A method for the precise detection and validation of spindle timing in rodents

Blaine Harper[1] and Jean-Marc Fellous[1,2,3]

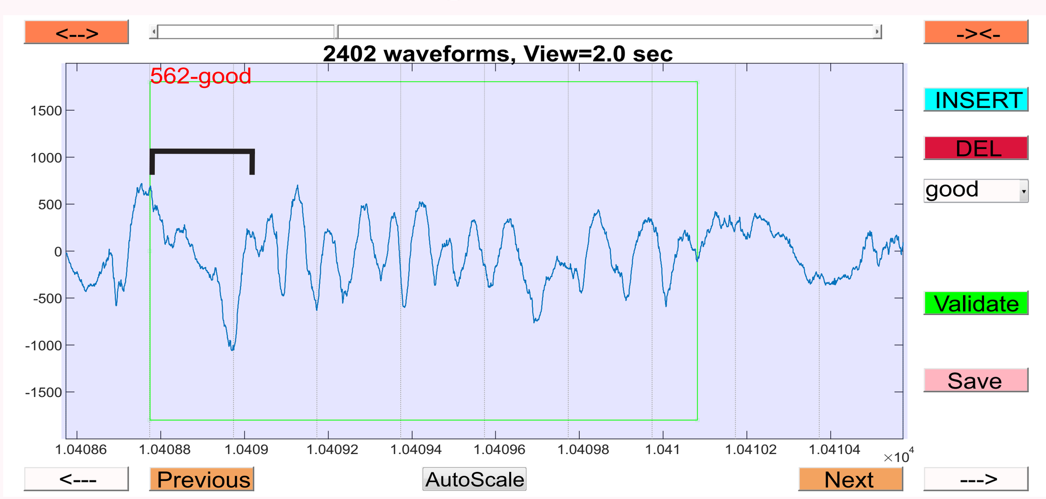
1 Department of Psychology, University of Arizona, 2 Biomedical Engineering Department, University of Arizona, 3 Program in Applied Mathematics, University of Arizona

Introduction

- The detection of spindle boundaries is critical in the field of memory consolidation[1,2].
- The values of algorithm parameters vary between studies using similar algorithms.
- Previous studies have not evaluated spindle detection quality in rodent data[3].
- Our study has two goals:

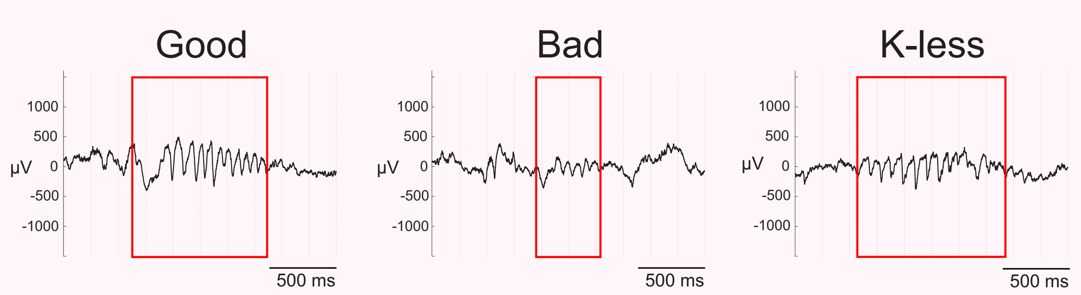
1. Create a database of rodent spindles scored by humans for algorithm assessment.
2. Establish a validation and testing procedure to optimize algorithm parameters.

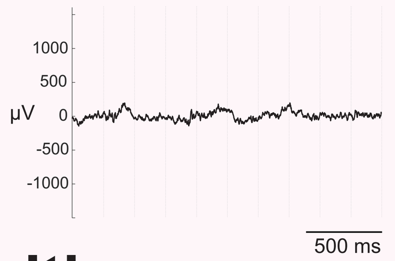
Manual Scoring



Spindle Evaluation Guidelines

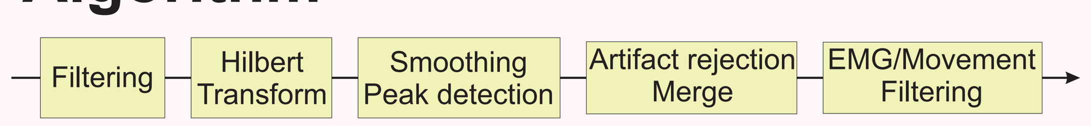
- frequency (~11-15 Hz)
- duration (at least 3 oscillatory cycles)
- amplitude
- presence of initial K-complex (see bracket)





Baseline EEG during sleep

Algorithm



- Validation data: 2.5-hour recording, 6 raters
- Testing data: six 2.5-hour recordings, 1 rater

Parameters Studied

- low cut frequency (7-12 Hz)
- high cut frequency (15-20 Hz)
- smoothing window (200-500 ms)
- peak detection threshold (1-3.5 standard deviations above mean)
- artifact rejection methods

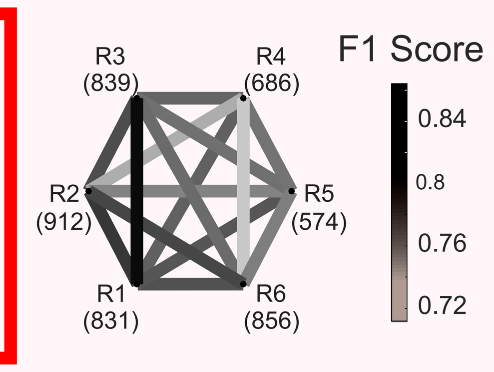
Ground Truth Construction

F1 Score

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

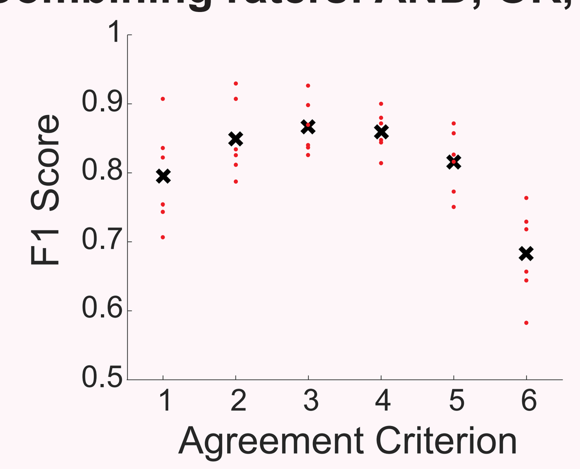
Recall: 1 - False Negatives

Precision: 1 - False Positives



- F1 score is calculated using 10 ms bins.
- Inter-rater agreement is high.
- Number of detected spindles varies across raters.

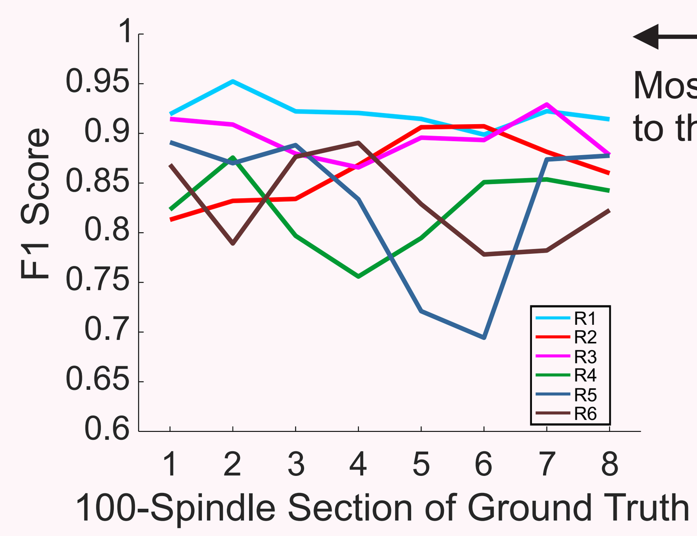
Combining raters: AND, OR, and AND/OR rules



AND rule: "all raters"
OR rule: "any rater"
AND/OR rule: "any X of these raters"

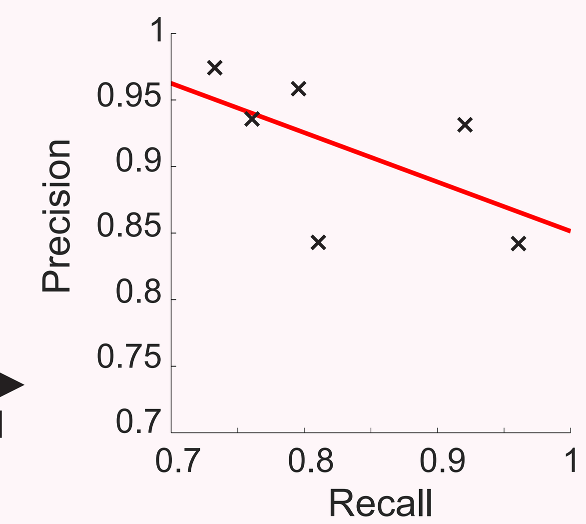
A combination AND/OR rule using three out of six raters performs best in our data.

Rater Performance



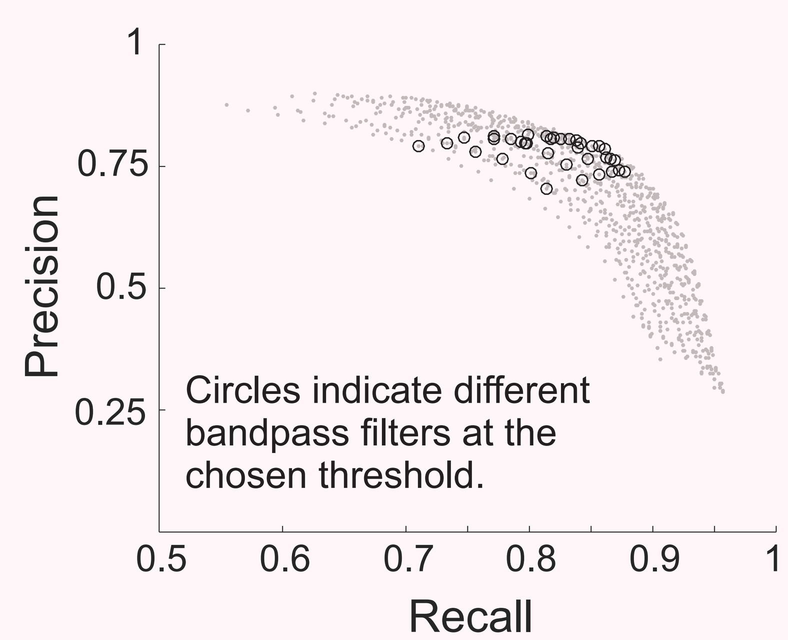
Most raters have consistent performance relative to the ground truth throughout the recording.

Human failures are biased toward false negatives.

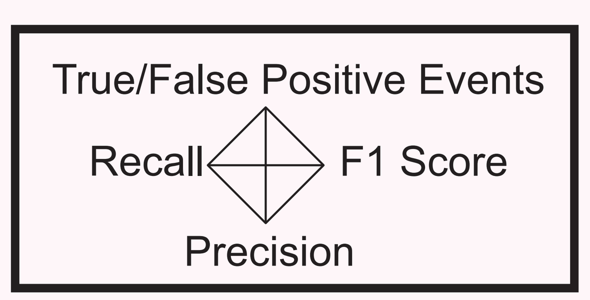


Performance on Validation Data

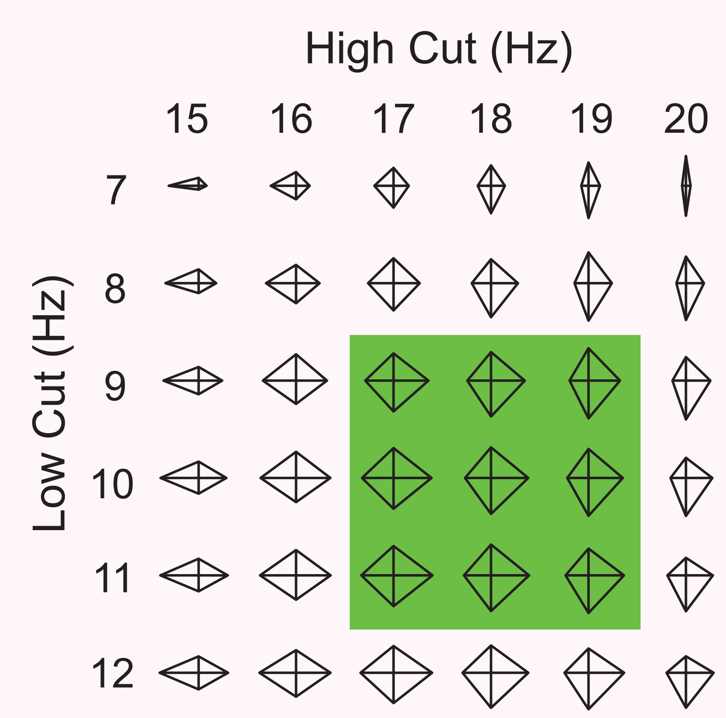
Parameter optimization: maximize $F1_{MAX}$ and minimize $F1_{MAX}$-$F1_{MIN}$



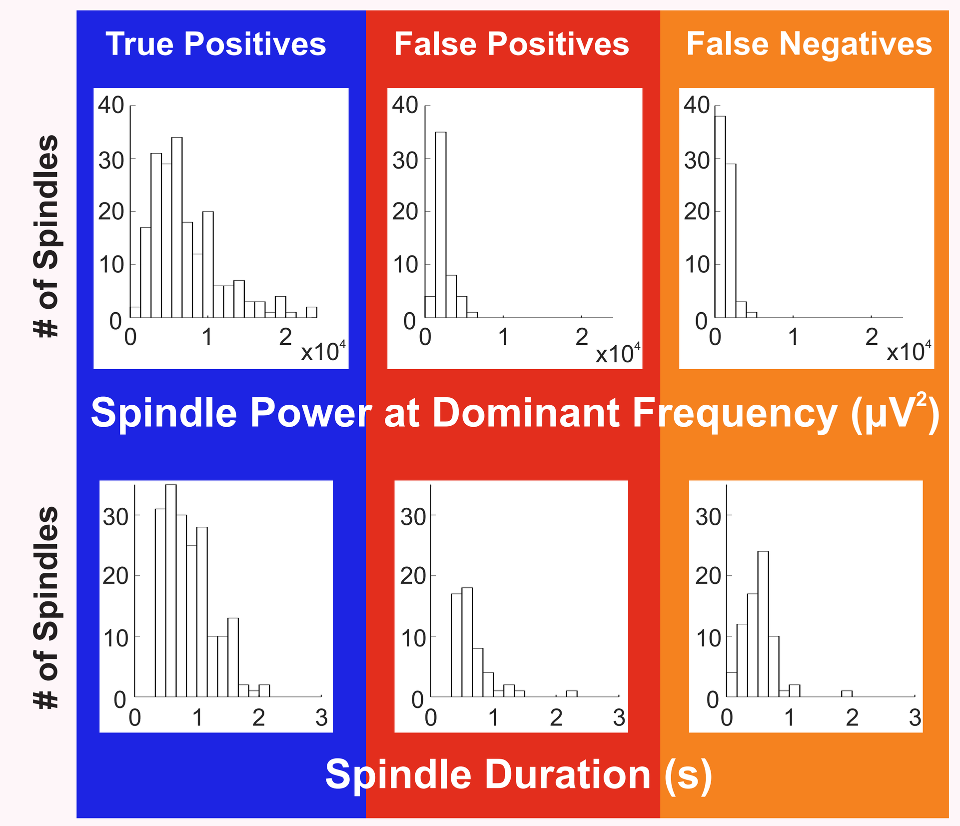
- Smoothing window had little influence on F1 score (4% or less).
- Some automatic extractions are biased toward false negatives, while others are biased toward false positives.



- Extractions with high F1 scores at 10 ms resolution also have low rates of false positive whole spindles.



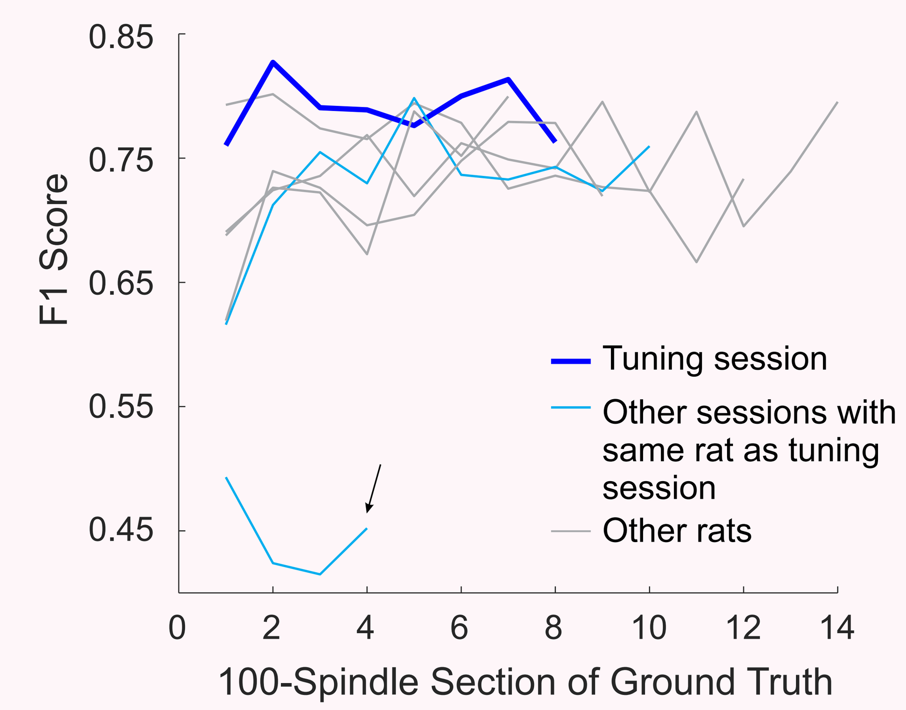
Automatic Detection Failures in Validation Data



- Both false positive and false negative spindles have lower power at their dominant frequency.
- Both false positive and false negative spindles tend to have durations shorter than 1 second.
- Spindle frequencies did not differ across true positives, false positives, and false negatives.

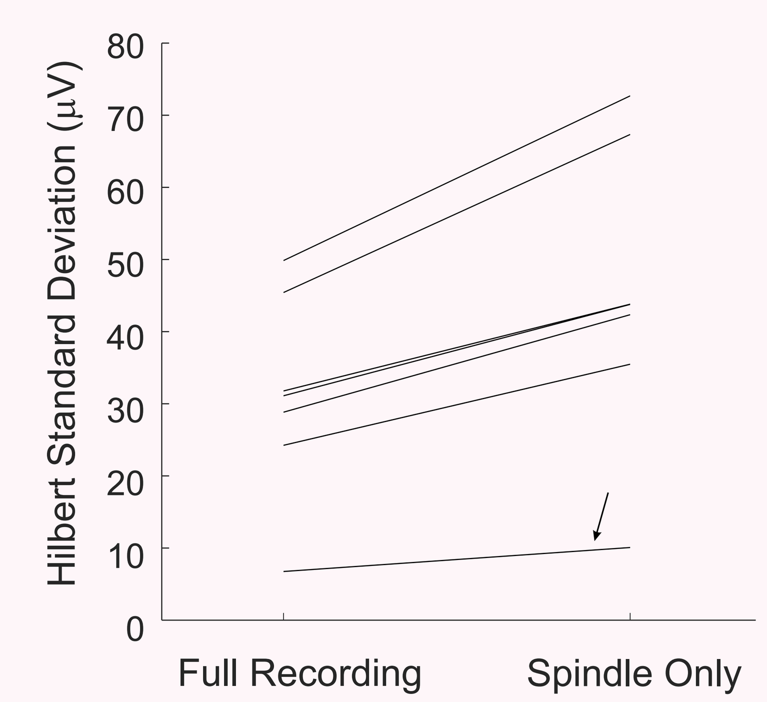
- Most automatically detected spindles have less than 100 ms of overlap failure.
- False positive overlap failures were more common than false negative ones for both the automatic detection and humans relative to ground truth.

Performance on Testing Data

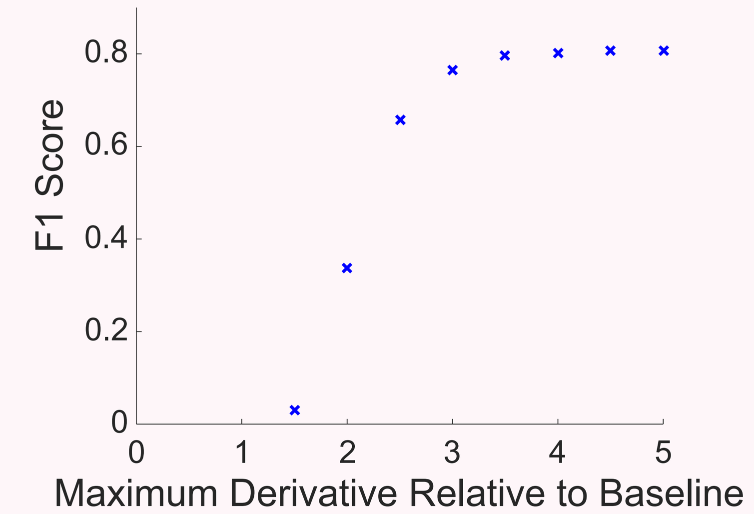


- Overall F1 score and consistency in performance during recordings were more variable in testing data.
- Performance did not vary systematically between rats or between different recording depths.

- Low absolute variability in the transform may flag sessions with low-quality extractions.



Artifact Rejection



- Rejection thresholds may not generalize across recordings of different quality.

Discussion

- We chose a criterion of three out of six raters to construct our ground truth.
- Our validation procedure identified parameters that generalized well to other recordings.
- Human failures favored false negatives.
- Different parameter sets in the automatic detection may favor false negatives or false positives.
- Putative artifact rejection measures disproportionately identified true positive events in good recordings.
- Data, code, and scoring guidelines will be posted at CRCNS.org and on our website.

Future Directions

- Validation and comparison of different spindle algorithms using rodent data
- Comparison of spindles and their automatic detection in rodents and humans
- Development of machine learning approaches to spindle detection[4]
- Assessment of spindle rating by humans (intra/inter-rater reliability)
- Development of new artifact identification measures and databases

References

1. Siapas & Wilson (1998) Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep.
2. Maingret et al. (2016) Hippocampo-cortical coupling mediates memory consolidation during sleep.
3. Warby S et. al (2014) Sleep spindle detection: crowdsourcing and evaluating performance of experts, non-experts, and automated methods.
4. Tan D et al. (2015) Sleep spindle detection using deep learning: a validation study based on crowdsourcing.

Acknowledgments

Support was provided by ONR MURI N000141310672, N000141612829, and N000141512838. We thank Dr. Marco Contreras, Dr. Bruce Harland, Erin Howard, and Raven Padgett for contributing spindle scoring, and the entire laboratory for helpful feedback.

Contact: blaineharper@email.arizona.edu