An Efficient and Robust Approach to Detect Auditory Evoked Brainstem Responses by Progressive Sub-average Cross-covariance Analysis

Haoyu Wang[1-4#], Bei Li[1#], Xu Ding[2-3#], Xueling Wang[1-3], Zhiwu Huang[1-3], Yunfeng Hua[1-4]*, Hao Wu[1-4]*

[1]Department of Otolaryngology-Head and Neck Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]Ear Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3]Shanghai Key Laboratory of Translational Medicine on Ear and Nose Diseases, Shanghai, China
[4]Shanghai Institute of Precision Medicine, Shanghai Ninth People's Hospital, Shanghai, China
[#] These authors contributed equally to this work
* Corresponding authors. E-mail: yunfeng.hua@outlook.com (Y.H.) and haowu@sh-jei.org (H.W.)

Abstract

Objective: Auditory brainstem response (ABR) is widely employed to assess auditory function of humans and lab animals. Despite attempts to automation, ABR threshold determination still relies on human visual inspection on average responses. The aim of this study is to develop a robust procedure for automatic and accurate threshold determination.

Design: Signal lag of maximal cross-covariance between sub-averaged sweeps was utilized to test the existence of time-locked components during progressive averaging by iterations. The minimal number of iterations required to reach the critical lag value versus sound level series was modeled to inform the estimation of ABR threshold.

Study Sample: Test datasets containing single ABR sweeps were acquired from eight mice and four human participants.

Results: To reach the critical lag value, weak responses at near-threshold levels

required more sweep averaging than strong responses at supra-threshold levels, allowing precise and robust determination of the ABR threshold. Moreover, by contrast to level averaging over fixed sweep number the algorithm saved up to 69 % and 36 % sweeps at supra-threshold sound levels for the threshold determination in mice and humans, respectively.

Conclusion: Implementation of this method in commercial devices will automate the ABR test in a more reliable and efficient fashion.

Introduction

The auditory brainstem responses (ABRs) are brain electrical potential changes due to synchronous neuronal activities evoked by supra-threshold acoustic stimuli (Jewett et al., 1970). These responses are detectable using non-invasive surface electrodes placed on the scalp of test subject. The ABR test is widely employed to assess the hearing function, particularly for infants, adults with learning disabilities and patients undergoing operation, to whom a test through communication or body movements is not applicable (Jacobson et al., 1990). Typical ABR waveform is composed of five peaks in the early onset of sound evoked potentials, corresponding to synchronous activities arising from auditory nerve, cochlear nucleus, superior olivary complex, lateral lemniscus and inferior colliculus, respectively. Thereby, features of the ABR waveform including peak latencies and amplitudes provide clinical-significant information, for instance hidden hearing loss (Mehraei et al., 2016; Ridley et al., 2018), tinnitus (Bramhall et al., 2018; Castaneda et al., 2019), site of lesions or tumors in the auditory system (Roeser et al., 2007) based on the way how the waveform is altered.

Although the ABR test itself is considered objective, the waveform recognition at near-threshold level is not always trivial. Currently, professionals are still required to supervise recording and interpret results, which account for the largest cost component of the test. Moreover, as accurate waveform recognition highly depends on the skill and the experience of the human assessors, often limited consensus occurs, especially when untypical waveform or high background noise are encountered

(Vidler and Parkert, 2004). As diagnosis of progressive and acquired hearing loss in preschool children (Lu et al., 2014; Lu et al., 2011) and categorizing temporal threshold shifts caused by noise exposure (Le Prell, 2019) require comprehensive hearing screening with high accuracy and robustness, there is an increasing demand for automatic approaches making the ABR test less labor-intensive.

The fundamental challenge of automated ABR analysis is high-level noise contamination. In order to improve the signal-to-noise ratio (SNR) for unambiguous waveform recognition, averaging over hundreds of sweeps is exclusively required. Hence, ABR results are usually provided with statistical measures like well-known single- and multiple-point F-ratio (F_{sp} and F_{mp}) for evaluating the signal and noise characteristics (Don and Elberling, 1994; Elberling and Don, 1984). Over decades, several methods were proposed to automate the procedure. Feature-based strategies detect responses by quantifying the similarity of measured waveforms to either existing templates (Valderrama et al., 2014) or waveforms evoked by adjacent sound levels (Suthakar and Liberman, 2019), as well as by features learned by artificial neural network from human-annotated datasets (Alpsan and Ozdamar, 1991; McKearney and MacKinnon, 2019). However, the detection accuracy is often limited owing to waveform heterogeneity, varying data quality, and inconsistent training data (McKearney and MacKinnon, 2019). Alternatively, statistics-based strategies quantify either inter-sweep variability using correlation function (Bershad and Rockmore, 1974; Weber and Fletcher, 1980; Xu et al., 1995) or the 'signal quality' through scoring procedures like F-ratios (Don and Elberling, 1994; Elberling and Don, 1984), but

reliable response thresholding only is only obtained under a narrow range of experiment conditions. Although some approaches mentioned above have already been integrated in commercial devices, nowadays reliable threshold determination still involves human supervision due to frequent detection errors.

In this study, we first tried to calibrate the single-sweep covariance measurement for ABR waveform detection and observed large inter-trial variability of both pairwise correlation coefficient and F_{sp}-ratio at the threshold level, limiting the reliability and accuracy of this approach. Next, we bypassed this limitation by proposing a novel procedure, in which time-locked ABR was detected with fixed cross-covariance criterion from progressively averaged responses by iterations and the executed iteration counts versus level function was modeled to interpolate for the threshold level beyond the sampling resolution of level series. Mouse and human ABR results suggested that the algorithm outcomes matched the expert assessments with a maximal aberration of 5 dB. Besides, in our hands the algorithm required up to 69 % and 36 % fewer sweeps for threshold determination in mice and humans, respectively.

Materials and methods

Animals, human participants and ethics

C57BL/6 mice were purchased from Sino-British SIPPR/BK Lab Animal Ltd. (Shanghai, China). The telomerase-knock-out mice were kindly donated by Prof. Lin Liu (Nankai University, China) and bred in house. Human participants were recruited from Shanghai Ninth People's Hospital and consent forms were signed before the experiment. This study is conduct at the Ear Institute and the Hearing and Speech Center of the hospital. All procedures were reviewed and approved by the Institutional Authority for Laboratory Animal Care (HKDL2018503) and the Hospital Ethics Committee for Medical Research (SH9H-2019-T79-1).

ABR recording

Mouse ABR sweeps were recorded using TDT RZ6/BioSigRZ system (Tuck-Davis Tech. Inc., US) in a sound-proof chamber. 7-week-old mice were anesthetized through intraperitoneal injection of hydrate chloride (500 mg/kg). During the recording, animal body temperature was maintained at 37 °C using a regulated heating pad (Harvard Apparatus, US) with rectal thermal probe. Electrical potentials were registered via subdermal needle electrodes (Rochester Electro-Med. Inc., US) placed at animal vertex (record electrode), left infra-auricular mastoid (reference electrode) and right shoulder region (ground electrode). 3-millisecond tone pips at 16 kHz were delivered via a multi-field magnetic speaker (Tuck-Davis Tech. Inc., US) positioned at 10 cm from the animal vertex. Evoked potentials upon repeated stimuli (20 Hz)

were sampled at 24 kHz. 500 sweeps were acquired at test sound level series starting from 90 to 0 dB with 5-dB decrement. For one animal (**Suppl. Figure 2**), we acquired potentials starting from +10 to –10 dB around predetermined threshold level with 1-dB decrement.

Human ABRs were recorded using commercial ABR device (Intelligent Hearing Systems, US) with Smart EP software from four volunteers aged 21-29 years without preselection based on their medical conditions. Click sound stimulation (100 μs duration, rectangular envelopes) was generated and presented monaurally through ER3 insert earphones with foam tips at a rate of 37.1/s with alternating polarity. Electrode impedance was less than 5 kΩ and interelectrode impedance was within ±1 kΩ. Artifact rejection level was < 31% (rejection voltage 31 μV) to reduce response contamination of high EEG levels and myogenic activities. The recorded potentials were sampled at 40 kHz and amplified by a factor of 100,000. 100 Hz (high-pass) and 3000 Hz (low-pass) filters were applied. Average responses over 500, 1000, and 2000 sweeps were acquired three times at level series starting from 60 to 0 dB with 5-dB step size.

Single-sweep covariance analysis

Mouse ABR sweeps were analyzed using self-written routines in MATLAB (MathWorks Inc., US). The rejection amplitude of artifacts was set to 55 μV and a high-pass filter (smoothing splines with a smoothing parameter of 0.5) was applied to the raw recordings. The pairwise correlation coefficients were computed using

MATLAB function from 350 sweeps within a temporal window of 0-6 ms post-stimulation onset, resulting in 61,075 coefficients (350×349/2 sweep pairs) at each sound level. Histograms of the coefficients were fitted with shifted normal distributions.

Progressive sub-average cross-covariance analysis

For mouse ABR, 350 sweeps were loaded progressively by seven iterations (50 sweeps per executed iteration). In each iteration, sweeps were randomly subdivided into two groups for separated averaging. Cross-correlation between sub-averages was computed to measure the signal lag of the maximal correlation-coefficient. Time-locked responses were detected based on the criterion whether the measured lag was fewer than 1 data point (equivalent to 1% of the analyzed record length, 1-6 ms post-stimulation onset). In order to reject cases of coincidently overlapped noise peaks with similar latency (false positives), three parallel runs were performed with regrouped sweeps. The executed iterations were counted to obtain the minimal sweep number required for detectable responses. The iteration upper limit was set to seven for total 350 recorded sweeps. Sigmoid and exponential function were employed to model the development of the normalized iteration counts (*NC*) with the presentation level (*S*). For modeling with sigmoid function (1), the best fit-parameter $\alpha = 0.6$. At function growth of 0.9 interpolation was perform to determine the threshold level (*T*). For modeling with exponential function (2), the best fit-parameter $\alpha' = 0.25$ and the function growth of 1 was used to interpolate for the threshold level.

$$NC_S = \frac{1}{1+e^{-\alpha(S-T)}} \qquad (1)$$

$$NC_S = e^{-\alpha'(S-T)} \qquad (2)$$

As for human ABR, average responses over 500, 1000 and 2000 sweeps were used as inputs for the algorithm variant (**Suppl. Figure 4A**). In each iteration three cross-correlation runs were performed with the combination of two out of three sub-averages which were updated with 500 more sweeps. The iteration upper limit was set to seven for maximum average over 3500 sweeps. Note that to generate average responses over 1500, 2500, 3000 and 3500 sweeps weighted averaging (3) was used, in which *avg*(M), *avg* (N) and *avg*(M+N) referred to averages over M, N and M plus N sweeps, respectively.

$$avg(M+N) = \frac{M \bullet avg(M) + N \bullet avg(N)}{M+N} \qquad (3)$$

The lag criterion for the time-locked component was fewer than 7 data points (equivalent to 2 % of the analyzed record length, 0-10 ms post-stimulation onset). Same as for the mouse ABR, 0.9 was used to interpolate for the threshold level on the best-fit sigmoid growth.

Human assessment of ABR threshold

Human ground truth of the ABR threshold was obtained from visual inspection of waveforms at level series by five clinicians independently. Each of them was provided with average responses either over fixed sweep number (the conventional averaging) or over those used in the algorithm (the algorithm averaging). They were supposed to determine the lowest level with the presence of ABR waveform. They were blinded to

the test subject identities and the assessments from others. Means of the determined thresholds were used as ground truth to compare with algorithm outcomes.

Results

ABR Thresholding by covariance analysis of single-sweeps

ABRs are usually embedded in high-level background activities and system noise, and thereby smooth baseline and clear waveform, if present, require averaging over hundreds of sweeps. To generate test datasets, we recorded from mice single-sweeps upon 16 kHz tone-pips at different sound levels (**Figure 1A**). After exclusion of those with movement-induced excessive amplitudes, more than 350 sweeps were obtained and filtered sequentially through a high-pass filter (Nishida et al., 1993) to remove low frequency fluctuations in the short-latency component (**Suppl. Figure 1**). At each presentation level, pairwise correlation-coefficients (CCs, a statistical measure of waveform similarity) were sampled. Positively shifted distributions were obtained with increased sound levels, indicating elevated degrees of correlation (**Figure 1B**). Together with prior study (Galbraith, 2001), this result suggests that single-sweeps of supra-threshold ABRs are highly correlated.

Next, we attempted to calibrate the CC criterion for reliable ABR thresholding. Recordings were performed at level series starting from 90 to 0 dB on three wild-type (wt) of normal hearing and five telomerase-knock-out mice (terc$^{-/-}$) with high risk of deafness. Example ABR runs were illustrated in **Figure 2A** and the medians (peak positions) of pairwise CCs were plotted as a function of the level series (**Figure 2B**). Shifted rising phases of about 30 dB were observed when compared the CC median – level function of the terc$^{-/-}$ animal to that of the wt animal, consistent with the elevated threshold of the terc$^{-/-}$ animal. Together with early theoretical work (Bershad and

Rockmore, 1974), this result confirmed CC as a promising measure for detecting supra-threshold responses. In order to obtain the boundary value of CC for supra-threshold responses, we plotted the CC medians as a function of levels relative to the human-determined thresholds (**Figure 2C**). Large inter-trial variability (0.0119 ± 0.0034, mean ± s.d.), which was presumably caused varying SNRs of recordings under different experimental conditions (such as electrode position, depth of anesthesia, etc.), was observed and compromised a reliable ABR thresholding via this approach. Similarly, when F_{sp}-ratio was employed in the replacement of CC (**Figure 2D** & **E**), large variation (3.09 ± 1.53, mean ± s.d.) persisted in the measured F_{sp}-rations at the threshold level.

Threshold detection using cross-correlation test during ongoing averaging

It had been shown above that both CC and F_{sp}-ratio measured at the threshold level differed from trial to trial, presumably due to varying SNRs of recordings. Within each trial, however, constant noise level is expected and weak response thereby requires more averaging than strong response to reach the same SNR level. Based on this idea, we designed a novel procedure for threshold determination by modeling the minimal required sweep number for an average response to reach stationary SNR criterion as a function of level series (**Figure 3A**). In detail, single-sweeps were randomly divided into two groups **(Figure 3A – (I))** and the signal lag for maximal peak overlap between group averages (**Figure 3B** for averages of all sweeps; **Figure 3C** for sub-averages) was searched using cross-correlation function **(Figure 3A – (II))**.

For time-locked ABRs, irrespective of waveform and amplitude, a neglectable lag (here the lag criterion k < 1 data point) was expected **(Figure 3A – (III),** see **Figure 3D** for measured lags between sub-averages). To reject false positives caused by coincidently overlapped noise peaks with similar latency, three parallel runs were performed with regrouping. Upon detection of time-locked response, the procedure was then repeated with sweeps recorded at lower sound level; when time-locked response was under detected or absent, additional sweeps were added iteratively to increase total number for the random-split averaging until the iteration upper limit (I_{max} = 7) was reached **(Figure 3A – (IV))**. As illustrated in **Figure 3E**, we found that the counts of executed iterations increased rapidly when approaching to the threshold level. That was expected, because to reach the lag criterion weak responses of small amplitudes required more sweep averaging to suppress noise peaks. Here the iteration upper limit was crucial to avoid infinity iterations, as time-locked responses were absent from the averages at sub-threshold levels.

So far, we showed that this procedure was able to determine the lowest supra-threshold level (last level with iteration counts fewer than upper limit) and highest sub-threshold level (first level with reached iteration upper limit). Further attempt was made to interpolate for the threshold beyond the step size of sampled levels (**Figure 3F**). From one animal, we acquired single-sweeps with level series of 1-dB- instead of 5-dB step size. By fitting the rising phase of the normalized iteration counts with different functions, the threshold was found at 1 on the exponential growth and 0.9 on the sigmoid growth **(Suppl. Figure 2)**. In our hands, fitting with

sigmoid function was more reliable than that of exponential function, especially when limited data points were present at the rising phase. Thus, we used 0.9 on the sigmoid growth for threshold interpolation with 5-dB-spaced level series. In addition, this procedure was not limited to the detection method of ABR waveform. Similar results were obtained when measured mean F_{sp}-ratio at the threshold **(**F_{sp}-threshold = 3.09, **Figure 2E)** was utilized in the replacement of the lag criterion for response detection (**Suppl. Figure 3**).

Accuracy and efficiency of the new method for threshold determination

To evaluate the performance of the new method, we compared the algorithm-determined thresholds with human ground-truth on both mouse and human ABR. Note that for human ABR, sub-averages were separately acquired instead of generated by random-split averaging, as single-sweep registration mode was not applicable on the commercial devices we used (see **Suppl. Figure 4A** for the modified flowchart and **Materials and Methods** for details). Close matches were found between the algorithm-determined thresholds and the readouts of five independent clinicians (2.5 ± 1.6 dB for mouse ABR and 0.75 ± 0.95 for human ABR, mean ± s.d., **Figure 4A** & **Suppl. Figure 5A**), suggesting a detection accuracy of near-human-performance. Besides, upon detected responses ongoing averaging was stopped by the algorithm at supra-threshold levels and fewer sweeps in a response-amplitude-dependent manner did not appear to influence the waveform recognition by human (**Figure 4B** & **Suppl. Figure 5B**). When compared to

averaging with fixed sweep number, the new method required 69.39 ± 17.45 % and 36.19 ± 17.53 % (mean ± s.d.) fewer sweeps in mouse and human ABR, respectively, without compromising the detection accuracy (**Figure. 4C** & **Suppl. Figure 5C**).

Discussion

Since decades attempts have been made to automate ABR analysis. Several techniques were proposed including correlation function for detecting time-locked waveform in random fluctuations. In this study, we first reinvestigated the pairwise CCs of single-sweeps as a function of level series: (1) constant peak broadening of CC distributions across sound levels was observed (**Figure 1B**). Considering the broadening as a consequence of random correlation and anti-correlation of noise peaks with evoked responses, this result implied high fidelity of auditory transmission and stable noise level for single experiment; (2) within single recording session, CC peak position increased monotonically with elevated test sound level **(Figure 2B)**. Often slow increase or even decline in CCs was observed at 30 dB above the threshold level. The reason was unclear and might suggest none-linear change of response peak amplitudes at those levels; (3) the obtained CCs at the threshold level showed large trial-to-trial variations (**Figure 2C**). Together with similar result of quantification using F_{sp}-ratio **(Figure 2D & E)**, we concluded that this approach is sensitive to experimental settings (for instance electrode position, anesthesia depth, etc.) and thereby may not be suitable for reliable response detection across multiple test subjects or recording sessions.

To bypass this limitation, we proposed a novel strategy. While keeping a stationary waveform detection criterion, we asked the minimal sweep number for an average response to reach that criterion and in turn the threshold was informed by development of this number with level series. By contrast to prior studies, in which

signal lag of maximal correlation was used to detect auditory evoked responses in two consecutive recordings (Xu et al., 1995), here the lag criterion was applied to random-split averaged responses, so that multiple test runs were possible by regrouping. At near-threshold levels, we observed a fast increase of the minimal sweep number (reported as the executed iteration counts in the algorithm) for sub-averaged responses to reach the lag criterion (**Figure 3F**). Since varying noise level only contributes to the baseline sweep numbers, the location of the characteristic fast rising is supposed to be threshold-specific and insensitive in different experimental trials. Besides, neither the selection of the lag criterion nor maximal iteration number played critical role for the threshold determination, because the measured lags were much larger at sub-threshold than supra-threshold level (**Figure 3D**) and shifting iteration upper limit caused only little change of the corresponding level due to the exponential-like growth (**Figure 3E**, **Suppl. Figure 2**). These features made our approach robust and almost calibration-free. Next, we showed that precise threshold determination was possible by modeling and interpolation beyond the 5-dB step size of sampled levels (**Figure 3F**), in our case up to 1 dB. Further development of this method is to combine with sampling strategy of progressively reduced level step size and increased sweep averaging per iteration at near-threshold sound levels in order to estimate the model in a more precise and efficient way (Cebulla and Sturzebecher, 2015).

For both mouse and human ABR, the new method was proven reliable in threshold determination and required fewer sweeps at supra-threshold levels. This

selective reduction did not cause increased difficulty in waveform assessment by human (**Figure 4B** & **Suppl. Figure 5B**), as SNR of responses at near-threshold is most crucial for the threshold determination. Such feature is attractive in two respects. First, it provides minimal quality control for unambiguous waveform recognition for both human and machine. Further, standardized data will benefit machine-learning-based approaches by minimizing discrepancy in training data (McKearney and MacKinnon, 2019). Second, when to stop averaging is an important decision to make during ABR recording, the new method not only makes the ABR test more efficient by avoiding prolonged acquisition and redundant data, but also specifies the confidence level of waveform interpretation (Don and Elberling, 1996; Madsen et al., 2018). Based on our quantification, up to 69 % sweeps in mouse ABR and 36 % in human ABR were saved when ongoing averaging was stopping upon detected responses (**Figure 4C** & **Suppl. Figure 5C**), implying the potentials of bursting the regular ABR test in speed and efficiency.

Acknowledgements

We would like to thank Yun Li, Kun Han, Yan Ren, Lu Yang and Haifeng Li from Shanghai Ninth People's Hospital for help with ABR assessment. We would like to thank Drs. Guangming Chen and Lin Liu for providing terc$^{-/-}$ mice. This study was supported by The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (QD2018015 to Y.H.), Elite Program at Shanghai Ninth People's Hospital (JY201802 to Y.H., Z.H.), The National Science Foundation for Young Scientists of China (81800901 to Y.H., 81700903 to B.L) and SHIPM-mu Fund (JC201808 to B.L.).

Declarations of interest

The authors have declared that no conflicts of interest exist.

Author contribution

Y.H., H.W. conceptualized and supervised the study. B.L., X.D. acquired the ABR data; Y.H., H.Y.W. designed the algorithm; Y.H., H.Y.W., X.W., Z.H. analyzed and interpreted the data. Y.H., H.Y.W. wrote the manuscript with help of B.L. and X.D.; All authors commented on the manuscript.

Figure Legends

Figure 1. Pairwise CCs of single-sweep mouse ABRs. **(A)** Example sweeps recorded at four different sound levels (grey lines). Average responses (colored lines) were

obtained from 350 sweeps. **(B)** Distributions of pairwise CCs from (A) were fitted with normal distributions (colored lines, mean full-width-at-haft-maximum was 0.1778 ± 0.0154, mean ± s.d.) with varying peak positions.

Figure 2. Quantification of pairwise CC and F_{sp}-ratio at the threshold level. **(A)** Example average responses of 350 sweeps recorded from a wt animal (black lines) and a terc$^{-/-}$ animal (red lines). Attenuated ABR amplitudes were observed: > 25 dB for the wt animal (pointed by black asterisk) and > 55 dB for the terc$^{-/-}$ animal (pointed by double red asterisks). **(B)** Plot of the pairwise CC medians vs. the level series. A shifted rising of CCs by about 30 dB in level was observed between the wt (black dots) and the terc$^{-/-}$ animal (red cycles). **(C)** Plot of pairwise CC medians vs. levels relative to the threshold. Datasets of three wt (black dots) and five terc$^{-/-}$ animal (red cycles) were centered to the human-determined thresholds (pointed by asterisks). Mean of CC medians at the threshold (level = 0) was 0.0119 ± 0.0034 (mean ± s.d., varying from 0.0090 to 0.0199). **(D)** Similar to (B), plot of F_{sp}-ratios vs. the level series. **(E)** Similar to (C), plot of F_{sp}-ratios vs. levels relative to the threshold. Mean of F_{sp}-ratios at the threshold was 3.09 ± 1.53 (mean ± s.d.).

Figure 3. Progressive sub-average cross-covariance test of mouse ABR. **(A)** Flowchart of the algorithm, in which argA, argB, argC referred to the subgroup averages (I), xcorr the cross-correlation operation (II), k the lag criterion for time-locked responses (here k = 1 data point, 1% of record length, III), r1, r2, r3 the

outcomes of three parallel runs, as well as I_{max} the iteration upper limit (I_{max} = 7 for total 350 sweeps, IV). **(B)** Example average responses of 350 single-sweeps recorded from mouse. The human-assessed threshold was 30 dB (black asterisk). **(C)** Averaged responses (black and grey lines) of random-split sweeps at the level series. **(D)** Mean lags for maximal overlap between sub-averages from three parallel cross-correlation runs were plotted as a function of the level series. At supra-threshold levels (dots) small lags were obtained, while large amplitude and significant inter-trial variability (31.11 ± 16.12 data points, mean ± s.d.) were observed at sub-threshold levels (cycles). **(E)** Increased number of sweeps (50 sweeps per iteration) were used for random-split averaging, until the lag criterion (k < 1 data point) was fulfilled. Executed iteration counts (dots) were plotted as a function of the level series. If the iteration upper limit (I_{max}, dash red line) was reached at two consecutive levels (cycles), the time-locked response was considered absent from the average and the algorithm was aborted to stop testing at lower levels (triangles). **(F)** The normalized iteration counts vs. level function was fitted by a sigmoid function and 0.9 on the growth was used to interpolate for the ABR threshold between adjacent levels (see **Suppl. Figure 2** for the calibration).

Figure 4. Performance of the new method on mouse ABR. **(A** & **B)** Average responses (n = 8) were generated by averaging total 350 sweeps (A, conventional averaging) or those used in the executed iterations (B, algorithm averaging). The mean thresholds determined independently by five clinicians were consistent with the

algorithm outcomes (aberration δ = 2.50 ± 1.60 dB in A and 2.13 ± 2.47 dB in B, mean ± s.d.). Linear fit: adjust R^2 = 0.9587 in A and 0.9713 in B. **(C)** Comparison between the sweep number used in the conventional averaging (left bar) and the algorithm averaging (right bar). Note that the sweeps were counted at all supra-threshold and two highest sub-threshold level for both averaging modes. The algorithm averaging used 69.39 ± 17.45 % (mean ± s.d.) fewer sweeps than the conventional averaging.

References

Alpsan, D., and Ozdamar, O. (1991). Brain-Stem Auditory Evoked-Potential Classification by Backpropagation Networks. Ieee Ijcnn, 1266-1271.

Bershad, N.J., and Rockmore, A.J. (1974). On estimating signal-to-noise ratio using the sample correlation coefficient. IEEE Trans Inf Theory *IT-20*, 112-113.

Bramhall, N.F., Konrad-Martin, D., and McMillan, G.P. (2018). Tinnitus and Auditory Perception After a History of Noise Exposure: Relationship to Auditory Brainstem Response Measures. Ear Hear *39*, 881-894.

Castaneda, R., Natarajan, S., Yule Jeong, S., Na Hong, B., and Ho Kang, T. (2019). Electrophysiological changes in auditory evoked potentials in rats with salicylate-induced tinnitus. Brain Res.

Cebulla, M., and Sturzebecher, E. (2015). Automated auditory response detection: Further improvement of the statistical test strategy by using progressive test steps of iteration. Int J Audiol *54*, 568-572.

Don, M., and Elberling, C. (1994). Evaluating residual background noise in human auditory brain-stem responses. J Acoust Soc Am *96*, 2746-2757.

Don, M., and Elberling, C. (1996). Use of quantitative measures of auditory brain-stem response peak amplitude and residual background noise in the decision to stop averaging. J Acoust Soc Am *99*, 491-499.

Elberling, C., and Don, M. (1984). Quality estimation of averaged auditory brainstem responses. Scand Audiol *13*, 187-197.

Galbraith, G.C. (2001). Enhanced brainstem and cortical evoked response amplitudes: single-trial covariance analysis. Percept Mot Skills *92*, 659-672.

Jacobson, J.T., Jacobson, C.A., and Spahr, R.C. (1990). Automated and conventional ABR screening techniques in high-risk infants. J Am Acad Audiol *1*, 187-195.

Jewett, D.L., Romano, M.N., and Williston, J.S. (1970). Human auditory evoked potentials: possible brain stem components detected on the scalp. Science *167*, 1517-1518.

Le Prell, C.G. (2019). Effects of noise exposure on auditory brainstem response and speech-in-noise

tasks: a review of the literature. Int J Audiol *58*, S3-S32.

Lu, J., Huang, Z., Ma, Y., Li, Y., Mei, L., Yao, G., Wang, Y., Shen, X., and Wu, H. (2014). Comparison between hearing screening-detected cases and sporadic cases of delayed-onset hearing loss in preschool-age children. Int J Audiol *53*, 229-234.

Lu, J., Huang, Z., Yang, T., Li, Y., Mei, L., Xiang, M., Chai, Y., Li, X., Li, L., Yao, G., *et al.* (2011). Screening for delayed-onset hearing loss in preschool children who previously passed the newborn hearing screening. Int J Pediatr Otorhinolaryngol *75*, 1045-1049.

Madsen, S.M.K., Harte, J.M., Elberling, C., and Dau, T. (2018). Accuracy of averaged auditory brainstem response amplitude and latency estimates. Int J Audiol *57*, 345-353.

McKearney, R.M., and MacKinnon, R.C. (2019). Objective auditory brainstem response classification using machine learning. Int J Audiol, 1-7.

Mehraei, G., Hickox, A.E., Bharadwaj, H.M., Goldberg, H., Verhulst, S., Liberman, M.C., and Shinn-Cunningham, B.G. (2016). Auditory Brainstem Response Latency in Noise as a Marker of Cochlear Synaptopathy. J Neurosci *36*, 3755-3764.

Nishida, S., Nakamura, M., and Shibasaki, H. (1993). Method for single-trial recording of somatosensory evoked potentials. J Biomed Eng *15*, 257-262.

Ridley, C.L., Kopun, J.G., Neely, S.T., Gorga, M.P., and Rasetshwane, D.M. (2018). Using Thresholds in Noise to Identify Hidden Hearing Loss in Humans. Ear Hear *39*, 829-844.

Roeser, R.J., Valente, M., and Hosford-Dunn, H. (2007). Audiology. Diagnosis, 2nd edn (New York: Thieme).

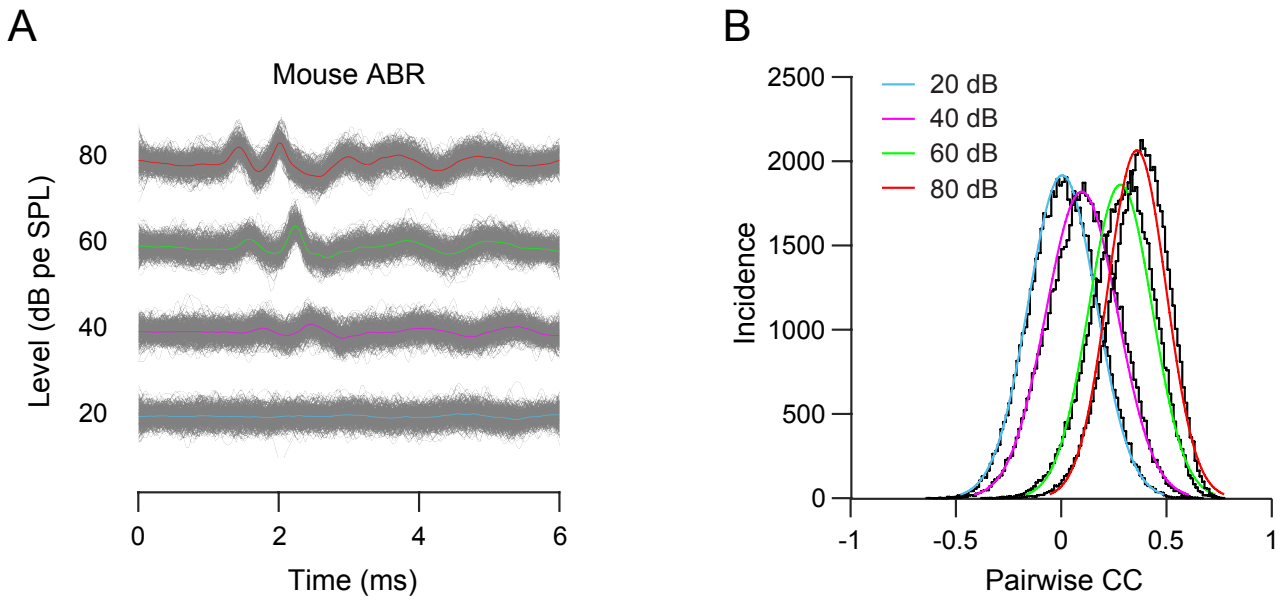
Suthakar, K., and Liberman, M.C. (2019). A simple algorithm for objective threshold determination of auditory brainstem responses. Hear Res *381*, 107782.

Valderrama, J.T., de la Torre, A., Alvarez, I., Segura, J.C., Thornton, A.R., Sainz, M., and Vargas, J.L. (2014). Automatic quality assessment and peak identification of auditory brainstem responses with fitted parametric peaks. Comput Methods Programs Biomed *114*, 262-275.

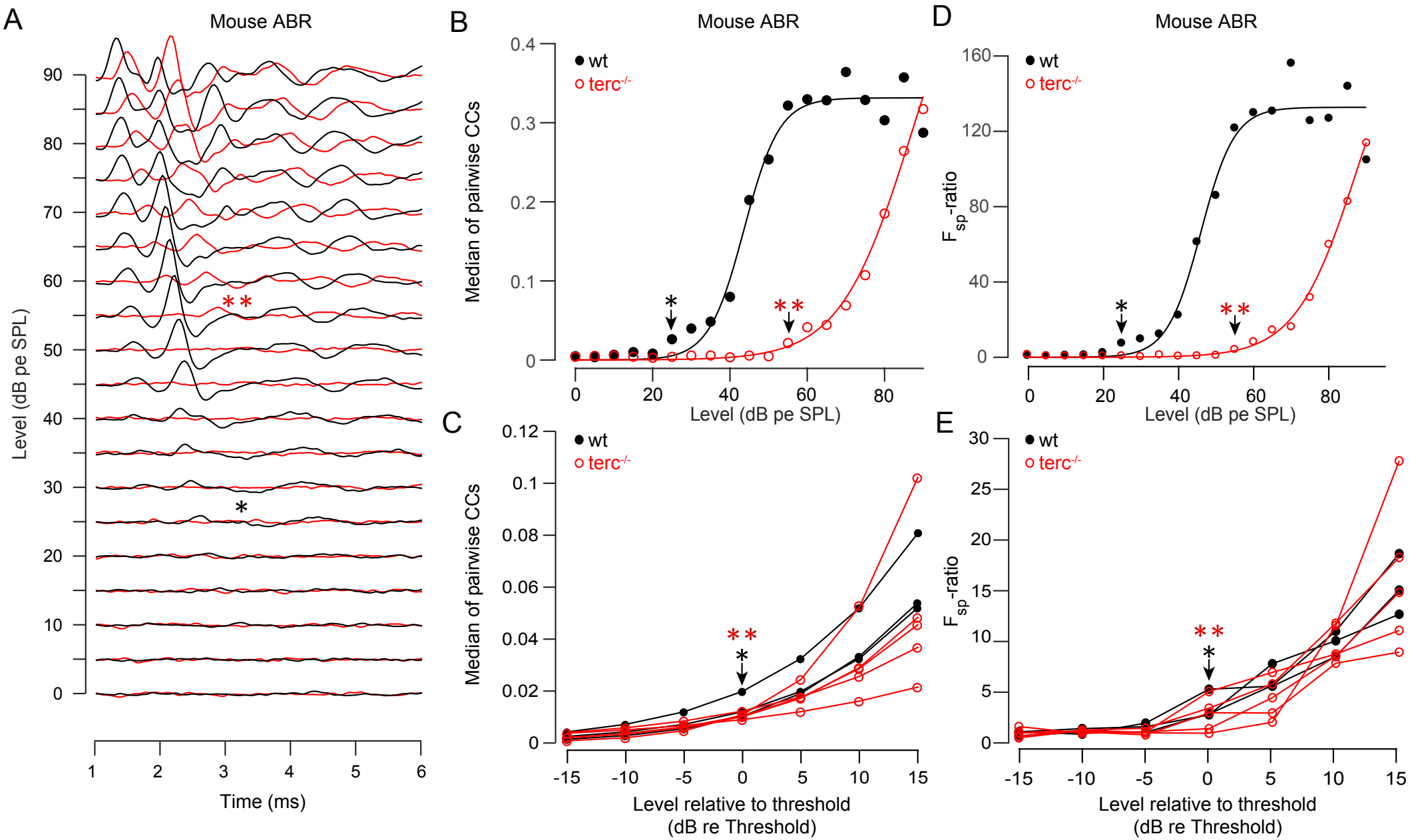
Vidler, M., and Parkert, D. (2004). Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. Int J Audiol *43*, 417-429.

Weber, B.A., and Fletcher, G.L. (1980). A computerized scoring procedure for auditory brainstem response audiometry. Ear Hear *1*, 233-236.

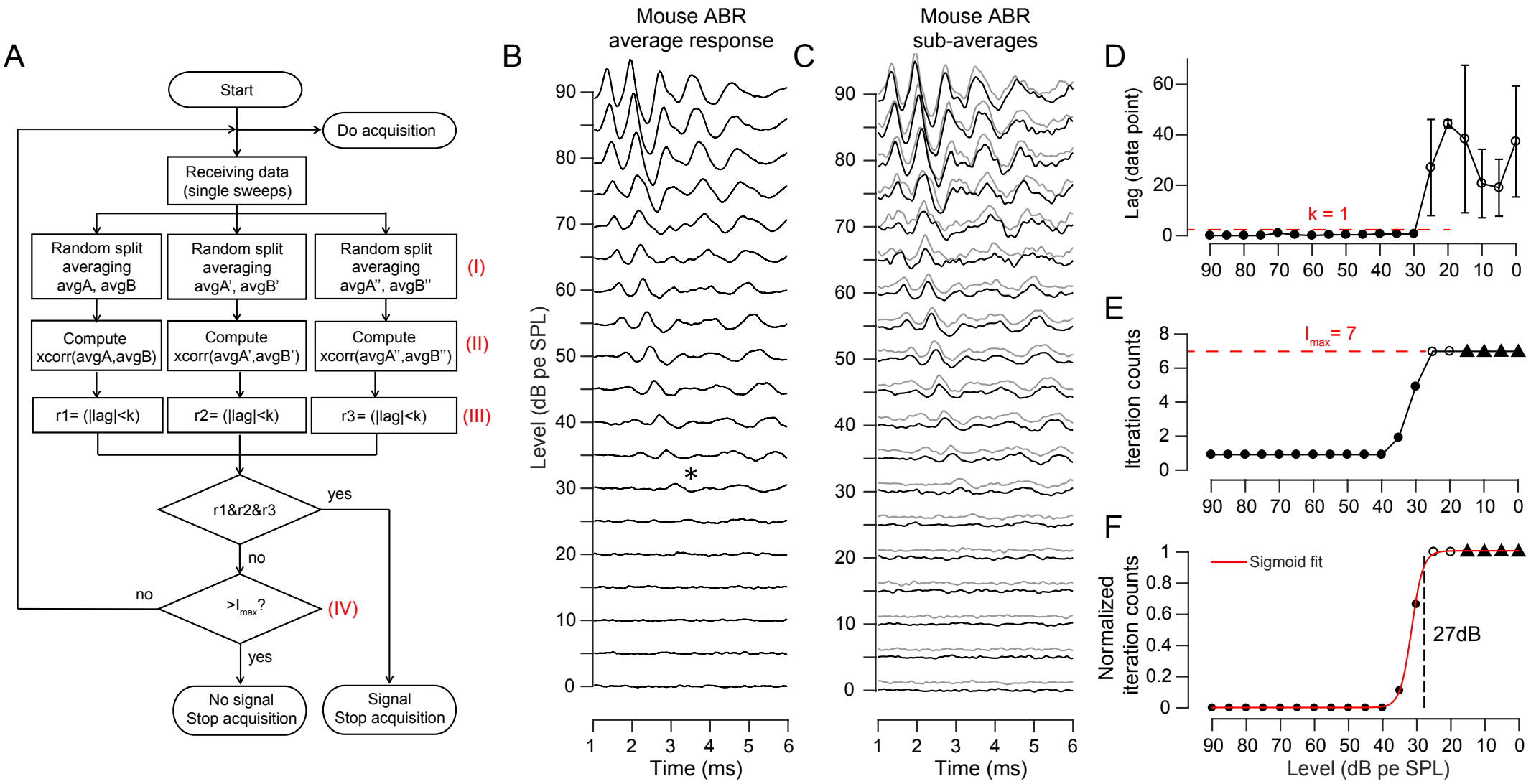
Xu, Z.M., De Vel, E., Vinck, B., and Van Cauwenberge, P. (1995). Application of cross-correlation function in the evaluation of objective MLR thresholds in the low and middle frequencies. Scand Audiol *24*, 231-236.



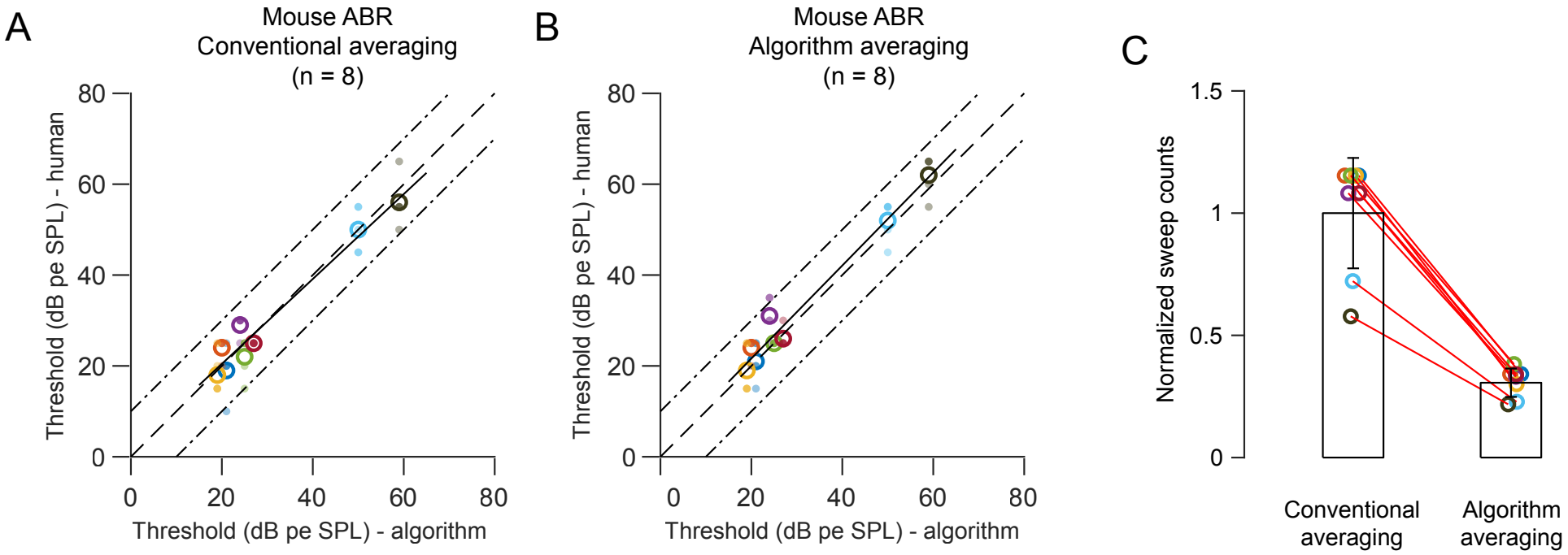
Wang et al., Fig 1



Wang et al., Fig 2



Wang et al., Fig 3



Wang et al., Fig 4