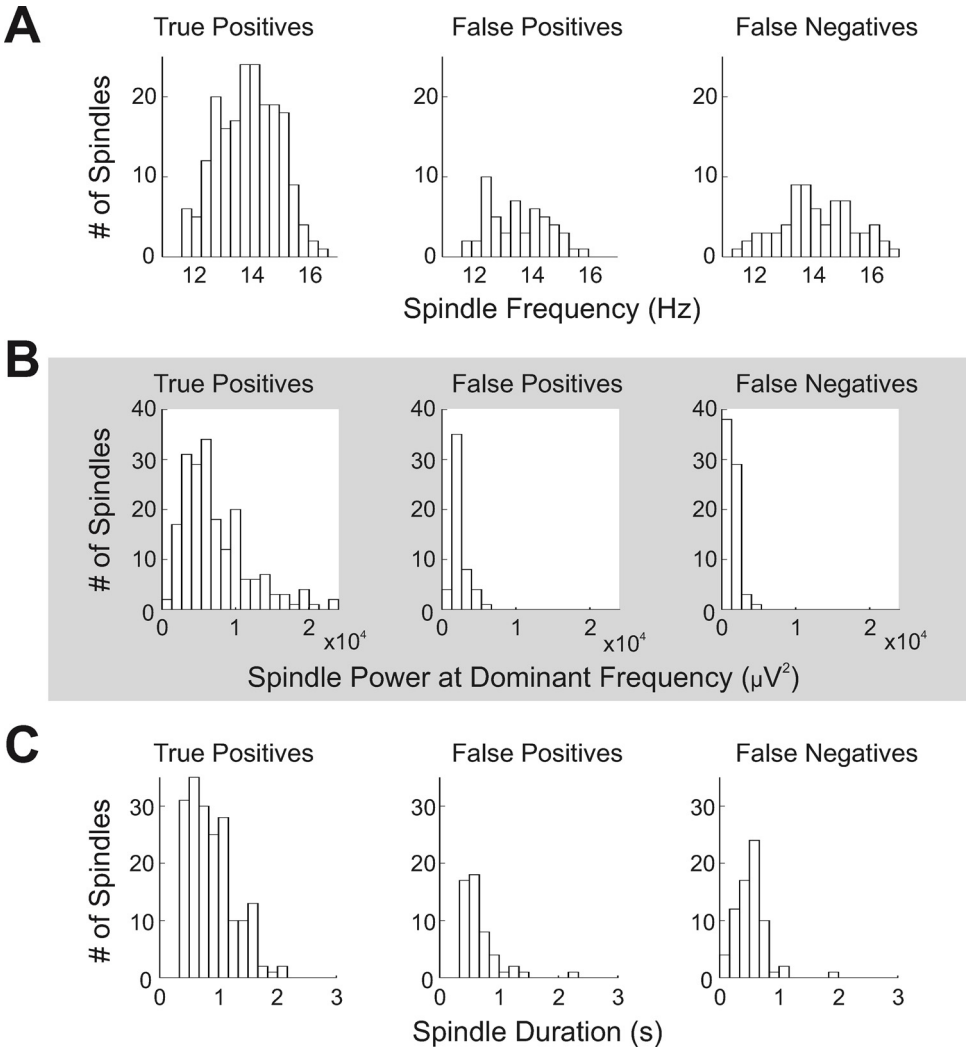

Fig. 6. Hard failures of the automatic detection algorithm. A: Frequency of true positive (left), false positive (center), and false negative (right) spindles. B: Power of spindles at dominant frequency in each category from A. C: Duration of spindles in each category from A.

an agreement of three human raters out of six. These raters do not need to be experts at spindle extraction and need only basic training. Computing the ground truth on the basis of too few or too many raters, or requiring agreement between all raters in a pool of any size, may have a deleterious effect on its assessment as measured by the F1 score. This problem can be avoided by comparing rater agreement with the ground truth and choosing the agreement criterion with the highest and most consistent individual rater agreement, similar to Warby et al. (2014). The agreement criterion should be optimized for different rater pools in order to adjust to higher or lower rates of rater error. We show

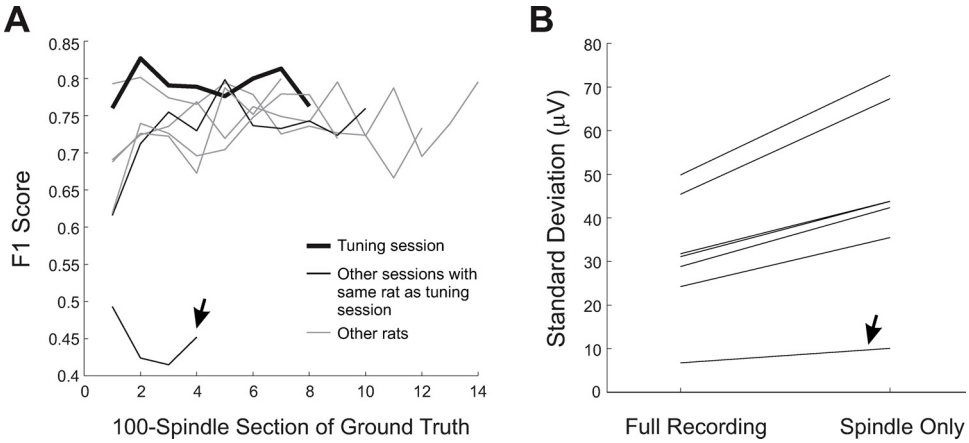

Fig. 7. Generalization of the chosen parameter set to other recordings. A: Automatic detection F1 scores per section of 100 single-rater ground truth spindles in recordings from different days and rats. B: Hilbert transform variability in the same sample of recordings.

that under these conditions, individual raters had a high and temporally consistent F1 score when compared to ground truth. Notably, our average inter-rater F1 score matched the highest inter-rater reliability reported in a study of human spindle detection by expert sleep technicians (Wendt et al., 2015).

We also showed that spindle density could be well approximated by the individual scoring of most raters, as evidenced by the similarity in multiple raters' detections of spindle density over the course of the recording. Our ground truth based on three out of six raters produced an effective estimate that captured the same temporal dynamics of spindle density as human raters. Taking together both rater and ground truth estimates of spindle density, and the high F1 scores comparing human raters and the ground truth, it appears that the ground truth captures the timing of individual spindles as well. These results suggest that a ground truth based on three out of six raters with basic training can both successfully estimate spindle density and the precise timing of individual spindle occurrence.

While it is always possible to construct a ground truth by aggregating across spindles marked by individual raters, one limitation of this approach is the possibility that minor but consistent variations across raters may contaminate the estimate. For example, during periods when spindles occur in rapid sequences, there may be overlap in the boundaries of spindles marked by different raters. Consequently, the ground truth identifies times that should generally be detected as spindle-rich— but not the precise timing of the start or end of each spindle. This complicates the meaning of any comparison metric, including the F1 score: meaningful, yet brief, gaps between spindles detected by individual human raters or an algorithm could count as deviations from perfect agreement. This affects ground truths constructed even in circumstances of high inter-rater reliability. In cases of low agreement between human raters, it would add to the existing problem of rater influence on F1 score (O'Reilly and Nielsen, 2015). Attempting to find algorithm parameters that achieve a higher F1 score than individual humans could therefore compromise the advantage of methods that detect spindle boundaries when compared to spectral methods that reliably capture coarser spindle-rich epochs. However, these epochs can be identified by measuring inter-spindle intervals in detections performed by an algorithm or by individual raters, and can be excluded from analyses or parameter tuning if so desired.

Using the recording used for tuning, the performance of the automatic detection algorithm was similar to that obtained by human rater agreement. The ideal case would be to achieve near-perfect recall, minimizing false negatives, before affecting precision with higher rates of false positives. The balance observed using the parameter sweep in this study occurs around values of 0.8 for both quantities. This produced a dense region of F1 scores just within two standard deviations below the mean F1 score of our human raters relative to the ground truth.

The inaccuracies of our automatic detection consisted mostly of soft failures. These failures were related to the precise start and end of a spindle rather than to whether they occurred at all. Spindles that were missed or that were incorrectly detected (hard failures) had relatively low power and were of shorter duration than average. Such features may prompt the design of future algorithms that could focus specifically on transient low-power events.

Because our analysis focused on continuous time series, our F1 scores calculated agreement at a resolution of 10 ms. This differed from by-event strategies, which classify whole spindles as true positives using an overlap threshold, used for some F1 calculations in studies reporting expert (Warby et al., 2014; Wendt et al., 2015) and non-expert agreement (Zhao et al., 2017). High ratios of true positive spindles to false positive spindles were observed for a wide range of parameter sets with variable F1 scores, suggesting that parameter tuning should emphasize criteria related to F1 score.

The comparison of rater performance with ground truth in 100-spindle sections allows us to understand how raters varied over the course of one recording. This variability can be contrasted to that of a given rater assessing the same session multiple times, which would demonstrate how individual raters learn from experience and whether they converge onto a stable rating. Unlike in our study, where most raters agree with each other consistently in a given recording session, a study of intra-rater variability would require that multiple sessions be used to ensure that the results were not dependent on the specifics of a particular sleep epoch. Raters could spend either a short amount of time between rating sessions, allowing for evaluation of rater reliability within similar circumstances, or a longer amount of time to assess reliability in the long term. The latter condition has been partially addressed in previous spindle detection literature in humans (Wendt et al., 2015), but not in rodents. We left additional study of intra-rater reliability for further work.

We were surprised to find that the optimal extraction threshold identified during our parameter tuning was higher than the thresholds used in some studies (Eschenko et al., 2006; Mölle et al., 2011; Sullivan et al., 2014). To our knowledge, the manner in which these thresholds were set was *ad hoc*, with little or no justification provided. This suggests that previous spindle detection methods using low thresholds may have been biased towards the detection of more spindles than were actually occurring, and that they therefore included a large number of false positives, the consequence of which is difficult to estimate.

At the same time, actual differences between recordings may warrant different parameters. Without an established procedure to optimize parameters for spindle detection, the amount of detection error may vary across studies and remains unquantified. The method introduced here accounts for the need for flexibility between studies and provides rationale for the chosen parameters, allowing the adoption of a common framework for parameter selection. It is compatible with automatic detections using the algorithm discussed here and variations on it. To support the evaluation of additional detection approaches as well as the parameters of a given detection, the data will be freely available for comparison of detectors specifically in the context of rodent recordings.

As noted above, automatic detection with a tuned parameter set may show weaker correspondence to a human ground truth in recordings that were not used for optimization. Performance was slightly lower in the sample of additional recordings tested, but remained comparable for most recordings. Simplifications of the procedure may be appropriate in some contexts. To reduce computation time, it is possible to run two reduced parameter sweeps while maintaining sufficient coverage of the parameter space: first, a coarse sweep with large intervals (e.g., threshold values varied in increments of 0.5 instead of 0.1), and then a fine sweep between the two parameter values with the best outcomes. This is recommended only for algorithms that do not exhibit local minima/maxima in F1 score. To reduce the amount of human scoring required for parameter tuning or post-hoc verification of a chosen parameter set, it is also possible to limit the ground truth sample to the first 200–300 rater-detected spindles in a recording. However, it is important to note that tuning using this approach must still be conducted independently of data used for analyses.

Conflict of interest statement

The authors have no conflict of interest to declare.

Funding

Support was provided by ONR MURIN000141310672, N000141612829, and N000141512838 (JMF)

Acknowledgment

The authors thank Dr. Marco Contreras, Dr. Bruce Harland, Erin Howard, and Raven Padgett for contributing to spindle scoring.

References

Buzsáki, G., 2006. Rhythms of the Brain. Oxford University Press, New York; Oxford.

Clemens, Z., Molle, M., Eross, L., Barsi, P., Halasz, P., Born, J., 2007. Temporal coupling of parahippocampal ripples, sleep spindles and slow oscillations in humans. Brain 130, 2868–2878.

Contreras, M., Pelc, T., Llofriu, M., Weitzenfeld, A., Fellous, J., 2018. The ventral hippocampus is involved in multi-goal obstacle-rich spatial navigation. Hippocampus 1–14.

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

Eschenko, O., Molle, M., Born, J., Sara, S.J., 2006. Elevated sleep spindle density after learning or after retrieval in rats. J. Neurosci. 26, 12914–12920.

Fogel, S.M., Smith, C.T., 2006. Learning-dependent changes in sleep spindles and Stage 2 sleep. J. Sleep Res. 15, 250–255.

Gais, S., Molle, M., Helms, K., Born, J., 2002. Learning-dependent increases in sleep spindle density. J. Neurosci. 22, 6830–6834.

Gothard, K.M., Skaggs, W.E., Moore, K.M., McNaughton, B.L., 1996. Binding of hippocampal CA1 neural activity to multiple reference frames in a landmark-based navigation task. J. Neurosci. 16, 823–835.

Harper, B., Sampson, A., Sejnowski, T.J., Fellous, J.-M., 2016. Sleep spindles and single-cell reactivation in the rodent medial prefrontal cortex during context-dependent memory reconsolidation. Annual Meeting of the Society for Neuroscience.

Helfrich, R.F., Mander, B.A., Jagust, W.J., Knight, R.T., Walker, M.P., 2018. Old brains come uncoupled in sleep: slow wave-spindle synchrony, brain atrophy, and forgetting. Neuron 97 (221-230), e224.

Iber, C., Ancoli-Israel, S., Chesson, A.L., Quan, S., 2007. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.

Jiang, X., Shamie, I., W KD, Friedman, D., Dugan, P., Devinsky, O., Eskandar, E., Cash, S.S., Thesen, T., Halgren, E., 2017. Replay of large-scale spatio-temporal patterns from waking during subsequent NREM sleep in human cortex. Sci. Rep. 7, 17380.

Lachner-Piza, D., Epitashvili, N., Schulze-Bonhage, A., Stieglitz, T., Jacobs, J., Dümpelmann, M., 2018. A single channel sleep-spindle detector based on multivariate classification of EEG epochs: MUSSDET. J. Neurosci. Methods 297, 31–43.

Lajnef, T., O'Reilly, C., Combrisson, E., Chaibi, S., Eichenlaub, J.B., Ruby, P.M., Aguera, P.E., Samet, M., Kachouri, A., Frenette, S., Carrier, J., Jerbi, K., 2017. Meet spinky: an open-source spindle and K-complex detection toolbox validated on the open-access montreal archive of sleep studies (MASS). Front. Neuroinform. 11, 15.

Latchoumane, C.V., Ngo, H.V., Born, J., Shin, H.S., 2017. Thalamic spindles promote memory formation during sleep through triple phase-locking of cortical, thalamic, and hippocampal rhythms. Neuron 95 (424-435), e426.

Luthi, A., 2014. Sleep spindles: where they come from, what they do. Neuroscientist 20, 243–256.

Mander, B.A., Rao, V., Lu, B., Saletin, J.M., Ancoli-Israel, S., Jagust, W.J., Walker, M.P., 2014. Impaired prefrontal sleep spindle regulation of hippocampal-dependent learning in older adults. Cerebral Cortex (New York, NY: 1991) 24, 3301–3309.

Molle, 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, 10941–10947.

Mölle, M., Bergmann, T.O., Marshall, L., Born, J., 2011. Fast and slow spindles during the sleep slow oscillation: disparate coalescence and engagement in memory processing. Sleep 34, 1411–1421.

Muller, L., Piantoni, G., Koller, D., Cash, S.S., Halgren, E., Sejnowski, T.J., 2016. Rotating waves during human sleep spindles organize global patterns of activity that repeat precisely through the night. eLife 5.

Niknazar, M., Krishnan, G.P., Bazhenov, M., Mednick, S.C., 2015. Coupling of Thalamocortical Sleep Oscillations Are Important for Memory Consolidation in Humans. PLoS One 10, e0144720.

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

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

Siapas, A.G., Wilson, M.A., 1998. Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. Neuron 21, 1123–1128.

Sirota, A., Csicsvari, J., Buhl, D., Buzsaki, G., 2003. Communication between neocortex and hippocampus during sleep in rodents. Proc. Natl. Acad. Sci. U. S. A. 100, 2065–2069.

Sullivan, D., Mizuseki, K., Sorgi, A., Buzsaki, G., 2014. Comparison of sleep spindles and theta oscillations in the hippocampus. J. Neurosci. 34, 662–674.

Tamminen, J., Payne, J.D., Stickgold, R., Wamsley, E.J., Gaskell, M.G., 2010. Sleep spindle activity is associated with the integration of new memories and existing knowledge. J. Neurosci. 30, 14356–14360.

Tsanas, A., Clifford, G.D., 2015. Stage-independent, single lead EEG sleep spindle detection using the continuous wavelet transform and local weighted smoothing. Front. Hum. Neurosci. 9, 181.

Valdes, J.L., McNaughton, B.L., Fellous, J.M., 2015. Offline reactivation of experience-dependent neuronal firing patterns in the rat ventral tegmental area. J. Neurophysiol. 114, 1183–1195.

Wallant, D.C., Maquet, P., Phillips, C., 2016. Sleep spindles as an electrographic element: description and automatic detection methods. Neural Plast. 2016, 6783812.

Warby, S.C., Wendt, S.L., Welinder, P., Munk, E.G., Carrillo, O., Sorensen, H.B., 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, 385–392.

Wendt, S.L., Welinder, P., Sorensen, H.B., Peppard, P.E., Jennum, P., Perona, P., Mignot, E., Warby, S.C., 2015. Inter-expert and intra-expert reliability in sleep spindle scoring. Clin. Neurophysiol. 126, 1548–1556.

Wierzynski, C.M., Lubenov, E.V., Gu, M., Siapas, A.G., 2009. State-dependent spike-timing relationships between hippocampal and prefrontal circuits during sleep. Neuron 61, 587–596.

Zhao, R., Sun, J., Zhang, X., Wu, H., Liu, P., Yang, X., Qin, W., 2017. Sleep spindle detection based on non-experts: a validation study. PLoS One 12, e0177437.