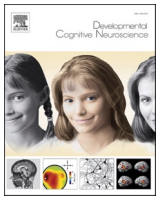
Developmental Cognitive Neuroscience 54 (2022) 101068

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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



NEAR: An artifact removal pipeline for human newborn EEG data

Velu Prabhakar Kumaravel [a,b], Elisabetta Farella [a], Eugenio Parise [b], Marco Buiatti [b,*]

[a] Fondazione Bruno Kessler, Trento, Italy
[b] CIMeC, Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy

ARTICLE INFO

Keywords:
EEG
Newborns
Infants
Artifact removal
Local Outlier Factor
Artifact Subspace Reconstruction

ABSTRACT

Electroencephalography (EEG) is arising as a valuable method to investigate neurocognitive functions shortly after birth. However, obtaining high-quality EEG data from human newborn recordings is challenging. Compared to adults and older infants, datasets are typically much shorter due to newborns' limited attentional span and much noisier due to non-stereotyped artifacts mainly caused by uncontrollable movements. We propose Newborn EEG Artifact Removal (NEAR), a pipeline for EEG artifact removal designed explicitly for human newborns. NEAR is based on two key steps: 1) A novel bad channel detection tool based on the Local Outlier Factor (LOF), a robust outlier detection algorithm; 2) A parameter calibration procedure for adapting to newborn EEG data the algorithm Artifacts Subspace Reconstruction (ASR), developed for artifact removal in mobile adult EEG. Tests on simulated data showed that NEAR outperforms existing methods in removing representative newborn non-stereotypical artifacts. NEAR was validated on two developmental populations (newborns and 9-month-old infants) recorded with two different experimental designs (frequency-tagging and ERP). Results show that NEAR artifact removal successfully reproduces established EEG responses from noisy datasets, with a higher statistical significance than the one obtained by existing artifact removal methods. The EEGLAB-based NEAR pipeline is freely available at https://github.com/vpKumaravel/NEAR.

1. Introduction

Studying human newborns in the first days of life provides key insights on the neurocognitive predispositions that humans are endowed with before interacting with the outside world. While most of the research in this field is behavioural, the recent availability of high-quality Electroencephalography (EEG) systems suitable for newborns opened the way to an increasing number of investigations on the neural bases of such predispositions with EEG (Beauchemin et al., 2011; Buiatti et al., 2019; Fifer et al., 2010; Ronga et al., 2021).

However, analyzing newborn EEG data is a challenging task, especially in the case of visual stimulation, because of two main factors: 1) Due to newborns' limited attentional span, the data segments during which newborns effectively attend to the stimuli are very short; 2) Since newborns are unconstrained, the most frequent artifacts are caused by a variety of movements (head, arms, frowning, sucking) which generate non-stereotyped artifacts that constantly vary in topography and temporal dynamics. Because of these factors, artifact removal for newborn EEG data is an arbitrary and time-consuming task. Since most artifacts are non-stereotyped, ICA-based methods that are successful with adults (Mognon et al., 2011; Pion-Tonachini et al., 2019) or older infants (Leach et al., 2020) might not be equally efficient in this case because ICA captures only stereotyped artifacts (Onton et al., 2006).

One promising tool for correcting non-stereotyped artifacts is Artifact Subspace Reconstruction (ASR), an algorithm specifically designed to remove transient or large-amplitude artifacts of any nature (Kothe and Jung, 2016). However, ASR performance depends on some user-defined parameters that have not been established for developmental data. Moreover, both ASR and ICA require a preliminary bad channel detection step and, as we show in this paper, the ones proposed by several state-of-the-art methods are too strict for analyzing newborn EEG data, especially with frequency-tagging paradigms that are less affected by artifacts than ERP designs.

Here we propose NEAR (Newborn EEG Artifact Removal), a method for efficient artifact removal from raw newborn EEG data. Compared to existing methods for artifact removal, NEAR introduces two innovative features: First, a novel bad channel detection tool relying on the Local Outlier Factor (LOF), a robust, density-based local outlier detection algorithm (Breunig et al., 2000); Second, a standard procedure for calibrating the two user-defined key parameters of ASR to newborn EEG

* Correspondence to: CIMeC, Center for Mind/Brain Sciences, University of Trento, Piazza Manifattura 1, 38068 Rovereto, Italy.
E-mail address: marco.buiatti@unitn.it (M. Buiatti).

https://doi.org/10.1016/j.dcn.2022.101068
Received 18 June 2021; Received in revised form 15 November 2021; Accepted 13 January 2022
Available online 15 January 2022

data: ASR parameter *k* and ASR processing mode, which can be either correction or removal of the detected bad segments.

In this paper, we start by illustrating each processing step of NEAR within a full pre-processing pipeline transforming the raw data to artifact-free data ready to be analyzed. This pipeline includes a procedure for calibrating both the bad channel detection and the bad segment correction/removal parameters.

We then describe the three steps used to test NEAR performance:

1) As a proof-of-concept, by using the simulation toolbox SEREEGA (Krol et al., 2018), we tested NEAR on simulated, neurophysiologically plausible EEG data, including transient, high-amplitude artifacts predominantly found in EEG data of newborns and young infants;
2) We tested NEAR performance on newborn EEG data based on a frequency-tagging paradigm (Buiatti et al., 2019), an experimental design consisting of a periodic temporal stimulation and measuring the stimulus-related response with the EEG oscillations at the stimulation frequency. This design is increasingly used with infants (de Heering and Rossion, 2015; Kabdebon et al., 2015) and newborns (Buiatti et al., 2019) as it generally achieves a higher signal-to-noise ratio than Event-Related Potential (ERP) designs (Norcia et al., 2015).
3) To test NEAR performance on an older population with an event-related design and with data recorded in another lab, we also evaluated NEAR on infant EEG data recorded on 9-months-old infants with an ERP paradigm (Parise and Csibra, 2012).

We calibrated NEAR parameters on a training dataset for the tests on real data and validated NEAR on an independent dataset.

Validation also included comparison with state-of-the-art methods: EEGLAB's *clean_rawdata* function (for bad channel detection only) and two automated artifact removal pipelines designed for developmental research, HAPPE (Gabard-Durnam et al., 2018) and MADE (Debnath et al., 2020).

NEAR scripts are made freely available for the users, along with an anonymized example dataset. A user-friendly step-by-step tutorial of the entire pipeline for the use of NEAR is provided in the Appendix.

2. Materials and methods

2.1. Training and Test datasets

2.1.1. Newborn datasets

Newborn Training and Test datasets belong to two studies performed at the Neonatal Neuroimaging Unit (CIMeC, University of Trento) installed in the maternity ward of Rovereto Hospital "Santa Maria del Carmine" (Rovereto, Italy). Both studies were approved by the local ethical committee for clinical research (Comitato Etico per le Sperimentazioni Cliniche, Azienda Provinciale Servizi Sanitari, Province of Trento, Italy); parents were informed about the content and goal of the study and gave their written informed consent.

Both datasets were recorded by an EGI EEG system (GES400, Electrical Geodesic, Inc, Eugene, OR, USA) with 125 channels. Scalp voltages were referenced to the vertex, amplified and digitized at 250 Hz. Electrode impedances were kept below 100 kΩ. Newborns were tested in a calm, dimly illuminated space in the maternity ward, seated on the lap of a trained researcher in front of a 60 cm × 33.8 cm LCD screen (distance eyes-screen: about 30 cm) while wearing the EEG cap. Video recording from a hidden camera on the top of the screen provided online monitoring of the infant. The newborn's parents, when present, were off the sight of the infant (separated by a curtain) and instructed to keep silent during the recordings. For both datasets, visual stimuli were presented dynamically with sinusoidal contrast modulation (the visibility of each stimulus gradually rises with respect to the gray background from 0% at the beginning of the cycle to 100% at mid-cycle, then gradually decreases to 0% towards the end of the cycle, see Fig. 1 in (Buiatti et al., 2019), at a rate of 0.8 Hz (frequency-tagging paradigm). We used sinusoidal contrast modulation instead of squared on–off dynamics, both to minimize nonlinear effects in the brain frequency response (Norcia et al., 2015) and to make the stimulation more pleasant to the babies (de Heering and Rossion, 2015). The slow presentation rate (0.8 Hz) was chosen to ensure that newborns fully perceived the stimulus at each cycle of the periodic, peekaboo-like presentation.

The Training Dataset is part of an ongoing study investigating the neural bases of number perception in newborns (Buiatti et al., in preparation). Visual stimuli consisted in a set of 4 or 12 coloured simple geometrical shapes, presented in blocks of 50 s or until the subject stopped attending them; shape, number and spatial arrangement were constant within each block and randomly changed between blocks. For the whole duration of the study, an auditory stimulation consisting of sequences of syllables was simultaneously presented (the response to the auditory stimulation will not be considered here). For the purpose of this paper, the Training Dataset includes all the subjects that attended at least 15 s of visual stimulation, independently whether they attended one or both number conditions (11 newborns, six males; mean age 40 ± 16 h; all were healthy [APGAR(1 min) ≥ 8, APGAR(5 min) = 10 for all subjects] and born full-term (gestation age, 39.9 ± 0.9 wk).

The Test Dataset belongs to a study investigating the cortical bases of facelike pattern processing (Buiatti et al., 2019). Visual stimuli consisted of a white head-shaped form containing three black squares and differed only in the spatial configuration of the three squares to form the three stimuli (upright face, inverted face, and scrambled face). Stimuli were presented in blocks of 50 s or until the subject stopped attending them. Subjects were 10 healthy newborns (six males; mean age 60 ± 22 h). All were healthy [APGAR(1 min) ≥ 8, APGAR(5 min) = 10 for all subjects] and born full-term (gestation age, 39.7 ± 1.5 wk). Further details in (Buiatti et al., 2019).

2.1.2. Infant datasets

Infant Training and Test datasets belong to a study investigating semantic understanding of common nouns in preverbal infants, performed at the Cognitive Development Center (CDC, Central European University) and whose results are published in (Parise and Csibra, 2012). Ethical approvals were obtained from the ethics committee of the Central European University, Budapest; parents were informed about the content and goal of the study and gave their written informed consent. All infants were born full term (gestational age: 37–41 weeks) in the normal weight range (> 2500 g).

Both the Training and the Test dataset included 14 healthy infants (Training: 6 females; mean age = 278 days, range = 266–285 days. Test: 5 females; mean age = 277 days, range = 269–286 days). Both datasets were acquired using an EGI amplifier (GES 300, Electrical Geodesic, Inc, Eugene, OR, USA) at a sampling rate of 500 Hz with a low-pass filter at 200 Hz. Continuous EEG was recorded by 125-channel Geodesic Sensor Nets referenced to the vertex. Infants were tested in a calm, dimly illuminated room in the CDC BabyLab, sitting on a high chair 70 cm in front of a 9-inch, 800 × 600, 100 Hz CRT monitor. The infants were video-recorded throughout the session from a hidden camera placed below the presentation monitor. The infant's mother and an experimenter sat on chairs at either side of the infant.

For both datasets, each trial started with a live auditory stimulus delivered either by the experimenter (Training dataset) or by the infant's mother (Test dataset) while a dynamic fixation stimulus (a colorful rectangle 343 ×363 pixels) was presented on top of an occluder. After the live auditory stimulus ended, the fixation stimulus stopped moving, and the display remained frozen for 600–800 ms. Then the fixation stimulus disappeared, and the occluder started to fall forward (a 90° rotation on the basis-hinge) revealing an object behind it (see Fig. 1 in Parise and Csibra, 2012). The object, laying on a black background, was fully visible for 1000 ms before the occluder began to rise, hiding the object again. This was followed by an intertrial interval lasting 1100

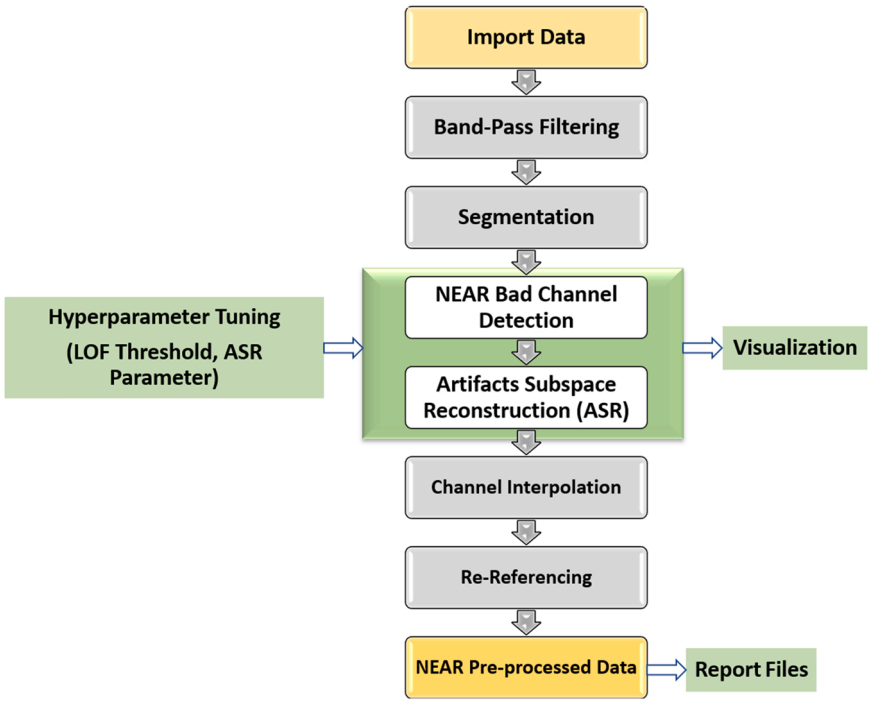


Fig. 1. Schematic representation of NEAR pipeline. Green boxes indicate the artifact removal part. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to 1300 ms. The pictures of 15 different objects were used; their average size was 302.5 × 321.6 pixels. Trials were presented as long as the infants were attentive. The minimum inclusion criterion for the infants was at least 10 artifact-free trials in each of two experimental conditions. Further details in (Parise and Csibra, 2012).

2.2. NEAR

NEAR preprocessing pipeline consists of a set of custom MATLAB scripts that can be executed as a fully automated EEG preprocessing within the EEGLAB (Delorme and Makeig, 2004) framework. The core, innovative parts of the pipeline integrating EEGLAB scripts with original custom scripts consist in the artifact removal processing block: preliminary calibration of the bad channel detection threshold (LOF) and of the ASR cut-off parameter, bad channel detection using LOF algorithm and correction/removal of bad segments using ASR, both endowed with original visualization of the outcomes. In addition, we provided the scripts (based on EEGLAB functions) for a fully automated EEG processing from raw to clean data: importing and filtering raw data, interpolation of removed channels and re-referencing. Depending on the application requirements, these auxiliary steps can easily be modified. Fig. 1 shows the steps involved in the NEAR preprocessing pipeline. In the following sections, we describe each step in detail. A step-by-step tutorial including figures illustrating the main steps of NEAR artifact removal is presented in the Appendix of this manuscript.

2.2.1. Import raw data

NEAR supports import functionality of four main formats: .mff, .raw, .set and .edf. We considered these formats because most developmental EEG raw data fall into one of these categories. For other formats, users can import the data with EEGLAB importing tools and use NEAR with the resulting EEGLAB format .set.

2.2.2. Band-pass filtering

The principle underlying band-pass filtering is that it is convenient for all the subsequent analyses to keep the frequency range of the signal that we want to analyze and discard the higher and lower frequencies, especially those that likely contain artifacts. At the higher end, it is beneficial to use a low-pass filter with a cut-off frequency below the power line (50 Hz or 60 Hz) to avoid line noise. At the lower end, i.e. below 1 Hz, the EEG signal typically contains eye movement, respiration and heart-beat artifacts. We, therefore, recommend using the highest high-pass filter cut-off frequency that preserves the signal of interest, paying attention to the width of the filter roll-off and, in the case of ERP designs, to the risk of introducing spurious effects (Acunzo et al., 2012).

For the newborn data, we applied a low-pass FIR filter with a cut-off frequency at 40 Hz (by using EEGLAB's default filter). Since for the analysis considered in this paper, we need to preserve the frequency components down to 0.5 Hz (see Section Neural measure: FTR), we used a non-causal high pass filter between 0.15 and 0.3 Hz and a stop-band attenuation of 80 dB.

The infant data were band-pass filtered between 0.3 and 30 Hz by using the default EEGLAB filter.

2.2.3. Data segmentation

For studies involving a stimulation paradigm, a key pre-processing step for identifying the relevant data in newborn/infant EEG recordings is to restrict data analysis to the intervals during which newborns/infants were effectively attending to the stimuli. We, therefore, recommend segmenting the data related to stimulation (i.e. segmenting stimulation periods for continuous stimulation or segmenting event-related epochs in case of event-related designs). Furthermore, for visual stimulations, we strongly recommend recording newborns/infants with a camera or an eye-tracker and light conditions guaranteeing clear monitoring of eye movements, and devoting careful attention to the identification of the effective looking times. This pre-processing step is crucial because not only it minimizes noise in the data, but it also removes data segments associated with unattended intervals that are usually very artifacted, potentially causing biases in the subsequent artifact analysis. For this reason, this step is performed before detecting bad channels and segments.

For resting-state EEG studies as well, our scripts can be adapted to retain good segments (or remove bad segments) of data known apriori. See Appendix, Step 4 for details.

2.2.4. Bad channel detection

Bad channel detection in newborn/infant EEG data is a challenging step because, due to typically short preparation time devoted to lowering electrode impedance and frequent movement artifacts, electrode contact and stability is generally much lower than in adult data. In particular, after some preliminary tests on our data, we realized that the existing methods of bad channel detection are generally too strict for newborn EEG data. To overcome this issue, we implemented an algorithm in which the core step is a novel bad channel detection method based on LOF, a robust, data-driven outlier detector. The three steps of the algorithm are as follows:

2.2.4.1. Flat signals. Because of defective contact with the scalp or disconnection from the recording device, sometimes electrodes record a flat signal. To remove these channels, we adopted the function *clean_flatlines* from the EEGLAB *clean_rawdata* plugin (https://github.com/sccn/clean_rawdata). A channel is marked as flat by default if it records a flat signal for more than 5 consecutive seconds.

2.2.4.2. Local Outlier Factor (LOF). Traditional outlier detection methods based on statistical measures such as mean, median, IQR or mean absolute deviation are too sensitive to outliers in the context of newborn EEG. To tackle this challenge, we introduce (to the authors' knowledge, for the first time in the context of EEG data analysis) a robust unsupervised method called Local Outlier Factor (LOF), a density-based data-driven approach (Breunig et al., 2000) to detect and remove bad channels. This technique operates in a multidimensional channel space where the "distance" between channels is computed as a robust distance estimation (Squared Euclidean distance ('seuclidean' in MATLAB)) between the activity vectors associated to each channel (i.e., the time series of each EEG signal) (not to be confused with the physical distance between the channels). Precisely, it assigns each channel a degree of "local outlierness" depending on how isolated the channel is with respect to its *k* neighbor channels (Fix and Hodges, 1989).

To demonstrate the efficiency of the LOF algorithm, we show Fig. 2 that contains sample data clusters for illustration purposes. Suppose C_1 and C_2 are two main clusters and two additional objects o_1 and o_2. As shown in Fig. 2, both objects o_1 and o_2 are outliers for the respective clusters C_1 and C_2. While most statistical-based and distance-based algorithms would correctly capture the o_1 as an outlier, LOF being a local density-based approach is capable of identifying objects like o_2 as well.

LOF algorithm is implemented as follows:

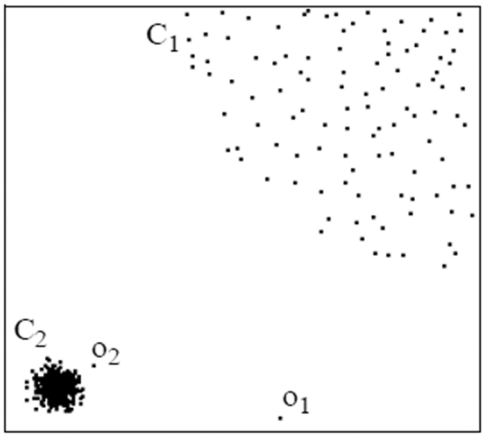


Fig. 2. A sample dataset that contains two clusters of data (C_1 and C_2) and two outlier objects (o_1 and o_2).

1) For each channel *p*, LOF algorithm identifies *k* nearest neighbors based on a distance metric (by default, Squared Euclidean in NEAR pipeline).

2) A reachability distance is computed between a channel *p* and each neighbor. For example, let us consider channel *o* that falls within the *k* neighbors of channel *p*. Then, the reachability distance between *p* and *o* is computed as follows:

$$reachability\_dist(p, \quad o) = \max\{\text{k-distance(o)}, \quad d(p,o)\}$$

where *d(p,o)* is the actual distance between two channel vectors.

Intuitively, if channel *p* is far away from *o*, then the reachability distance is simply their actual distance. Instead, if they are "sufficiently" close, the actual distance is replaced by the *k-distance* of channel *o*.

3) Once, the reachability distances of each channel with respect to its neighbors is computed, the local reachability density (LRD) is determined as follows:

$$LRD_k(p) = \frac{1}{\left(\frac{\sum_{o \in N_k(p)} reachability\_dist_k(p,o)}{|N_k(p)|}\right)}$$

To put it in words, LRD of the channel *p* is the inverse of the average reachability distance based on the *k*-nearest neighbors of *p*. Intuitively, channel *p* will have a lower LRD if it were an outlier (i.e., bad) channel because it is not easily reachable by its neighbors.

4) Then, the local outlier factor (LOF) is computed as follows:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{LRD_k(o)}{LRD_k(p)}}{|N_k(p)|}$$

LOF of channel p is the average of the ratio of the LRD of p and those of *p*'s *k*-nearest neighbors. The lower *p*'s LRD is, and the higher the LRD of *p*'s *k*-nearest neighbors are, the higher is the LOF value of *p* (and therefore possibly an outlier). In other words, an outlier channel would display a lower LRD (therefore, larger in distance) compared to its neighbours (on average).

As it can be noted, *k* is the hyperparameter in the computations of LOF. In this work, we used the Natural Neighbors algorithm (Zhu et al., 2016) to compute a data-driven *k* value. We adapted the MATLAB-based implementation of the LOF algorithm (Density-based Outlier Detection Algorithms, https://github.com/BlueBirdHouse/DDoutlier. Retrieved June 3, 2021.) in order to make it compatible with the EEGLAB data structure. Further, following a systematic study of the application of LOF to EEG data (Kumaravel et al., 2021), we found the Standardized Euclidean distance metric (defined as '*seuclidean*' in MATLAB) outperforming the default Euclidean metric to compute the *k*-distance (*knnsearch* function, MATLAB).

Once LOF is computed for all the channels in a given dataset, it is important to set a threshold to separate the inlier channels from the outlier ones. In the original theory, if a channel exhibits an LOF of more than 1.5 (section 7.3 in Breunig et al., 2000), it shall be considered as an outlier. For the application to EEG data that we are considering in this work, we proposed an adaptive approach by estimating the optimal threshold value for LOF from a Training Dataset on the basis of a standard scoring of the bad channels (See Section 3.2.1).

2.2.4.3. Periodogram Analysis. To detect channels that predominantly

recorded motion-related artifacts that manifest as an increase in power in Beta range, and a decrease in power in Delta and Alpha ranges (Georgieva et al., 2020), we implemented a bad channel detection method based on a spectral measure (*periodogram* function, MATLAB). For the datasets analyzed in this paper, we noted that while this method captures the most significant bad channels, they were already detected by at least one of the previous two steps. Therefore, we kept this method as optional in NEAR's bad channel detection plugin (see Fig. A1 in Appendix).

2.2.5. Artifact removal using Artifact Subspace Reconstruction (ASR)

ASR is an automated artifact removal technique to detect/remove transient high-amplitude artifacts in continuous EEG data (Kothe and Jung, 2016). It is available as an open-source EEGLAB plug-in function *clean_rawdata*. ASR has been tested extensively both on simulated data and on real EEG acquired using mobile setup from adult participants (Kumaravel et al., 2021; Mullen et al., 2015). Thanks to its efficient artifact removal, ASR is now considered as one of the default preprocessing algorithms within the EEGLAB framework. However, ASR has only been evaluated on adult EEG data thus far (Blum et al., 2019; Chang et al., 2020; Mullen et al., 2015). For the first time, in this work, we evaluate ASR on noisier developmental EEG data and propose it as one of the core blocks in our pipeline. In addition, we propose a calibration procedure for adapting ASR algorithm to developmental data. ASR processes artifacts in three steps that are briefly described as follows (for more detailed technical documentation, please refer to Kothe and Jung, 2016, and Chang et al., 2020).

2.2.5.1. ASR algorithm.

1) First, ASR identifies cleaner data portions according to a predefined robust statistical distribution of EEG-like data.

2) Then, ASR performs Principal Component Analysis (PCA) on the obtained cleaner segments of data to extract a rejection threshold, defined as follows:

$$T_i = \mu_i + k * \sigma_i$$

where i is the Principal Component (PC) index, μ and σ are the corresponding mean and variance and k is the user-defined multiplicative SD factor (also known as ASR cut-off parameter).

3) With the extracted threshold T, ASR identifies the artifacts subspace on the original data and reconstructs them based on the statistics obtained using the cleaner portions of the data.

To calibrate ASR to newborn/infant EEG data, we analyzed two crucial user-defined parameters of ASR:

2.2.5.2. ASR cut-off parameter (k). ASR defines an upper-bound threshold for a PC representing EEG-like components based on the mean and variance of PCs extracted from the cleaner portion of the data. Therefore, the components exceeding this threshold are most likely artifactual. The threshold is computed as defined in step 2) of the previous subsection. It can be observed that a lower k implies a lower threshold and therefore a strict artifact detection (i.e. more artifacts are detected); a higher k implies a looser cleaning of the data (i.e. less artifacts are detected). For adult EEG, the optimal k values lie in the range between 20 and 30 (Chang et al., 2020). As mentioned before, to the best of our knowledge, the ASR parameter k has never been evaluated on developmental data.

2.2.5.3. Processing mode. Using the *clean_rawdata* plugin, ASR can be operated in two distinct modes: ASR Correction (hereafter, indicated as ASR_C throughout this manuscript) in which the bad portions of the data are corrected to 'EEG-like' data, and ASR Removal (indicated as ASR_R) in which the detected bad portions are removed from the data.

To calibrate these two parameters to newborn EEG data, a grid-search was performed on the Training Dataset (see Results).

2.2.6. Bad channel interpolation

The removed channels are interpolated from neighbouring channels by using EEGLAB's function *pop_interp*. As suggested by EEGLAB developers, we recommend using *spherical* interpolation. However, using the NEAR pipeline, it is possible to use other supported techniques such as *v4*.

2.2.7. Re-referencing

For re-referencing, NEAR provides options for both average re-referencing (recommended and most commonly used in developmental EEG studies) and re-referencing to a particular channel (e.g., Cz). For this task, NEAR uses EEGLAB's *pop_reref* function.

2.2.8. Calibration of artifact removal parameters

A key feature of NEAR is the preliminary calibration of its artifact removal parameters (LOF bad channel threshold, ASR parameter k and ASR processing mode). We provide scripts for this calibration and we highly recommend NEAR users to perform it on previously analyzed datasets from the same setup and experimental design as these parameters impact the quality of preprocessing (see Section 3.2.1 NEAR parameter calibration).

2.2.9. Other functionalities of NEAR

NEAR supports both single-subject processing and batch-processing (in case of multiple subjects). The relevant scripts for these functionalities can be found in the repository.

Finally, NEAR supports saving functionality and provides a comprehensive report that summarizes the preprocessing done on each of the input EEG files. This report might be useful to review the effects of preprocessing done on the raw input EEG.

2.3. Validation tools

2.3.1. Simulated data

Simulated data were generated with SEREEGA (Krol et al., 2018), a Matlab-based toolbox that simulates EEG datasets consisting in neurophysiologically realistic continuous and/or event-related brain activity. We generated two datasets simulating newborn EEG data with a frequency-tagging stimulation as in (Buiatti et al., 2019), and with an event-related stimulation similar to the one in (Parise and Csibra, 2012), respectively. More specifically, we generated a 64-channels EEG dataset with the following components:

2.3.1.1. Component 1. A stimulus response, in the form of a sinusoidal Steady-State Visual Evoked Potential (SSVEP) (stimulation frequency = 0.8 Hz) for the frequency-tagging stimulation, and in the form of an event-related potential (latency=300 ms) for the event-related stimulation. Both responses were localized in two bilateral sources in the early visual cortex (MNI coordinates: [-8-76 10] and [8-76 10]).

2.3.1.2. Component 2. Event-unrelated ongoing EEG activity originating in 62 randomly selected cortical sources, plus in the 2 sources of the first component located in the early visual cortex. Such activity is generated as Brown noise (power spectrum increasing as $1/f^2$ for $f \rightarrow 0$), mimicking the one observed in newborns (Fransson et al., 2013). Importantly, the signal-to-noise ratio between component 1 and component 2 was of the same order of magnitude as the one measured on real, artifact-free newborn EEG data. The first two components represent the *ground truth*.

2.3.1.3. Component 3. Artifacts in (5 randomly chosen) single channels consisting in intermittent potential shifts and flat signals mimicking electrical discontinuities, and low-frequency fluctuations (0–10 Hz) mimicking local bad contacts and movement artifacts;.

2.3.1.4. Component 4. Transient high-amplitude artifacts involving all the channels in the form of intermittent abrupt potential shifts or smoother Gaussian-like fluctuations, where both the amplitude at each channel and the duration varies randomly for each transient artifact (mean duration=1.6 s). Durations and amplitudes are of the same order of magnitude as those observed in real newborn data. This component mimics motion artifacts, which are very frequent in newborns.

Fig. 3 shows that this simulation well represents the main features of newborn EEG ongoing activity and artifacts. The scripts generating the simulation datasets are available at https://github.com/vpKumaravel/NEAR and the simulated datasets described in the Results are available here: https://osf.io/79mzg/.

2.3.2. Standard semi-automatic expert artifact removal procedure

As a reference for a standard semi-automatic artifact processing performed by experts (hereafter abbreviated as *standard*), we report the original procedures of artifact rejection performed in the original papers from which the newborn and infant datasets are taken (Buiatti et al., 2019; Parise and Csibra, 2012). Their bad channel scoring will be taken as the reference standard scoring both for the calibration and the validation of NEAR's bad channel detection algorithm.

2.3.2.1. Newborns. Bad channels were detected on both datasets after band-pass-filtering and segmentation. Channels were marked as bad if they 1) had a standard deviation (computed on the whole data length by using the TrimOutlier toolbox: https://sccn.ucsd.edu/wiki/TrimOutlier) higher than 150 μV (to detect channels with high-amplitude artifacts) or lower than 1 μV (to detect flat or weakly responsive channels); 2) showed artifactual patterns after accurate visual inspection of the time course and power spectrum plots of suspicious channels and comparison with their neighbours.

Once bad channels were removed, identification of bad data segments was based on 1) the detection of amplitude jumps exceeding ± 200 μV; 2) the presence of paroxysmal artifacts after accurate visual inspection of the time course and topography of the EEG data.

2.3.2.2. Infants. Both infant datasets were automatically and manually edited. Automatic data rejection for body and eyes movements was performed whenever the average amplitude of a 80 ms sliding window exceeded ± 200 μV at any channel. A bad channel score was obtained by considering as bad the channels that were marked as rejected for at least 40% of the epochs. Bad channels were automatically interpolated in epochs in which ≤ 10% of the channels contained artifacts; epochs in which > 10% of the channels contained artifacts were automatically rejected. Data was then manually edited by visual inspection of each individual epoch.

2.3.3. Other bad channel detection methods

To validate the performance of NEAR's channel rejection tool against existing methods, we considered the following three state-of-the-art bad channel removal methods:

1) The default EEGLAB function *clean_rawdata* (CRD, https://github.com/sccn/clean_rawdata) detects flat-line channels, channels contaminated with high-frequency noise and channels uncorrelated with its neighbors.
2) HAPPE (Gabard-Durnam et al., 2018) uses EEGLAB *pop_rejchan* function to detect bad channels based on amplitude and spectral thresholding (z-score threshold=3 instead of EEGLAB default 5), running it twice to avoid residual bad channels.
3) FASTER (Nolan et al., 2010) detects bad channels by computing the temporal correlation between channels, their variance, and a score based on the Hurst exponent.

2.3.4. Other automated pipelines for artifact removal in developmental EEG

2.3.4.1. MADE. The Maryland Analysis for Developmental EEG (MADE) is an automated standardized pre-processing pipeline specifically developed for developmental populations (Debnath et al., 2020). MADE uses FASTER (Nolan et al., 2010) to remove bad channels and ICA to correct data from artifacts. Bad ICs are classified automatically using Adjusted-ADJUST (Leach et al., 2020), a modified version of ADJUST (Mognon et al., 2011) developed specifically for infant data. Residual

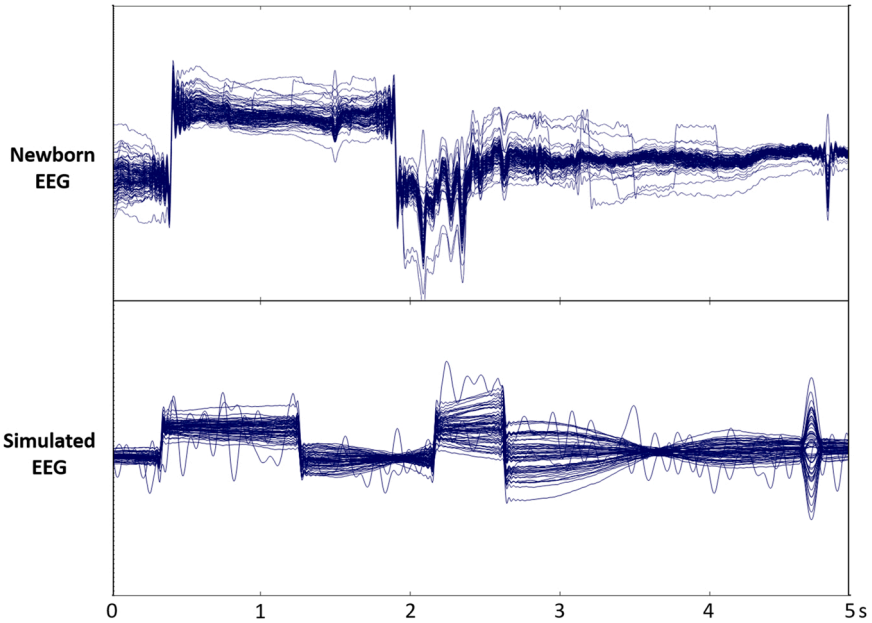


Fig. 3. Top Panel: Newborn EEG data from (Buiatti et al., 2019). Bottom Panel: Simulated EEG data (*ground truth* plus artifacts). Data are shown in butterfly mode (all electrode signals overlapped).

epochs contaminated by ocular artifacts are removed by employing a predefined amplitude threshold. MADE has been validated on infants starting from 1 year of age to childhood (3–6 years old) and late adolescence (16 years old).

2.3.4.2. HAPPE. The Harvard Automated Processing Pipeline for EEG (HAPPE) is a standardized automated pipeline for developmental EEG that contains high degree of artifact contamination and often short recording lengths (Gabard-Durnam et al., 2018). HAPPE pipeline consists of 9 steps including bad channel rejection using pop_rejchan.m (Delorme et al., 2015) and a wavelet-integrated ICA decomposition to recover artifactual segments. Bad ICs are classified automatically using MARA (Winkler et al., 2014). HAPPE has been validated on resting-state developmental EEG data (age between 3 and 36 months). As HAPPE is not suitable for event-related designs (Gabard-Durnam et al., 2018), we will compare it with NEAR on the continuous datasets only.

2.3.5. Neural measures for calibration and validation

2.3.5.1. Frequency-tagging designs: FTR. To compute a signal-to-noise ratio of the stimulus-related EEG response for both ASR parameter calibration and overall NEAR validation, we used the same measure defined in (Buiatti et al., 2019). EEG data were segmented in partially overlapping epochs of 10 s (overlap varied between one-half and three-fourths of epoch length to adjust to the variable length of clean data segments). For each electrode, the Fourier transform $F(f)$ of each epoch was calculated using a fast Fourier transform algorithm (MATLAB function FFT). To avoid rejecting data segments shorter than 10 s but still potentially containing relevant neural signals, zero-padding to 10 s was applied before FFT for data segments between 5 s and 10 s. Data segments shorter than 5 s were discarded. The power spectrum was calculated from these Fourier coefficients as the average over epochs of the single-epoch power spectrum:

$$PS(f) = \langle F(f) \times F^*(f) \rangle_{ep}$$

The Frequency-Tagged Response (FTR) at the tag frequency (0.8 Hz) was calculated as the ratio between the power spectrum at the tagged frequency and the background power, i.e. the value at 0.8 Hz of the power-law fit of the power spectrum estimated from the six neighboring frequency bins ($\pm$ 0.3 Hz), where the power-law fit was computed by fitting a line to the logarithm of the power at the six neighboring frequency bins (MATLAB function *Polyfit*).

2.3.5.2. Event-related potential designs: SNR(ERP). As a signal-to-noise ratio (SNR) of the ERP for ASR parameter calibration, we computed the one based on the Standardized Measurement Error (SME) recently proposed by (Luck et al., 2021). The SME is an estimate of the noise in the measure of an ERP score (computed on a time window and a set of electrodes) based on its trial-by-trial variability:

$$SME = \frac{SD(ERP_{tr})}{\sqrt{N}}$$

where $SD(ERP_{tr})$ denotes the standard deviation (across trials) of the single-trial ERP averaged over a time window and a set of electrodes, and N is the number of epochs. For each subject, the *SNR(ERP)* is the ratio between the ERP (averaged over trials) and the *SME*.

3. Results

3.1. Validation of NEAR on simulated data

We first validated NEAR on two synthetic EEG datasets simulating EEG signals that contain a SSVEP at 0.8 Hz like in (Buiatti et al., 2019) (*frequency-tagging dataset*) and an ERP response similar to the one recorded in (Parise and Csibra, 2012) (*ERP dataset*), respectively. Both datasets also include three key components of newborn/infant EEG data: Brown-noise-like background EEG, artifacts in single channels mimicking bad or unstable electrode contacts, transient high-amplitude fluctuations across most of the channels mimicking motion artifacts. Signal-to-noise ratios, data duration and proportion of artifacts are similar to the ones of real data (Buiatti et al., 2019; Parise and Csibra, 2012). Since it is difficult to incorporate enough variability to generate realistically different training and test datasets within the simulation framework, we set NEAR parameters to predefined values: LOF threshold= 2 and ASR parameter k = 20.

3.1.1. Frequency-tagging dataset

The ground truth data (SSVEP plus Brown-noise-like background EEG) shows a clear peak in the power spectrum at the stimulation frequency (0.8 Hz) that stands out of the background EEG power spectrum (blue line in Fig. 4). The topography of the associated FTR at 0.8 Hz shows a neat posterior medial activation (Fig. 4, bottom panel) fully compatible with the early visual cortex sources generated in the simulation (for details, see Section 2.3.1). Artifacts cause a massive positive shift of the power spectrum at low frequencies, almost completely masking the SSVEP response peak (red line in Fig. 4). Consequently, the topography of the FTR at 0.8 Hz does not show any clear posterior activation (Fig. 4, bottom panel).

NEAR bad channel detection algorithm efficiently captured the simulated 5 bad channels (and no additional channels). ASR_R was very efficient in removing all the transient artifacted segments from the data: the resulting peak at the stimulation frequency in the power spectrum almost overlaps with the one of the ground truth data (yellow line in Fig. 4), and the topography of the FTR at 0.8 Hz is very similar to the one of the ground truth (Fig. 4, bottom panel). ASR_C performance was slightly inferior: while the power spectrum peak was recovered, its amplitude was lower than the ground truth, and the overall power spectrum at low frequencies was shifted to lower values (magenta line in Fig. 4). This could depend on the fact that while all transient artifacts were correctly detected and removed by ASR, the correction also suppressed part of the SSVEP and of the background EEG. Nevertheless, the FTR topography was very similar to the ground truth one, even if with a slightly lower amplitude (Fig. 4, bottom panel).

For comparison with state-of-the-art methods for artifact removal, we also tested the ICA-based artifact removal pipelines of MADE (Debnath et al., 2020) and HAPPE (Gabard-Durnam et al., 2018). MADE was not able to correct or remove almost any of the transient artifacts, as shown by its power spectrum (green line in Fig. 4) and its FTR topography at 0.8 Hz (Fig. 4, bottom panel), that are both very similar to the ones of the contaminated data. HAPPE was more successful: it corrected most of the low-frequency artifacts (cyan line in Fig. 4) and FTR topography shows a posterior activation similar to that of the ground truth, although with a much lower amplitude than the ground truth and NEAR processing with ASR in both modalities (Fig. 4, bottom panel). The rationale behind this reduction in overall amplitude might be due to the wavelet-based ICA thresholding, as also highlighted by the authors (Gabard-Durnam et al., 2018).

3.1.2. ERP dataset

Results on the ERP dataset were very similar to the ones of the frequency-tagging dataset. The ground truth ERP mildly fluctuates around zero until 200 ms, then rises at 300 ms (the peak latency) and decreases again afterwards (blue line, top panel of Fig. 5). Its topography at the peak latency is neatly posterior (bottom panel, Fig. 5). Artifacts cause the ERP to spuriously rise even before the stimulus onset, and although the ERP peak is visible in the posterior electrodes (red line, top panel of Fig. 5), the topography at the peak latency is very noisy (bottom panel, Fig. 5).

NEAR bad channel detection algorithm efficiently captured the simulated 5 bad channels (and no additional channels). ASR_R was very efficient also in this case in removing all the transient artifacted

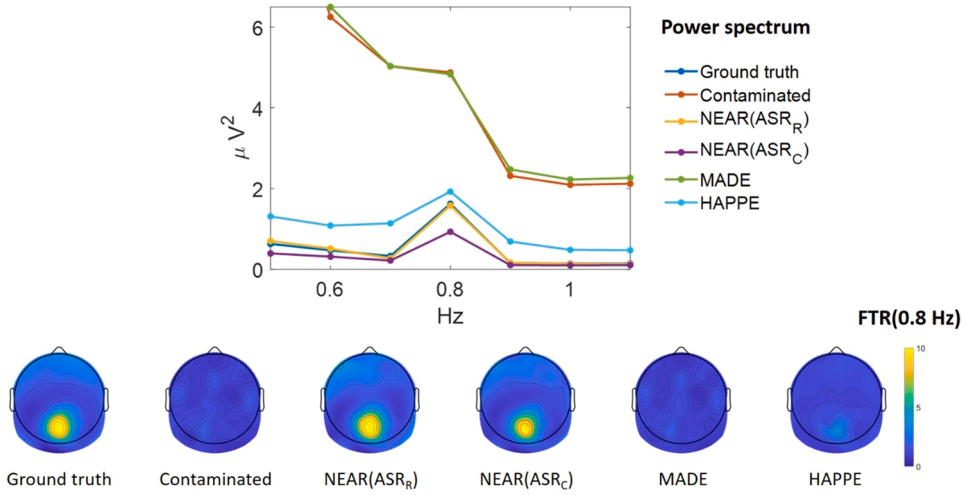


Fig. 4. Top panel: Power spectrum of the simulated *frequency-tagging dataset* between 0.5 and 1.1 Hz, averaged over the electrodes showing the largest FTR amplitude in the *ground truth* data (PO3, POz, PO4). Bottom panel: Topography of the FTR (defined in Section 2.3.5) at 0.8 Hz (the stimulation frequency). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

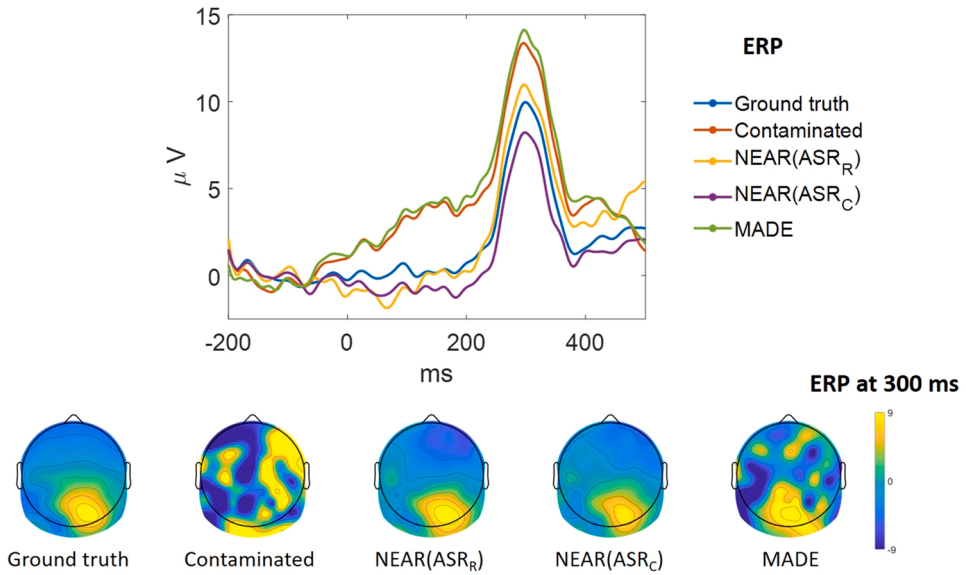


Fig. 5. Top panel: ERP of the simulated *ERP dataset* around the onset of the simulated stimulus, averaged over the electrodes showing the largest ERP amplitude in the *ground truth* data (PO3, POz, PO4). Bottom panel: Topography of the ERP averaged between 275 and 325 ms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

segments from the data: the ERP peak at 300 ms almost overlaps with the one of the ground truth, even if the ERP profile is a bit noisier at higher latencies (yellow line, top panel of Fig. 5), possibly an effect of the lower number of trials. The topography at 300 ms is very similar to the one of the ground truth (Fig. 5, bottom panel). ASR_C ERP peak has a lower amplitude than the ground truth but the ERP profile outside the peak is very clean with low fluctuations around zero (magenta line in Fig. 5). The ERP topography at 300 ms is as neat as the one of the ground truth (Fig. 5, bottom panel).

In comparison, also in this case, MADE could not remove the artifacts on the electrodes showing the posterior activation (green line, top panel of Fig. 5). However, its topography at 300 ms shows moderate success in removing artifacts from other electrodes, though much less successfully than NEAR (Fig. 5, bottom panel).

3.2. Validation of NEAR on newborn data

3.2.1. NEAR parameter calibration

We first calibrated the parameters of bad channel detection and ASR on the Training Dataset.

3.2.1.1. Calibration of LOF bad channel detection. NEAR's bad channel detection algorithm (Flat lines + LOF) was tested by comparing it to the *standard* bad channel detection score implemented in the original paper ((Buiatti et al., 2019), see Section 2.3.2 for details) with the quality metric F1 Score defined as

$$\text{F1 Score} = \frac{2 * TP}{2 * TP + FP + FN}$$

where TP, FP and FN indicate the number of True Positives, False Positives and False Negatives respectively (Dalianis, 2018).

By changing the LOF threshold from 1 to 10 in steps of 0.1, we found

that the maximal F1 Score was achieved with a threshold of 2.5 (Fig. 6). We therefore selected this value for performing bad channel detection on the newborn Test Dataset.

3.2.1.2. Calibration of ASR. To identify the optimal ASR parameter *k* and processing mode, we applied ASR on the newborn Training Dataset while systematically varying ASR parameter *k* between 1 and 100 for both modes of processing (bad segment removal (ASR_R) and correction (ASR_C)). As a validation measure, after a preliminary bad segment removal by visual inspection, we identified a broad occipital cluster of electrodes showing a visual response (Fig. 7, top inset); we then computed the average visual response FTR (see Materials and Methods) in this predefined occipital cluster for each *k* and processing mode. Results show that both processing modes achieve a similar maximum value of FTR by ($t(10) = -0.28$, $P = 0.78$), but for different *k* values: $k = 24$ for removal mode, $k = 13$ for correction mode. One possible explanation of this difference is that while for *k* between 20 and 30 the correction is not very effective, for $k < 15$ the removal mode rejects too many segments, providing too few samples for a robust computation of FTR. Since the two processing modes provide equivalent results for their optimal *k,* we will test both modes in the validation phase.

3.2.2. NEAR validation

3.2.2.1. NEAR bad channel detection. Once the optimal parameters were identified by calibration, we used them to validate NEAR artifact removal on the newborn Test Dataset. First, we validated NEAR's bad channel detection method by evaluating its performance in matching the bad channel scoring implemented in the original study (Buiatti et al., 2019), here considered as the ground truth (see Section 2.3.2 for details). We also compared its performance with one of the three state-of-the-art methods: the default EEGLAB function *clean_rawdata* (CRD) and the bad channel detection methods used by two popular pipelines specifically implemented for infant EEG data, HAPPE (Gabard-Durnam et al., 2018) and MADE (Debnath et al., 2020) (the latter using FASTER bad channel detection tool (Nolan et al., 2010)). As shown in Table 1, the number of bad channels in the ground truth widely varies between subjects, from a minimum of zero to a maximum of 14. Results (Table 1) show that NEAR is the tool that best captures this high variability (F1 score = 0.81). All the other methods tend to mark more bad channels (therefore, more false positives with respect to the ground truth).

3.2.2.2. Overall NEAR validation. Then we validated the overall performance of NEAR artifact removal by testing whether the EEG data cleaned by NEAR showed the statistical significance of the two main neural responses described in (Buiatti et al., 2019): 1) The EEG response to the overall visual stimulation, by comparing the power at the tag frequency with the background power estimated at the same frequency in the occipital cluster of electrodes identified in (Buiatti et al., 2019) (Fig. 2A therein); 2) The facelike pattern response, comparing the FTR to facelike stimuli with the one to inverted facelike patterns in the posterior cluster of electrodes illustrated in (Buiatti et al., 2019) (Fig. 3A therein). We also compared NEAR performance with the one obtained using a standard artifact processing as in the original paper (Buiatti et al., 2019) and with the two artifact removal pipelines for developmental data[1] (MADE (Debnath et al., 2020) and HAPPE (Gabard-Durnam et al., 2018)).

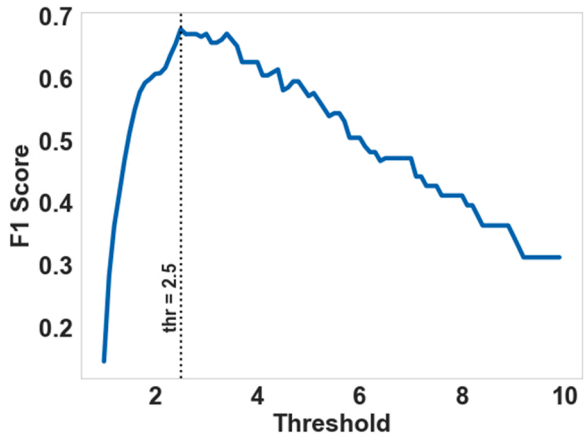


Fig. 6. Optimal threshold tuning for LOF using F1 Score as evaluation metric on the newborn Training Dataset. The highest F1 Score is obtained with a threshold of 2.5.

Removing artifacts with ASR (both processing modes) on the Test Dataset resulted in rejecting one subject because of too short clean segments for computing FTR. To ensure a fair comparison, we restricted the validation results to the remaining 9 subjects for all the considered methods.

For the visual response, standard processing resulted in a significant effect even with one less subject ($t(8) = 3.03$, $P = 0.016$) (Fig. 8, first row, left-hand panel). Compared to standard processing, ASR_R resulted in a somewhat lower power peak at the tag frequency accompanied by a similar decrease in the background power (Fig. 8, second row, left-hand panel), likely resulting from a more efficient noise reduction together with a slight power reduction. This minor difference impacted equivalently on the numerator and denominator of the FTR, obtaining a significant effect ($t(8) = 3.04$; $P = 0.016$) equivalent to the standard processing, and a response which is statistically indistinguishable from the standard mode (paired t-test between standard and ASR_R of the difference between power and background at the tag frequency across subjects: $t(8) = 0.034$, $P = 0.97$). The power spectrum resulting from ASR_C is further reduced, in particular at the tag frequency (Fig. 8, third row, left-hand panel), probably due to a slightly sub-optimal reconstruction of the steady-state response in bad segments. Nonetheless, the statistical effect is also significant ($t(8) = 2.60$, $P = 0.032$) and the response is only marginally lower than the standard mode (paired t-test as above: $t(8) = 1.91$, $P = 0.093$). On the contrary, the overall profile of the power spectrum resulting from MADE processing is notably higher than the one from the standard mode and with a much wider variance at low frequencies (<0.8 Hz), likely the effect of residual low-frequency artifacts (Fig. 8, fourth row, left-hand panel). Still, the visual response is statistically significant also in this case ($t(8) = 2.47$, $P = 0.039$), though marginally lower than the one obtained with ASR_R (paired t-test $t(8) = 1.95$, $P = 0.086$) and with standard correction (paired t-test $t(8) = 2.14$, $P = 0.065$). HAPPE (Fig. 8, fifth row, left-hand panel) also recovers a statistically significant peak of the visual response ($t(8) = 2.58$, $P = 0.033$) but it is significantly lower than for ASR_R ($t(8) = 3.04$, $P = 0.016$), ASR_C ($t(8) = 2.59$, $P = 0.032$) and standard processing ($t(8) = 3.01$, $P = 0.016$).

Validation on the facelike pattern response shows similar results. NEAR with ASR_R processing recovered a statistically significant effect ($t(8) = 2.79$, $P = 0.023$) and the facelike response was statistically equivalent to that obtained with the *standard* processing (paired t-test between standard and ASR_R of the difference between FTR for facelike and inverted facelike patterns across subjects: $t(8) = -0.38$, $P = 0.71$) (Fig. 8, first and second row, middle panel). Similar results are obtained with ASR_C: a significant facelike effect ($t(8) = 2.69$, $P = 0.027$), and no significant difference with standard mode ($t(8) = 1.67$, $P = 0.13$) (Fig. 8, third row, middle panel). However, MADE processing resulted in a shallower spectral peak (Fig. 8, fourth row, middle panel) and

[1] Neither MADE nor HAPPE are equipped with a method to detect flat channels. Since flat channels cause errors in the ICA classification algorithms used by these methods, we removed them before applying MADE and HAPPE to the data.

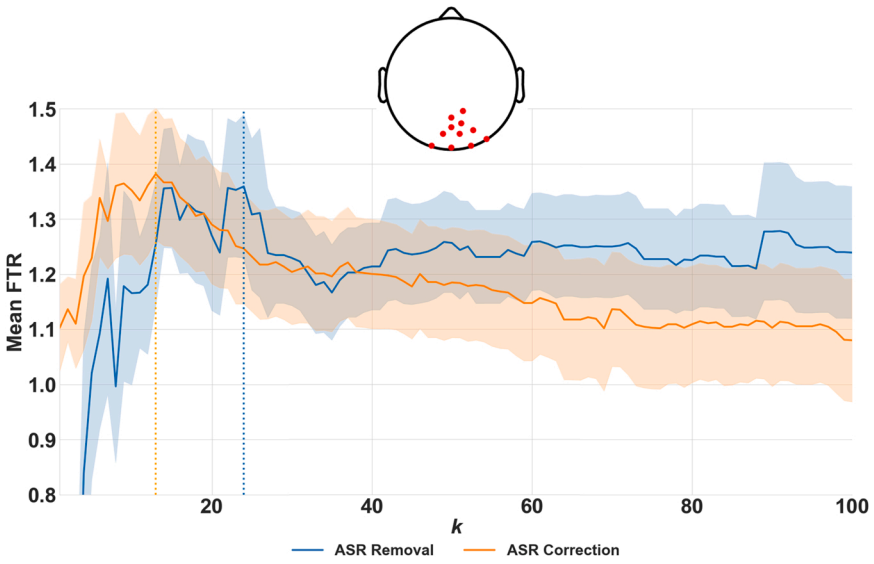


Fig. 7. A grid-search analysis to find the best ASR parameter settings on the newborn Training Dataset: Average visual response (FTR) on a predefined occipital cluster of electrodes (topography in top inset) as a function of ASR Parameter *k* and Processing Mode, computed on the Training Dataset (n = 11). The mean FTR is maximum at $k = 13$ for ASR Correction (ASR_C) and $k = 24$ for ASR Removal (ASR_R). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Performance of NEAR Bad Channel Detection tool compared to other methods (*standard* (Buiatti et al., 2019), *clean_rawdata* (CRD), HAPPE and FASTER). Top panel: Total number of detected channels for each subject. Bottom panel: Comparison of classification performance in matching standard bad channel detection (TP, FN, FP, TN, ACC indicate the number of True Positives, False Negatives, False Positives, True Negatives and Accuracy, respectively). NEAR shows the closest match with the *standard* bad channel detection score compared to other methods.

Bad Channel Detection										
Methods/Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Standard	1	2	3	0	1	7	2	14	7	13
CRD	5	8	8	6	0	13	6	22	9	9
HAPPE	9	8	15	8	18	13	6	2	6	6
FASTER	4	4	8	4	7	3	3	10	6	7
NEAR	2	2	3	1	1	5	1	23	8	14
Classification Metrics										
Methods	**TP**	**FN**	**FP**	**TN**	**ACC**			**F1-Score**		
CRD	3.3	1.7	5.6	113.4	94.11			0.47		
HAPPE	2.3	2.7	6.8	112.2	92.34			0.33		
FASTER	2.6	2.4	3	116	97.58			0.40		
NEAR	4.5	0.5	1.5	117.5	98.38			0.81		

recovered only a marginally significant facelike effect ($t(8) = 2.02$, $P = 0.078$), showing again a marginally significant difference compared to both ASR_R ($t(8) = 1.96$, $P = 0.085$) and standard processing ($t(8) = 1.88$, $P = 0.097$). HAPPE processing resulted in an even shallower peak (Fig. 8, fifth row, middle panel), failing to report a significant facelike effect ($t(8) = 1.21$, $P = 0.26$), although in this case the difference with NEAR methods is not significant (vs ASR_R: $t(8) = 1.59$, $P = 0.15$; vs ASR_C: $t(8) = 1.35$, $P = 0.21$), nor the one with standard *standard* processing ($t(8) = 1.98$, $P = 0.08$). These results are reflected in the single-subject responses (Fig. 8, right-hand column panel): While NEAR with both ASR_R and ASR_C recovered a preference for facelike patterns for all the subjects as with standard processing, two subjects following MADE and HAPPE processing showed an inverted effect.

3.3. Validation of NEAR on infant data

Although NEAR has been developed to remove artifacts in continuous newborn EEG data, here we show that NEAR is also efficient in removing artifacts from data of older infants and with event-related designs (thereby, proving its extensibility) by applying it to a 9-months-old EEG dataset recorded with an ERP paradigm (Parise and Csibra, 2012).

3.3.1. NEAR parameter calibration

Following the same procedure performed with newborns (Section 3.2.1), we calibrated LOF threshold for bad channel detection on the infant Training Dataset. By using F1 Score as the quality metric, the optimal LOF threshold obtained is 2, which is 0.5 lower than the one obtained on newborn data.

Likewise, the calibration of the ASR parameter *k* yielded an optimal value of $k = 21$ for ASR_R and $k = 3$ for ASR_C. Compared to what was obtained on newborns, this parameter is much lower for ASR_C than for ASR_R.

3.3.2. NEAR validation

3.3.2.1. NEAR bad channel detection. As done with the newborns data, we validated NEAR's bad channel detection method by evaluating its performance to match standard bad channel detection on the infant Test Dataset. For comparison, we also tested the performance of the three state-of-the-art methods: CRD, HAPPE and FASTER. As for newborns, NEAR's bad channel detection algorithm yielded the highest match with standard scoring: NEAR $F1 = 0.69$, HAPPE $F1 = 0.42$, FASTER $F1 = 0.33$, CRD $F1 = 0.33$.

3.3.2.2. Overall NEAR validation. Then, as for newborns, we validated the overall performance of NEAR pre-processing by direct comparison of

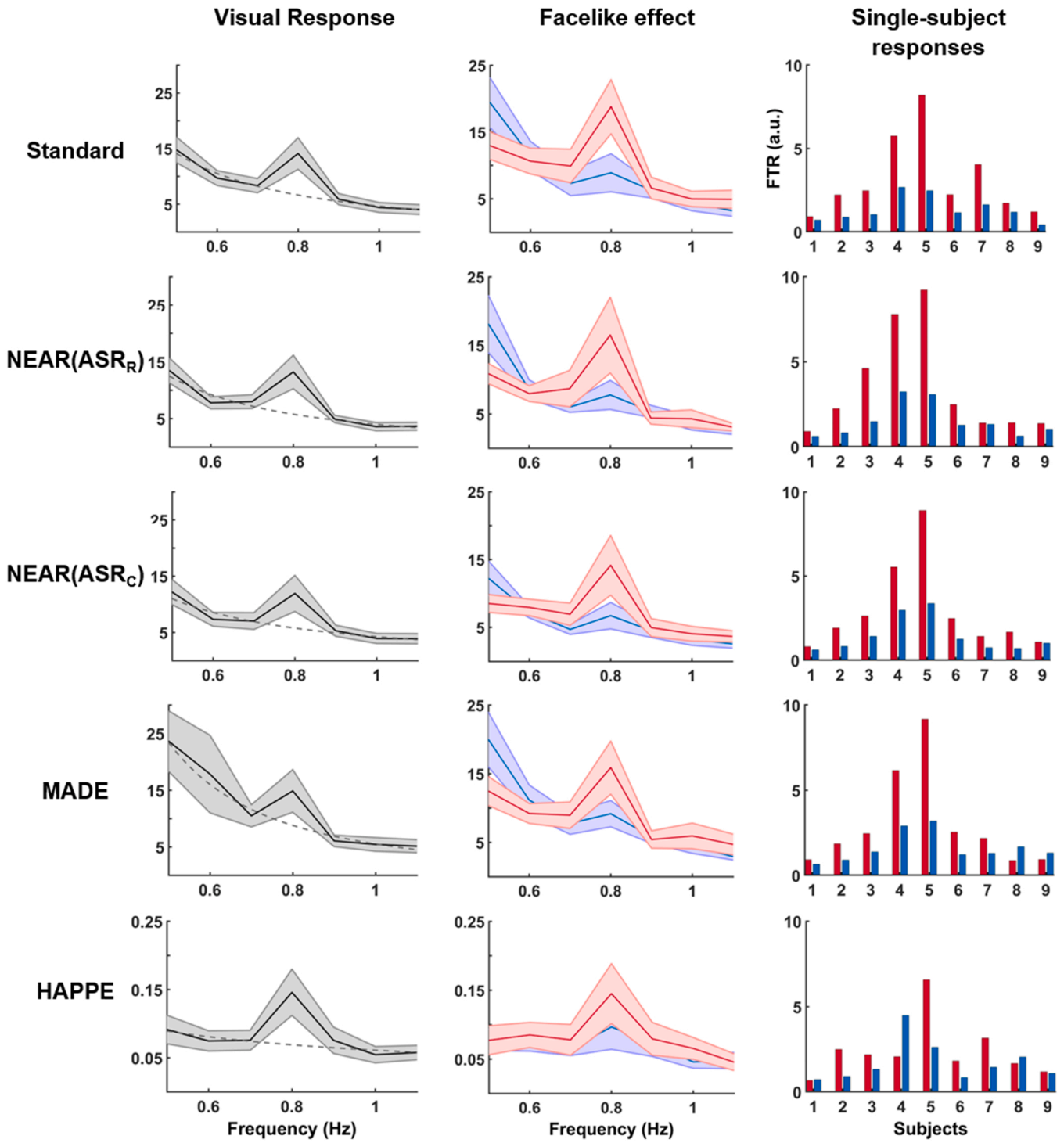


Fig. 8. Performance of NEAR in obtaining statistically significant neural responses from the raw newborn Test Dataset. Each row corresponds to an artifact removal method: *Standard* processing (Buiatti et al., 2019), NEAR using ASR_R, NEAR using ASR_C, MADE and HAPPE, respectively. Left-hand column: Power spectrum elicited by the overall visual stimulation. Shaded contour indicates the s.e.m. across subjects. The spectral peak at the tag frequency is statistically significant for all the methods, but NEAR using ASR_R obtains the highest t-value. Mid column: Power spectrum associated with upright (red line) and inverted (blue line) facelike stimuli. While the facelike effect is statistically significant for both NEAR processing modes, it is only marginally significant after MADE processing and not significant after HAPPE processing. Right-hand column: Single-subject FTR for upright (red bars) and inverted (blue bars) facelike images. The facelike effect is present in all subjects for NEAR processing, but not for MADE and HAPPE processing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the statistical significance of the main effect obtained by manual pre-processing in the original work (Parise and Csibra, 2012): a N400 differential response between incongruous and congruous conditions higher on the right region-of-interest than on the left one (where the regions of interest were identified by the electrodes between C3 and P3 and between C4 and P4, over the left and right hemisphere, respectively) (Fig. 9).

We also compared NEAR's performance with the state-of-the-art

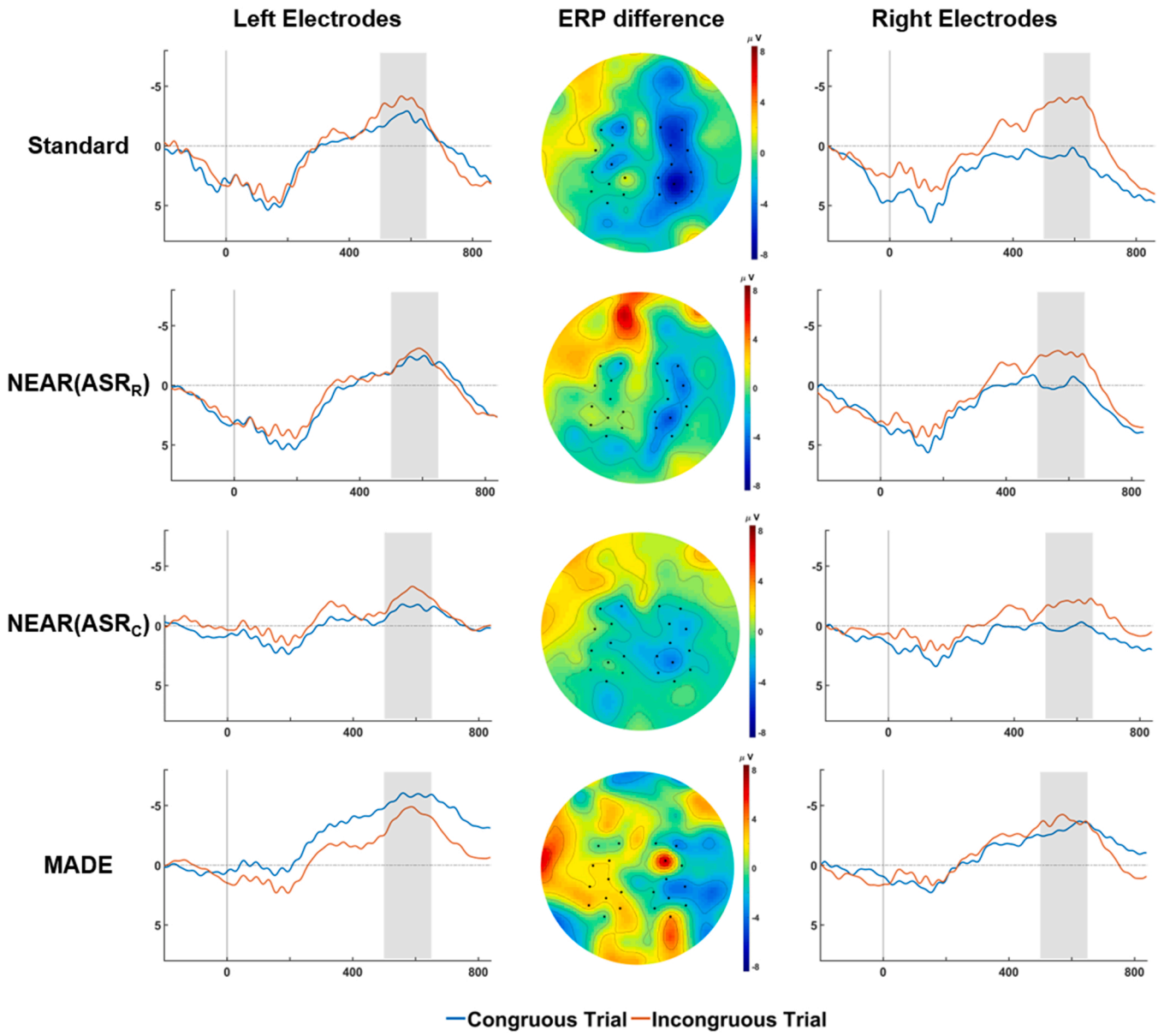


Fig. 9. Event-related potential (ERP) results for each of the processing modes: *Standard* processing (Parise and Csibra, 2012), NEAR using ASR_R, NEAR using ASR_C and MADE, respectively. The figure shows grand-average waveforms on congruous and incongruous trials in left (left-hand panels) and right (right-hand panels) regions of interest (marked by black points on the scalp maps). The gray shading indicates the time window of the infant N400 (500–650 ms), and the vertical line marks the time at which the object in each trial appeared from behind an occluder. The scalp maps (middle panels) depict the spatial distribution of the difference in ERP amplitude between incongruous and congruous trials in the given time window. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

artifact removal pipeline for developmental data MADE (Debnath et al., 2020).

Results show that the only method that recovered a significant ANOVA with factors condition and hemisphere is ASR_R ($F(1,13) = 5.13$, $P = 0.041$), while no significant effect is observed for ASR_C ($F(1,13) = 1.68$, $P = 0.22$), nor for MADE ($F(1,13) = 2.90$, $P = 0.11$). More specifically, NEAR using ASR_R yielded a clear congruency effect on the right hemisphere that was absent on the left hemisphere, similarly to *standard* processing (Fig. 9, first two rows). NEAR using ASR_C resulted in similar but shallower effects compared to ASR_R (Fig. 9, third row). MADE also exhibited a congruency difference between the hemisphere in the same direction, but with the congruous condition higher than for the other methods (Fig. 9, fourth row). However, no significant difference was found between the three methods on the size of the effect (paired t-test between the ERP difference between hemispheres of the difference between conditions of ASR_R vs MADE: $t(13) = 0.01$, $P = 0.99$; ASR_C vs MADE: $t(13) = -1.00$, $P = 0.33$; ASR_R vs ASR_C: $t(13) = 1.70$, $P = 0.11$).

4. Discussion

This paper presented NEAR, a pipeline that transforms artifacted raw developmental EEG data into clean data ready for downstream analysis. We demonstrated that NEAR's novel artifact removal procedure efficiently removes artifacts both from newborn and infant EEG data (high sensitivity), while preserving the EEG signal of neural origin (high specificity). NEAR will hopefully contribute to establish a more objective and reproducible preprocessing procedure within the developmental EEG community, a much-needed improvement considering the negative consequences of the variability of EEG data editing practices (Monroy et al., 2021). Hereafter we comment on some key aspects of

NEAR in the general context of EEG artifact removal.

4.1. An artifact removal method for non-stereotyped artifacts

The most problematic and predominant artifacts in newborn EEG data are non-stereotyped transient high-amplitude fluctuations involving variable sets of channels. By specifically simulating these artifacts, we showed that ASR processing included in NEAR is very efficient in detecting and removing them. On the other hand, ICA-based methods as MADE (Debnath et al., 2020) and (to a lesser extent) HAPPE (Gabard-Durnam et al., 2018) failed in processing these artifacts, most probably because they are developed to detect mainly the stereotyped ones. Notably, both MADE and HAPPE are more successful on newborn EEG data than on simulated data, possibly because real newborn EEG artifacts are more stereotyped (i.e. their spatial distribution and temporal profile are partly correlated across occurrences) than simulated ones, which are generated by random shuffling of the artifact topographical distribution. As discussed more extensively below, a combination of detection methods for non-stereotyped and stereotyped artifacts might be the solution to deal with the wide range of EEG artifacts, especially in developmental data.

4.2. ASR parameter calibration

One core tool used by NEAR is ASR (Kothe and Jung, 2016), an efficient algorithm that nevertheless depends on some user-defined parameters. The selection of these parameters is not univocal: the most systematic investigation on this issue (Chang et al., 2020) proposes that the optimal value of the ASR *k* parameter for adults lies "between 20 and 30", implicitly suggesting that it may be variable. Moreover, while ASR default processing mode is to correct the data from artifacts (a choice driven by the original aim of providing an efficient algorithm for real-time applications), the main developers of the EEGLAB software suggest removing the artifacted segments identified by ASR because the effects of ASR correction on the data "are not clearly understood" (https://eeglab.org/tutorials/06_RejectArtifacts/cleanrawdata.html, Retrieved June 7, 2021). Our study confirms that ASR performance significantly depends on the choice of both ASR Parameter *k* and processing mode (Fig. 7). The quality of developmental EEG data may vary substantially between different EEG setups, and different data analyses may require different thresholds. Therefore, we propose an adaptive approach to ASR: run ASR on a dataset previously collected with the same EEG setup and analyzed with the same analysis chosen for the current data and find the *k* and processing mode that best recover the EEG effects observed on that dataset. We provide a script for this calibration procedure and we recommend NEAR users to perform it before applying NEAR to newly recorded data. Once these parameters are identified, validation shown in this paper suggests that NEAR might be safely used in automatic mode. If a training dataset is not available, we recommend tuning NEAR parameters on at least a few subjects by measuring a well-known sensory response for both processing modes and *k* between 10 and 30. In any case, we strongly recommend that users keep monitoring the efficiency of NEAR (an easy task provided by the visualization tools along NEAR's pipeline and the report file), as unpredicted single-subject variations are always possible - automatic does not mean magic!

4.3. Artifact removal vs correction

Testing ASR on simulated data showed that removal mode is slightly more efficient than correction mode in cleaning the data from the effect of artifacts; results from the simulation suggest that while correction efficiently suppresses the artifacts, it also severely attenuates the underlying neural activity. This effect was consistently observed in the application of NEAR on both the newborn and infant data. This observation, together with the fact that the effects of ASR correction on the data are not fully understood yet, lead us to adopt EEGLAB recommendation to set the removal mode by default for off-line analysis, unless the performance of the correction mode on some training datasets shows significantly better results. In both processing modes, we recommend users to notice the amount of data being rejected (in case of ASR_R), or modified (in case of ASR_C) and the mean reduction of RMS variance in the processed signal. These values can be found in our report files. In particular for ASR_C, we recommend users to customize these values to set inclusion criteria for the subjects into the group-level analysis to avoid the risk of mostly relying on the reconstruction of heavily artefacted data.

4.4. Using NEAR on other experimental designs

NEAR has been trained and validated on a frequency-tagging paradigm by using a measure of the SSVEP and on an event-related design by using an ERP measure. The adaptive approach of NEAR provides a straightforward strategy to tune NEAR parameters to data recorded from other experimental designs that include event-related measures like time-frequency analysis or resting-state measures like (de)synchronization in specific frequency ranges or connectivity measures.

4.5. Combining NEAR with ICA for developmental EEG artifact removal

In comparison with NEAR, the pipeline for artifact removal of developmental data MADE performed moderately worse on newborn and infant data, mostly because it was not equally efficient in removing low-frequency artifacts. We see two possible reasons for this difference: 1) MADE's bad channel and bad segment identification tools were not calibrated to newborn EEG data; 2) As mentioned above, the benefit of removing artifacts by ICA is limited by the fact that most artifacts in newborn EEG data are non-stereotyped, therefore not easily captured by ICA. Nonetheless, ICA (and in particular Adjusted-ADJUST (Leach et al., 2020), the IC classifier developed for infant data and included in MADE) may be beneficial as a further processing step after NEAR because it could correct the data from residual stereotyped artifacts without any further data rejection. However, one issue that might be problematic for an efficient ICA decomposition of developmental EEG data with high-density systems is the very limited duration of the clean data segments. Rather than reducing the number of electrodes as in (Leach et al., 2020) (which would drastically decrease EEG spatial resolution, preventing a potential source reconstruction (Odabaee et al., 2014)), a possible solution would be to use PCA for dimensionality reduction. However, the application of PCA on EEG data has significant limitations (Artoni et al., 2018); therefore, investigations on alternative methods to run ICA on high-density EEG data of short duration would be very useful.

NEAR availability

NEAR is publicly available as open-source software at https://github.com/vpKumaravel/NEAR under GNU General Public License. An example anonymized dataset taken from the Test Dataset analyzed in the paper (data from (Buiatti et al., 2019)) has been deposited in the Open Science Framework and is freely available at https://osf.io/79mzg/. A step-by-step tutorial for the use of NEAR is included in the Appendix.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the three anonymous reviewers for their constructive comments and Barbara Pomiechowska for her help with data format exporting and sharing. This work was supported by the European Research Council Proof of Concept grant NeuroSoNew (842243).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.dcn.2022.101068.

References

Acunzo, D.J., MacKenzie, G., van Rossum, M.C.W., 2012. Systematic biases in early ERP and ERF components as a result of high-pass filtering. J. Neurosci. Methods 209 (1), 212–218. https://doi.org/10.1016/j.jneumeth.2012.06.011.

Artoni, F., Delorme, A., Makeig, S., 2018. Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition. NeuroImage 175, 176–187. https://doi.org/10.1016/j.neuroimage.2018.03.016.

Beauchemin, M., Gonzaíez-Frankenberger, B., Tremblay, J., Vannasing, P., Martınez-Montes, E., Belin, P., Beíand, R., Francoeur, D., Carceller, A.-M., Wallois, F., Lassonde, M., 2011. FEATURE ARTICLE mother and stranger: an electrophysiological study of voice processing in newborns. Cereb. Cortex August 21, 1705–1711. https://doi.org/10.1093/cercor/bhq242.

Blum, S., Mirkovic, B., Debener, S. , 2019. Evaluation of Riemannian ASR on cEEGrid data: an artifact correction method for BCIs, in: Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 3625–3630.

Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93–104. ⟨https://doi.org/10.1145/342009.335388⟩.

Buiatti, M., Di Giorgio, E., Piazza, M., Polloni, C., Menna, G., Taddei, F., Baldo, E., Vallortigara, G., 2019. Cortical route for facelike pattern processing in human newborns. Proc. Natl. Acad. Sci. USA 116 (10), 4625–4630. https://doi.org/10.1073/pnas.1812419116.

Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., Jung, T.-P., 2020. Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel EEG recordings. IEEE Trans. Biomed. Eng. 67 (4), 1114–1121. https://doi.org/10.1109/TBME.2019.2930186.

Dalianis, H., 2018. Evaluation Metrics and Evaluation BT - Clinical Text Mining: Secondary Use of Electronic Patient Records. In: Dalianis, H. (Ed.). Springer International Publishing, pp. 45–53. https://doi.org/10.1007/978-3-319-78503-5_6.

Debnath, R., Buzzell, G.A., Morales, S., Bowers, M.E., Leach, S.C., Fox, N.A., 2020. The Maryland analysis of developmental EEG (MADE) pipeline. Psychophysiology 57 (6), e13580. https://doi.org/10.1111/psyp.13580.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

Delorme, A., Miyakoshi, M., Jung, T.-P., Makeig, S., 2015. Grand average ERP-image plotting and statistics: a method for comparing variability in event-related single-trial EEG activities across subjects and conditions. J. Neurosci. Methods 250, 3–6. https://doi.org/10.1016/j.jneumeth.2014.10.003.

Fifer, W.P., Byrd, D.L., Kaku, M., Eigsti, I.-M., Isler, J.R., Grose-Fifer, J., Tarullo, A.R., Balsam, P.D., 2010. Newborn infants learn during sleep. Proc. Natl. Acad. Sci. 107 (22), 10320–10323. https://doi.org/10.1073/PNAS.1005061107.

Fix, E., Hodges, J.L., 1989. Discriminatory analysis. nonparametric discrimination: consistency properties. Int. Stat. Rev. / Rev. Int. De. Stat. 57 (3), 238–247. https://doi.org/10.2307/1403797.

Fransson, P., Metsäranta, M., Blennow, M., Åden, U., Lagercrantz, H., Vanhatalo, S., 2013. Early development of spatial patterns of power-law frequency scaling in fMRI resting-state and EEG data in the newborn brain. Cereb. Cortex 23 (3), 638–646. https://doi.org/10.1093/cercor/bhs047.

Gabard-Durnam, L.J., Mendez Leal, A.S., Wilkinson, C.L., Levin, A.R., 2018. The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): standardized processing software for developmental and high-artifact data. Front. Neurosci. 12, 97. ⟨https://www.frontiersin.org/article/10.3389/fnins.2018.00097⟩.

Georgieva, S., Lester, S., Noreika, V., Yilmaz, M.N., Wass, S., Leong, V., 2020. Toward the understanding of topographical and spectral signatures of infant movement artifacts in naturalistic EEG. Front. Neurosci. 14, 352. ⟨https://www.frontiersin.org/article/10.3389/fnins.2020.00352⟩.

de Heering, A., Rossion, B., 2015. Rapid categorization of natural face images in the infant right hemisphere. ELife 4, e06564. https://doi.org/10.7554/eLife.06564.

Kabdebon, C., Pena, M., Buiatti, M., Dehaene-Lambertz, G., 2015. Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. Brain Lang. 148, 25–36. https://doi.org/10.1016/j.bandl.2015.03.005.

Kothe, C.A. E., Jung, T. , 2016. Artifact removal techniques with signal reconstruction, Google Patents.

Krol, L.R., Pawlitzki, J., Lotte, F., Gramann, K., Zander, T.O., 2018. SEREEGA: simulating event-related EEG activity. J. Neurosci. Methods 309, 13–24. https://doi.org/10.1016/J.JNEUMETH.2018.08.001.

Kumaravel, V.P., Kartsch, V., Benatti, S., Vallortigara, G., Farella, E., Buiatti, M., 2021. Efficient artifact removal from low-density wearable EEG using artifacts subspace reconstruction, in: Proceedings of 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 333–336.

Leach, S.C., Morales, S., Bowers, M.E., Buzzell, G.A., Debnath, R., Beall, D., Fox, N.A., 2020. Adjusting ADJUST: optimizing the ADJUST algorithm for pediatric data using geodesic nets. Psychophysiology. https://doi.org/10.1111/psyp.13566.

Luck, S.J., Stewart, A.X., Simmons, A.M., Rhemtulla, M., 2021. Standardized measurement error: a universal metric of data quality for averaged event-related potentials. Psychophysiology, e13793. https://doi.org/10.31234/osf.io/jc3sd.

Mognon, A., Jovicich, J., Bruzzone, L., Buiatti, M., 2011. ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features. Psychophysiology. https://doi.org/10.1111/j.1469-8986.2010.01061.x.

Monroy, C., Domínguez-Martínez, E., Taylor, B., Portolés, O.M., Parise, E., Reid, V.M., 2021. Understanding the causes and consequences of variability in infant ERP editing practices. Dev. Psychobiol. https://doi.org/10.1002/dev.22217.

Mullen, T.R., Kothe, C.A.E., Chi, Y.M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.P., Cauwenberghs, G., 2015. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. IEEE Trans. Biomed. Eng. 62 (11) https://doi.org/10.1109/TBME.2015.2481482.

Nolan, H., Whelan, R., Reilly, R.B., 2010. FASTER: fully automated statistical thresholding for EEG artifact rejection. J. Neurosci. Methods 192 (1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015.

Norcia, A.M., Appelbaum, L.G., Ales, J.M., Cottereau, B.R., Rossion, B., 2015. The steady-state visual evoked potential in vision research: a review. J. Vis. 15 (6), 4. https://doi.org/10.1167/15.6.4.

Odabaee, M., Tokariev, A., Layeghy, S., Mesbah, M., Colditz, P.B., Ramon, C., Vanhatalo, S., 2014. Neonatal EEG at scalp is focal and implies high skull conductivity in realistic neonatal head models. NeuroImage 96, 73–80. https://doi.org/10.1016/j.neuroimage.2014.04.007.

Onton, J., Westerfield, M., Townsend, J., Makeig, S., 2006. Imaging human EEG dynamics using independent component analysis. In: Neuroscience and Biobehavioral Reviews, vol. 30. Pergamon, pp. 808–822. https://doi.org/10.1016/j.neubiorev.2006.06.007.

Parise, E., Csibra, G., 2012. Electrophysiological evidence for the understanding of maternal speech by 9-month-old infants. Psychol. Sci. 23 (7), 728–733. https://doi.org/10.1177/0956797612438734.

Pion-Tonachini, L., Kreutz-Delgado, K., Makeig, S., 2019. ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. NeuroImage 198, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026.

Ronga, I., Galigani, M., Bruno, V., Noel, J.-P., Gazzin, A., Perathoner, C., Serino, A., Garbarini, F., 2021. Spatial tuning of electrophysiological responses to multisensory stimuli reveals a primitive coding of the body boundaries in newborns. Proc. Natl. Acad. Sci. USA 118 (12). https://doi.org/10.1073/pnas.2024548118.

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., Tangermann, M., 2014. Robust artifactual independent component classification for BCI practitioners. J. Neural Eng. 11 (3), 035013 https://doi.org/10.1088/1741-2560/11/3/035013.

Zhu, Q., Feng, J., Huang, J., 2016. Natural neighbor: a self-adaptive neighborhood method without parameter K. Pattern Recognit. Lett. 80, 30–36. https://doi.org/10.1016/j.patrec.2016.05.007.