An Efficient and Robust Approach to Detect Auditory Evoked Responses using Adaptive Averaging

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

[1] Ear Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2] Shanghai Institute of Precision Medicine, Shanghai, China

[3] Department of Otolaryngology-Head and Neck Surgery, Shanghai ninth people's Hospital, Shanghai Jiao Tong University School of Medicine

[#] These authors contribute equally

* Corresponding authors

ABSTRACT

Objective: the auditory brainstem response is widely employed to evaluate hearing function of test subjects for both clinical and research purposes. Currently, hearing threshold estimation still relies on trained professionals to assess the audiograms, resulting in the largest cost component of the test. The objective of this study was to develop an automated approach to objectively and reliably detect the hearing thresholds.

Design: From eight mice and four human participants, we recorded for each sound level hundreds of single sweeps and asked how many sweeps were minimum required for a

response being detected from cumulative averages using correlation function. We named this procedure the adaptive averaging approach.

Result: We found an exponential increase in the required sweeps at near-threshold sound levels, modeling of which allowed objective and precise threshold estimation. The results from mouse and human recordings deviated consistently less than 5 dB from the expert-assessed values. Moreover, up to 69 % sweeps at suprathreshold sound tests were found redundant for the threshold estimation and could be avoided to improve test efficiency.

Conclusions: The adaptive averaging approach achieved objective and precise hearing threshold detection and implementation of this approach in commercial recording devices will automate the hearing test in a more reliable and cost-effective way.

INTRODUCTION

The auditory brainstem responses (ABRs) are brain electrical potential changes due to the synchronous neuronal activities evoked by suprathreshold acoustic stimuli (Jewett et al., 1970). In a hearing test, surface electrodes are placed on the scalp of the test subject to monitor the threshold sound level below which ABRs are no longer detectable. As a non-invasive technique, the ABR test has been worldwide employed in hospitals and clinics to exam auditory function, particularly for infants, adults with learning disabilities as well as 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 (up to ten milliseconds) of sound evoked potentials. These peaks appear to be one millisecond (ms) apart and have an amplitude of few hundreds nanovolt (nV), corresponding to synchronous activities arising from (I) auditory nerve, (II) cochlear nucleus, (III) superior olivary complex, (IV) lateral lemniscus and (V) inferior colliculus, respectively (Basbaum, 2008). Over decades, the ABR has not only been used to estimate the hearing threshold, also features of the waveform including peak latencies and amplitudes provide clinically significant information, for instance hidden hearing loss (Bramhall et al., 2018b; Mehraei et al., 2016; Ridley et al., 2018), tinnitus (Bramhall et al., 2018a; 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 sound levels is not always trivial and requires visual inspection from trained professionals under established guidelines. Currently, the test still demands audiologists or clinicians who are experienced in this technique to conduct recordings and interpret outcomes on site, which raises the largest cost component in such test. Moreover, according to a previous report (Vidler et al., 2004) cases with limited consensus about the assessed hearing thresholds are not rare, because the accuracy of waveform recognition highly depends on the skill and the experience of the assessors, especially when untypical waveforms or high background noise level are encountered. As extensive hearing test is employed for instance to diagnose progressive and acquired

hearing loss in preschool children (Lu et al., 2014; Lu et al., 2011) as well as to evaluate drug ototoxicity, there is a clear need for an automated approach of high accuracy and robustness to make the test less labor-intense and more objective.

Single-sweep ABR recordings are contaminated with background noise generated from muscle activity, adjacent power lines, etc. Therefore, it is almost exclusively required averaging over more than hundreds of sweeps to suppress random noise peaks, so that the signal-to-noise ratio (SNR) of the average responses is sufficient for an unambiguous ABR waveform recognition. In clinic, the waveform assessment is mainly based on visual inspecting of average responses and assisted by statistic measures like well-known single-point F ratio (F_{sp}) to evaluate the signal and noise characteristics (Elberling et al., 1984). Despite of recent progress using machine learning (McKearney et al., 2019), reliable assessments still heavily involve human supervision in a semi-objective fashion. In the last few decades, several methods have been proposed to automate the procedure. With the criterion whether sweep averages or single-sweeps are processed, they fall into two major categories, namely feature- and statistics-based detection strategies (Dobie, 1993). The former utilizes either quantification of the similarity between measured waveforms and existing templates from database (Valderrama et al., 2014), or feature recognition using trained artificial neural networks on human-annotated datasets (Alpsan et al., 1991; McKearney et al., 2019), but detection accuracy is still limited by the waveform heterogeneity, varying data quality, as well as inconsistent training data (McKearney et al., 2019). The latter

89 measures either inter-sweep variability to detect stable signals using correlation
90 function (Bershad et al., 1974; Weber et al., 1980; Xu et al., 1995) or level of residual
91 noise (Don et al., 1994; Elberling et al., 1984) to describe the 'quality' of the signal
92 through a scoring procedure, for which calibrations are needed for thresholding the
93 signals with arbitrary boundary. Although some of these techniques have already been
94 integrated in commercial devices, they are mainly employed for crosscheck purpose
95 due to frequent detection errors caused by noise-related artifacts.
96
97 To better understand the noise and ABR characteristics, we first recorded from mice at
98 each sound level hundreds of sweeps and found pairwise correlation coefficients of
99 those sweeps followed a normal distribution. Upon declined sound levels, a series of
100 peaks with shifted center towards zero were obtained as a result of progressively
101 overwhelmed responses in noise. This implies consistent threshold detection by
102 thresholding quantified correlation requires comparable SNRs across recordings and
103 frequent calibration, however neither of which is practical under daily experimental
104 settings. In the present study, we overcame this limitation by a novel approach using
105 adaptive averaging, in which sweeps were added iteratively for averaging in order to
106 test how many sweeps are required for stimulus-related waveform being detected
107 through cross correlation measurement. The hearing threshold was then objectively
108 estimated by modeling the change in the minimal sweep numbers at peri-threshold
109 sound levels. The autodetected thresholds were found deviated from the expert-
110 assessed ground truth consistently less than 5 dB. In addition, the minimal sweep

numbers are instructive for an efficient averaging strategy based on SNR adaptation. When implemented in commercial ABR devices, this automated approach could stop on-going averaging upon detected waveforms and save up to 69 % redundant sweeps according to our reconstructed experiments, and thereby shorten the test duration without compromising the accuracy of test results.

MATERIALS AND METHODS

Animals and Ethics

C57BL/6 mice (wt) were purchased from Sino-British SIPPR/BK Lab.Animal Ltd (Shanghai, China). The terc-/- mouse line was kindly donated by Lin Liu (Nankai University, China) and bred in house. This study is conduct at the Ear Institute and the Hearing and Speech Center of Shanghai Ninth People's Hospital. All procedures were reviewed and approved by the Institutional Authority for Laboratory Animal Care of the hospital (HKDL2018503).

ABR recordings

Mouse ABRs were recorded using a 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). Through the test, animal body temperature was maintained at around 37°C using a regulated heating pad (Harvard Apparatus, US) and rectal thermal probe. Electrical potential time courses were

133 registered via subdermal needle electrodes (Rochester Electro-Med. Inc., US) placed at
134 vertex (record electrode), left infra-auricular mastoid region (reference electrode) and
135 right shoulder region (ground electrode) of the test subjects. The acoustic stimuli of
136 3 ms tone pips at 16 kHz were delivered via a multi-field magnetic speaker (Tuck-Davis
137 Tech. Inc., US) positioned 10 cm from the vertex of the test subjects. The stimulus rate
138 was 21 per second and the evoked potentials were sampled at 24 kHz. 500 sweeps were
139 acquired at each test sound level starting from 90 dB to 0 dB with an increment of 5 dB.
140 For one animal, we acquired signals evoked by peri-threshold sound levels (±10 dB
141 with a 1-dB increment procedure).
142
143 Human ABRs were collected from four adult volunteers aged between 21 and 29 years-
144 old without history in abnormal auditory function using a commercial device
145 (Intelligent Hearing Systems, US) with Smart EP software. An electromagnetically
146 shielded insert earphone (ER-3) was applied to deliver click stimulation to the left ear
147 at a rate of 37.1 per second. The recorded potentials were amplified by a factor of
148 100,000 and filtered with a 100 Hz (high-pass) and 3000 Hz (low-pass) filter. Averaged
149 signals over 500, 1000, and 2000 sweeps were acquired 3 times at each test sound level
150 starting from 60 dB to 0 dB with an increment of 5 dB.
151
152 **Analysis of single-sweep correlation**
153 Analysis of inter-sweep correlation was conduct in mouse recordings. All acquired
154 single-sweep ABRs were analyzed using self-written routines in MATLAB

(MathWorks Inc., US). The artifact rejection level was set to 55 μV and a high-pass filter (smoothing splines with a smoothing parameter of 0.5, MATLAB function) was applied to the raw data. The pairwise correlation was calculated using MATLAB function from sweeps cropped with a temporal window of 1-6 ms post-stimulation onset, resulting in 61,075 correlation coefficients from 350 sweeps at each test sound levels. The histograms of the obtained coefficients were fitted with a normal distribution.

Adaptive averaging method

For the mouse dataset, each iteration recruited 50 sweeps into MATLAB workspace until all 350 sweeps were added. Within iterations all sweeps in the workspace were subdivided randomly into two groups for separated averaging, between which a cross correlation function was computed using MATLAB function. The existence of stimulus-related waveform was tested based on whether the obtained correlation has a maximum at a lag less than 1% of total signal length (1-6 ms post-stimulation onset). In each iteration the lag value was measured three times with regrouped sweeps to reject cases with coincidently overlapped signals at required lag value. For each sound level, it was measured at which iteration stimulus-related response was detected. The maximum number of executed iterations was set to 7, which is limited to total sweep numbers (after excessive amplitude exclusion). A sigmoid function (1) was employed to model the minimum required iteration numbers after normalization $N(S)$ at each sound level S and a growth of 0.9 in the function corresponds to the estimated hearing

threshold *T*. The constant *k* equals 0.6, which was obtained from the best fitting.

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

As for the human dataset, averaged responses over 500, 1000 and 2000 sweeps were loaded in each iteration. Lag value was measured three times with a combination of 2 out of 3 averaged responses of the same sweep number. The maximum executed iteration number was set to 7. For iterations requiring averages over 1500 or more than 2000 sweeps, the combined average responses *avg(M+N)* containing *M* plus *N* sweeps were obtained from the weighted averaging of *avg(M)* and *avg (N)* using the equation (2).

$$avg(M+N) = \frac{M \bullet avg(M) + N \bullet avg(N)}{M+N} \qquad (2)$$

The maximal lag value allowed for a true ABR waveform was 2 % of the total signal length (0−10 ms post-stimulation onset). Same as for the mouse data, hearing threshold was interpolated from a fitting with sigmoid function at a growth of 0.9.

Hearing Threshold Assessment

All hearing thresholds were verified by visual inspection of five independent clinicians blinded to the test subjects. For each experiment, recordings averaged either over fixed sweep counts (the conventional averaging) or according to the outcomes from the adaptive averaging. Averaged audiograms were provided in MATLAB figures which enable interactive zoom function for better waveform recognition. Results were collected and mean values were calculated using the 3/5 method with the highest and

the lowest estimation dropped.

RESULTS

Correlations between Single-sweep ABR Recordings

As ABRs are usually embedding in high level background activities as well as system noise, smooth baseline and clear waveform, if present, are usually obtained after an averaging over hundreds of sweeps (Fig 1 A). We recorded from mouse single-sweeps upon 16 kHz tone-pips at different levels of loudness. After a small fraction had been excluded due to excessive amplitudes caused by movement artifacts, more than 350 sweeps were obtained and then filtered through a high-pass filter (Nishida et al., 1993) to remove low frequency fluctuations in short latency ABR components (Suppl 1). Pairwise correlations were computed from the sweeps cropped with a temporal window of 1-6 ms post-stimulation onset. The obtained correlation coefficients (CCs) at each sound level were found following normal distributions (Fig 1 B). When sweeps were recorded upon suprathreshold sound levels, corresponding shifts of the CC distributions were observed, indicating increased degrees of positive correlation. In line with previous studies (Galbraith, 2001), this result confirmed that ABR waveforms in single sweeps are highly correlated and measurable using correlation function. Moreover, when fit with a single gaussian function, invariable peak width (0.1778 ± 0.0154) was obtained from each distribution, implying reliable auditory responses with identical waveform and an add-on effect of constant uncorrelated noises.

Objective Threshold Detection using Correlation Coefficients

Next, we tested whether one can use correlation measurements to detect objectively the hearing thresholds. Recordings were performed using acoustic stimuli varying from 90 dB to 0 dB in three wildtype mice (wt) of normal hearing and five telomerase-knock-out mice (terc-/-) of high percentage in deafness. Representative responses of a wt and a terc-/- animal were illustrated in Fig 2 A and the medians of CCs were plotted as a function of test sound levels (Fig 2 B). Between two rising phases, a shift of about 30 dB was found, consistent with the threshold elevation estimated by visual inspection. This result confirmed CCs as a promising measure to detect the hearing threshold, as derived from early theoretical work (Bershad et al., 1974). In order to obtain a critical test boundary between supra- and sub-threshold sound level, we computed from eight animal at each sound level the medians of CCs and aligned them to the visual inspected thresholds (Fig 2 C). Sweeps upon stimuli above hearing threshold were found to have stronger positive correlation with a median greater than 0.0119 ± 0.0034. Although this empirically determined value could be used to threshold detectable ABRs, the detection accuracy was limited by close values obtained at near-threshold levels as well as inter-trial variability (data not shown).

Adaptive Averaging Method

It is impractical to keep identical recording conditions (including electrode position,

243 depth of anesthesia, etc.) for multiple recording sessions or perform frequent calibration
244 before each experiment for daily use. In order to develop a more robust and applicable
245 approach, we designed a novel procedure to adapt sweep number for averaging to the
246 requirement for waveform detection. This means, instead of thresholding a detectable
247 positive correlation using pre-calibrated critical test boundary, the new approach asks
248 for how many sweeps are required to generate two average responses that are positively
249 correlated above certain criterion (Fig 3 A). As the noise level is suppressed
250 exponentially by averaging over increasing number of sweeps, this method adapts
251 intrinsically to varying data quality and works in a calibration-free fashion.
252
253 In detail, recorded sweeps were randomly divided into two groups and cross correlation
254 was calculated between group averages (Fig 3 C, see Fig 3 B for results using
255 conventional averaging). In the presence of stimulus-related waveforms, one expected
256 a neglectable lag (e.g. less than 1 % of analyzed data points) at which maximum
257 correlation was found (Fig 3 D). To reject false positives caused by coincident overlap
258 of noise peaks, each lag test required consistent results from three parallel
259 measurements on regrouped sweeps. For each failed lag test, additional sweeps were
260 added and the random split averaging was repeated, until a response was detected or
261 the maximum number of iterations was reached. As shown in Fig 3 E, the number of
262 executed iterations increased rapidly starting at near-threshold sound levels and
263 reaching its maximum at the first subthreshold sound level as expected. Although the
264 sharp transition allows direct readout of the lowest suprathreshold sound level, we

265 further tested whether precise estimation beyond the increment between test sound
266 levels could be achieved by modeling the result (Fig 3 F). From one animal, sweeps
267 were acquired upon peri-threshold sound levels with 1-dB increment and the increase
268 in executed iterations was fit with a sigmoid function. Similar fit parameters were
269 obtained from test datasets of 1-dB and 5-dB increment (Fig 3 F), indicating possible
270 interpolation of the threshold from regular 5-dB increment procedure to achieve 1-dB
271 detection precision based on the established model.
272
273
274 **Accuracy and Efficiency of the Adaptive Averaging Method**
275 To evaluate the performance of the proposed adaptive averaging approach, we applied
276 the procedure to ABR recordings from eight mice as well as four human test participants.
277 Note that for human test, average responses were collected multiple times with different
278 sweep numbers to simulate the random split averaging over sweeps (see Suppl 2 for the
279 variant of flowchart), as the commercial devices in our hospital did not support
280 exporting single-sweeps. When compared to the waveform assessments of five
281 independent clinicians, the new approach yielded correct hearing threshold detection
282 (deviated less than 5 dB from the expert-assessed ground truth) in 8/8 cases for mouse
283 and 4/4 for human (Fig 4 A), indicating a detection accuracy of near human
284 performance. In addition, visual inspection of the adaptive averaged responses and
285 those averaging over a fixed sweep counts yielded similar test results (Fig 4 A),
286 suggesting limited contribution of redundant sweeps in conventional averaging to

visual waveform recognition. When compared to the conventional averaging at all suprathreshold plus the first two subthreshold stimuli, the new method required up to 69.6 ± 5.8 % and 36.2 ± 21.1 % less sweeps for a correct threshold estimation in the mouse and human ABR tests, respectively (Fig 4 B).

Discussion

In the last few decades, attempts were made towards automated ABR analysis. Several techniques were proposed including the usage of correlation functions to discriminate stable components (ABR waveforms) from random fluctuations (system noise and background activities). In this study we found pairwise CCs between sweeps follows a normal distribution and medians of which reflect the power of stimulus-related waveform over baseline activities. The observed broadening of the distribution is probably caused by random fluctuations which by chance weak correlate or anti-correlate with the true responses. The observed constant peak widths for different test sound levels implies low variation in the ABR waveforms and comparable noise level within single recording session (Fig 1 B). This finding supports the notation that signal transmission in the auditory system is of high fidelity and SNR is the major challenge for detecting true hearing thresholds.

In practice, comparable SNRs are expected within single recording session except for electrode displacement or weak anesthesia (data not shown), allowing threshold

309 detection from a pre-calibrated system based on correlation measurement (Fig 2 C). It
310 is, however, unrealistic to keep the same experimental settings including the same
311 system, electrode position, anesthesia depth, etc. for different test subjects as well as in
312 multiple recording sessions, hampering a plug-in application of correlation-based
313 threshold autodetection. In fact, we showed CCs aligned to the hearing thresholds vary
314 between 0.009 and 0.0199 due to different SNRs under our experimental settings (Fig
315 2 C). This variation is about the same magnitude to those caused by different peri-
316 threshold sound levels, which leads to limited detection accuracy.
317
318 The basic idea of the adaptive averaging approach is to keep a standard waveform
319 detection criterion but adapt the SNR by iteratively increased sweep number for
320 averaging. In previous study of middle latency response analysis (Xu et al., 1995), it
321 has been proposed that neglectable lag in cross correlation between two consecutively
322 recorded average signals can serve as a good indication for suprathreshold responses,
323 because in contrast to stimulus-related responses, random distributed noise peaks may
324 superimpose at any time shifts. Here we adapted this idea to detect ABR waveforms in
325 two subgroup averages of sweeps. The underlying logic is as follows: one starts with
326 constant noise level and unknown signal power; each iteration recruits more sweeps in
327 the averaging to suppress progressively the noise level of the updated average responses
328 until the true ABR waveform is no longer overwhelmed in noise and becomes
329 detectable by cross correlation function. Thereby, one expects (1) converging lags to
330 zero requires averaging over less sweeps for recordings containing strong responses

than those with weak responses; (2) the absence of response leads to non-converging lags irrespective of sweep numbers used for averaging. These are supported by our observations that detection of responses upon suprathreshold stimuli requires minimum number of iterations (Fig 3 E), as well as three-fold measured lags show large variation at subthreshold sound levels (Fig 3 D). The steep increase in the number of executed iterations (equivalent to sweep numbers used in averaging) is characteristic for the peri-threshold responses and insensitive to the absolute values of SNRs, enabling a calibration-free automatic threshold detection. Further we showed modeling of the number of executed iterations enables more precise threshold estimation beyond the increment of test sound levels through interpolation, in our case up to 1 dB (Fig 3 F). Alternatively, the adaptive averaging approach enables a test strategy with progressively reduced increment at peri-threshold sound levels to achieve required precision.

For both mouse and human recordings, the automated approach is extreme reliable in detecting hearing thresholds within a maximum error of ± 5 dB from the expert-assessed ground truths provided by five independent clinicians (Fig 4 A). Besides, the observation variation is comparable between the visual inspected groups either using conventional or adaptive averaging, as well as between visual inspection and automatic detection (Fig 4 A). Based on these results we conclude that the automated approach achieves accurate and robust threshold estimation of near human performance. When to stop averaging is an important decision to make during the ABR data collection, it

not only makes the hearing test more efficient by avoiding prolonged acquisition and redundant data, but also specifies the confidence level of waveform interpretation (Don et al., 1996). Since the underlying principle of the new approach is to average just enough sweeps for the SNR above detection limit, intrinsically it returns instruction to stop averaging when a response is detected. Meanwhile, it is temporally efficient to terminate test when previous iterations with higher sound levels fail to detect a response, because in this case acquired sweeps at lower sound level is supposed to contain pure noise. The reconstructed experiments to compare required sweep number by the adaptive averaging and the fixed sweep counts in the conventional averaging show a reduction up to 69 % and 36 % for mouse and human ABR test, respectively (Fig 4 B). Since the maximum sweep count used here to rule out undetectable ABRs was chosen more than necessary, further improvement can be made by introducing single-point F ratio (Elberling et al., 1984) or residual noise analysis (Don et al., 1994) to terminate averaging. In fact, the acquisition speed of modern ABR devices is usually not a limiting factor (more than 2000 sweep per minute), maybe more importantly the adaptive averaging approach sets the minimum requirement for the data quality that ensures unambiguous waveform recognition by both human and machine.

The present work has demonstrated a robust and efficient approach to automate hearing threshold detection from recorded ABRs. In contrast to other existing techniques, the basic idea of our approach is not to quantify how likely the recording contains a response, but for the first time to model how many sweeps are required for an averaged

response to counteract the background noise and become detectable by cross correlation function. By doing so, we bypassed the long-lasting challenges caused by varying SNR of ABR recordings and in addition obtained a strategy to stop averaging in an unsupervised fashion. Hence, the proposed method has a remarkable potential of being implemented in commercial ABR devices to make hearing test more reliable and cost-effective.

References

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

Basbaum, A.I. 2008. The senses : a comprehensive reference Elsevier, Amsterdam ; New York.

Bershad, N.J., 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., McMillan, G.P. 2018a. Tinnitus and Auditory Perception After a History of Noise Exposure: Relationship to Auditory Brainstem Response Measures. Ear Hear 39, 881-894.

Bramhall, N.F., McMillan, G.P., Kujawa, S.G., Konrad-Martin, D. 2018b. Use of non-invasive measures to predict cochlear synapse counts. Hear Res 370, 113-119.

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

Dobie, R.A. 1993. Objective response detection. Ear Hear 14, 31-5.

Don, M., Elberling, C. 1994. Evaluating residual background noise in human auditory brain-stem responses. J Acoust Soc Am 96, 2746-57.

Don, M., 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, 491-9.

Elberling, C., Don, M. 1984. Quality estimation of averaged auditory brainstem responses. Scand Audiol 13, 187-97.

Galbraith, G.C. 2001. Enhanced brainstem and cortical evoked response amplitudes: single-trial covariance analysis. Percept Mot Skills 92, 659-72.

Jacobson, J.T., Jacobson, C.A., Spahr, R.C. 1990. Automated and conventional ABR screening techniques in high-risk infants. J Am Acad Audiol 1, 187-95.

Jewett, D.L., Romano, M.N., Williston, J.S. 1970. Human auditory evoked potentials: possible brain stem components detected on the scalp. Science 167, 1517-8.

Lu, J., Huang, Z., Ma, Y., Li, Y., Mei, L., Yao, G., Wang, Y., Shen, X., 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, 229-34.

Lu, J., Huang, Z., Yang, T., Li, Y., Mei, L., Xiang, M., Chai, Y., Li, X., Li, L., Yao, G., Wang, Y., Shen, X., Wu, H. 2011. Screening for delayed-onset hearing loss in preschool children who previously passed the newborn hearing screening. Int J Pediatr Otorhinolaryngol 75, 1045-9.

McKearney, R.M., 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., Shinn-Cunningham, B.G. 2016. Auditory Brainstem Response Latency in Noise as a Marker of Cochlear Synaptopathy. J Neurosci 36, 3755-64.

Nishida, S., Nakamura, M., Shibasaki, H. 1993. Method for single-trial recording of somatosensory evoked potentials. J Biomed Eng 15, 257-62.

Ridley, C.L., Kopun, J.G., Neely, S.T., Gorga, M.P., 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., Hosford-Dunn, H. 2007. Audiology. Diagnosis. 2nd ed. Thieme, New York.

Valderrama, J.T., de la Torre, A., Alvarez, I., Segura, J.C., Thornton, A.R., Sainz, M., Vargas, J.L. 2014. Automatic quality assessment and peak identification of auditory brainstem responses with fitted parametric peaks. Comput Methods Programs Biomed 114, 262-75.

Vidler, M., 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, 417-29.

Weber, B.A., Fletcher, G.L. 1980. A computerized scoring procedure for auditory brainstem response audiometry. Ear Hear 1, 233-6.

Xu, Z.M., De Vel, E., Vinck, B., 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, 231-6.

Acknowledgement

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 visual inspection. We would like to thank Dr. Guangming Chen and Dr. 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.), The Elite Program at Shanghai Ninth People's Hospital (JY201802 to Y.H., Z.H.), B.L was supported by The National Science Foundation for Young Scientists of China (No81700903) and The

SHIPM-mu Fund from Shanghai Institute of Precision Medicine (JC201808).

Declarations of interest

The authors have declared that no conflicts of interest exist.

Author contribution

Y.H., H.W designed the study; B.L. and X.D. acquired the data; Y.H. and H.Y.W created the software; Y.H., H.Y.W., X.W, and Z.H., analyzed and interpreted the data. Y.H., H.Y.W., H.W drafted and revised the work.

Figure Legends

Figure 1. Pairwise correlation coefficients of single-sweep ABR recordings. (A) sweeps were recorded at 4 different sound levels (grey lines) and characteristic ABR waveforms, if exist, were obtained via averaging (colored lines). (B) Histograms of pairwise correlation coefficients that were computed from the recorded sweeps in A (black lines). A normal distribution (colored lines) was used to fit each histogram. The mean value of the resulted full width at haft maximum was 0.1778 ± 0.0154 (mean ± s.d.).

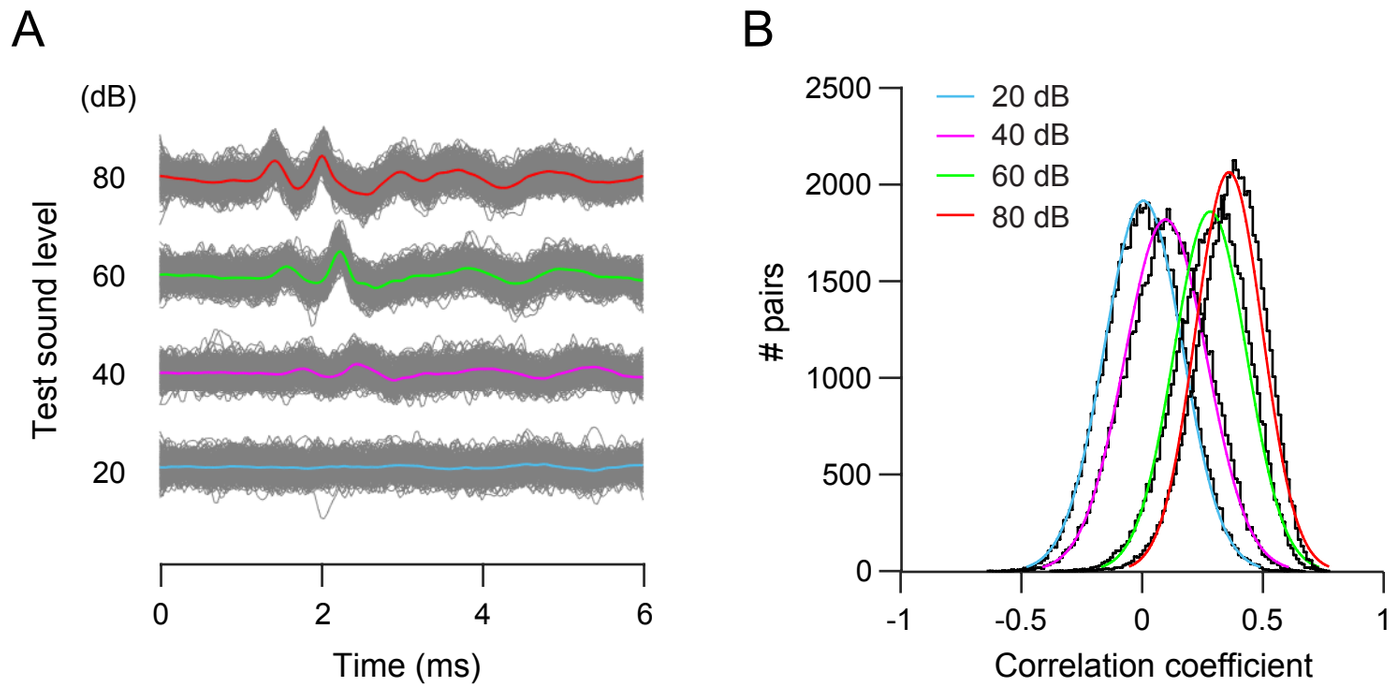
Figure 2. Threshold estimation using correlation coefficients. (A) Representative audiograms averaged over 350 sweeps were recorded from mice of normal (black lines) or impaired hearing (red lines). Waveforms with different peak amplitudes were observed for recordings at suprathreshold stimuli (above 25 dB for animal of normal hearing and 55 dB for animal of impaired hearing). (B) The same sweeps as for A were used to compute pairwise correlation coefficients at each sound level, respectively. The median correlation coefficients were plotted against corresponding test sound levels and fitted with a sigmoid function. Delayed raising of the medians upon increasing sound levels was observed in animal with impaired hearing, which is in line with the threshold elevation illustrated in A. (C) Median correlation coefficients at peri-threshold sound levels were measured from eight mice. Each curve was aligned to visual inspected thresholds at zero. Mean value of the median correlation coefficients at hearing threshold was 0.0119 ± 0.0034 (mean ± s.d.) varying between 0.0090 and 0.0199.

Figure 3. Threshold detection using adaptive averaging method (A) Flowchart of the adaptive averaging method. (B) Representative mouse audiogram averaged over 350 sweeps. The visual inspected threshold was between 25 dB and 30 dB. (C) Subgroup averages at each sound level (black versus grey lines) were obtained by averaging randomly split two groups of sweeps. The intermediate averages were updated by iterations for measuring the cross correlation. (D) The absolute lags in data point of the maximum cross correlation were three-fold measured at each test sound level. For

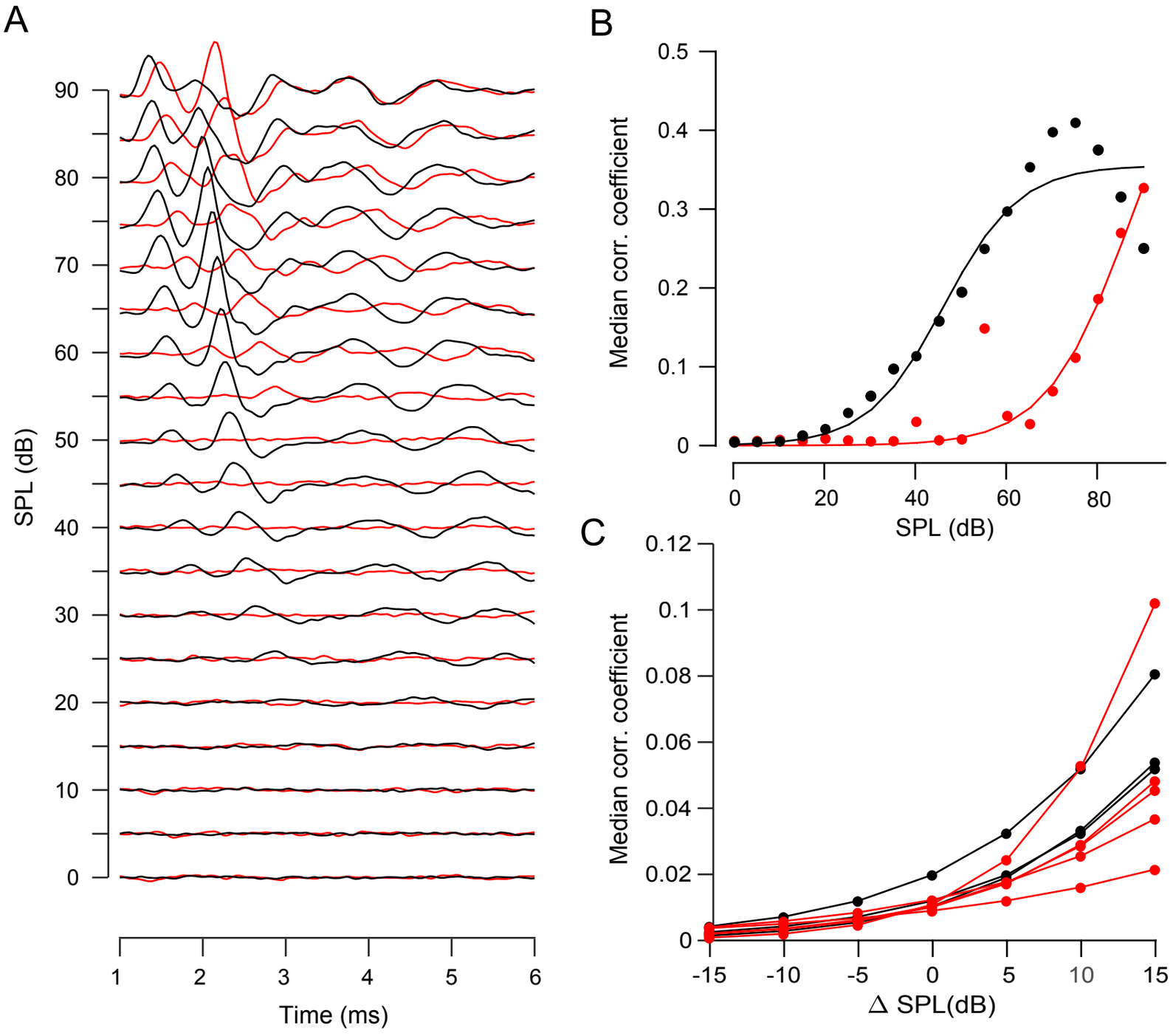
recordings upon suprathreshold stimuli (dots) a consistent small absolute lag values (no more than 1 data point) were obtained, while non-zero values with large variations (31.11 ± 16.12) were observed for cases of subthreshold stimuli (cycles). (E) In each iteration 50 new sweeps were added into the data pool for random split averaging, until the updated lag value was no more than k data point (in this case k = 1). Minimum required iterations were plotted for tests upon suprathreshold (dots) and subthreshold sound levels (cycles) with an upper limit at 7 to avoid infinite iterations. It is recommended to stop a test, when the hearing threshold is confirmed by two consecutive runs without detectable waveforms (black cycles), so that iterations at lower test sound levels are saved (red cycles). (F) The stepwise increase in required iterations was fitted by a sigmoid function after normalization, allowing threshold interpolation beyond the increment of test sound levels. As calibrated with peri-threshold test with 1-dB increment, the most accurate threshold estimation corresponds to a growth of 0.9 in the best fitted function.

Figure 4. Performance of the adaptive averaging method in detecting hearing threshold. (A) ABR recordings from mice (n=8) and human participants (n=4) were evaluated by 5 independent clinicians or the automated approach. As ground truth, the mean thresholds were used which were based on human inspection of conventional averaged responses with the highest and the lowest value excluded. The thresholds reported by clinicians based on the mouse ABR audiograms using conventional averaging (fixed sweep counts, black dots) and the adaptive averaged responses (varying sweep numbers,

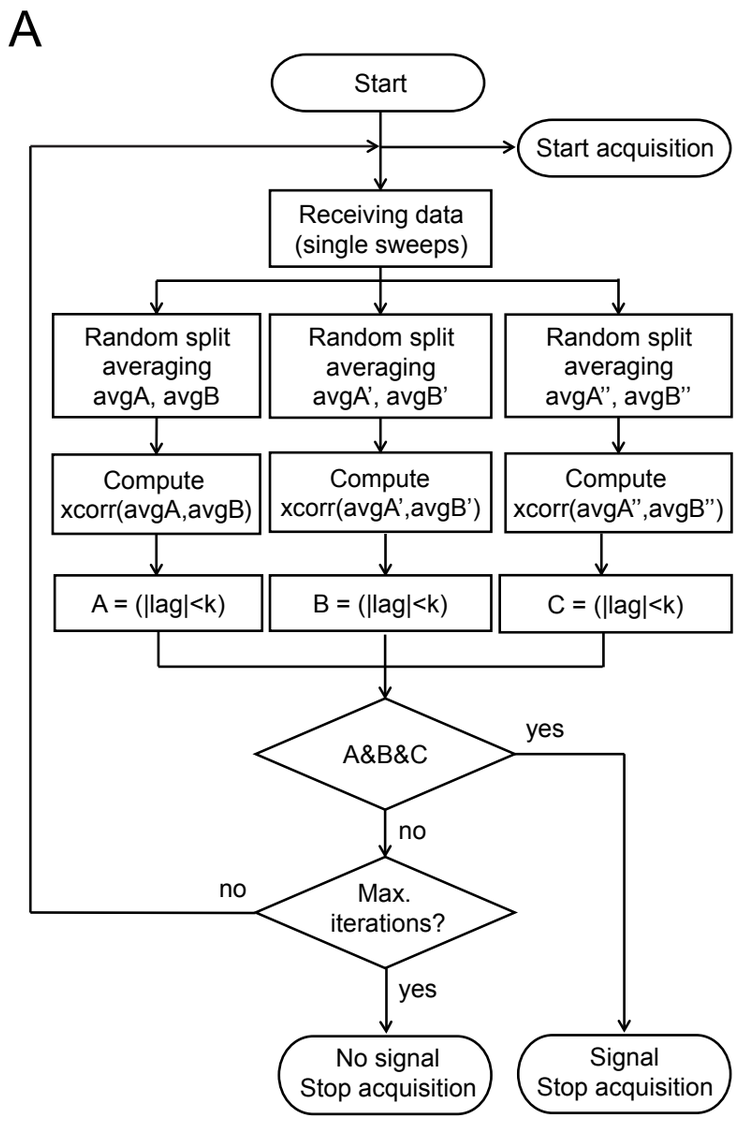
514 red dots) deviated from the expert-assessed ground truth by -0.250 ± 2.719 dB and
515 1.875 ± 3.487 dB, respectively, while it was 0.000 ± 3.590 dB for the automatic detected
516 threshold (red cycles). Similar results were obtained from human ABR audiograms with
517 estimation deviations of -0.083 ± 1.200 dB for the conventional averaging (black dots)
518 and 0.083 ± 1.200 dB for the adaptive averaging (red dots) through visual inspection,
519 compared to 0.167±1.258 dB for the automated approach (red cycles). (B) Total sweep
520 numbers required for the adaptive averaging (black boxes) were compared to those used
521 in the conventional averaging (red boxes). Within each experiment, sweep numbers at
522 all suprathreshold levels and the first two subthreshold levels were counted and
523 normalized to the default setting for averaging (350 sweeps for mouse ABR and 3000
524 sweeps for human ABR). The adaptive averaging required 30.4 ± 5.8 % sweeps in
525 mouse ABR tests (left) and 63.8 ± 21.1% in human ABR tests (right) to perform
526 successful threshold detection. Error bar represents standard deviation.

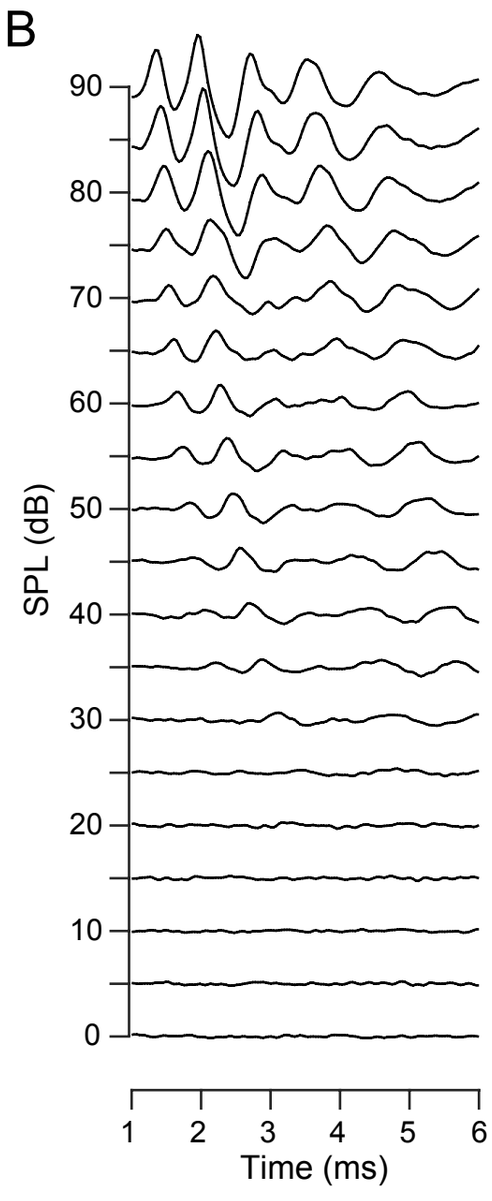


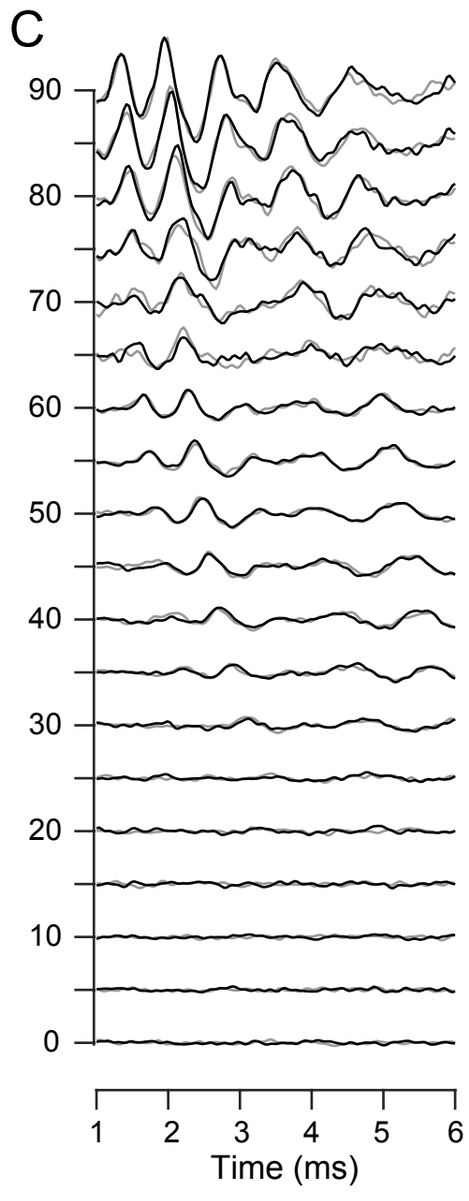
Wang et al., Fig 1

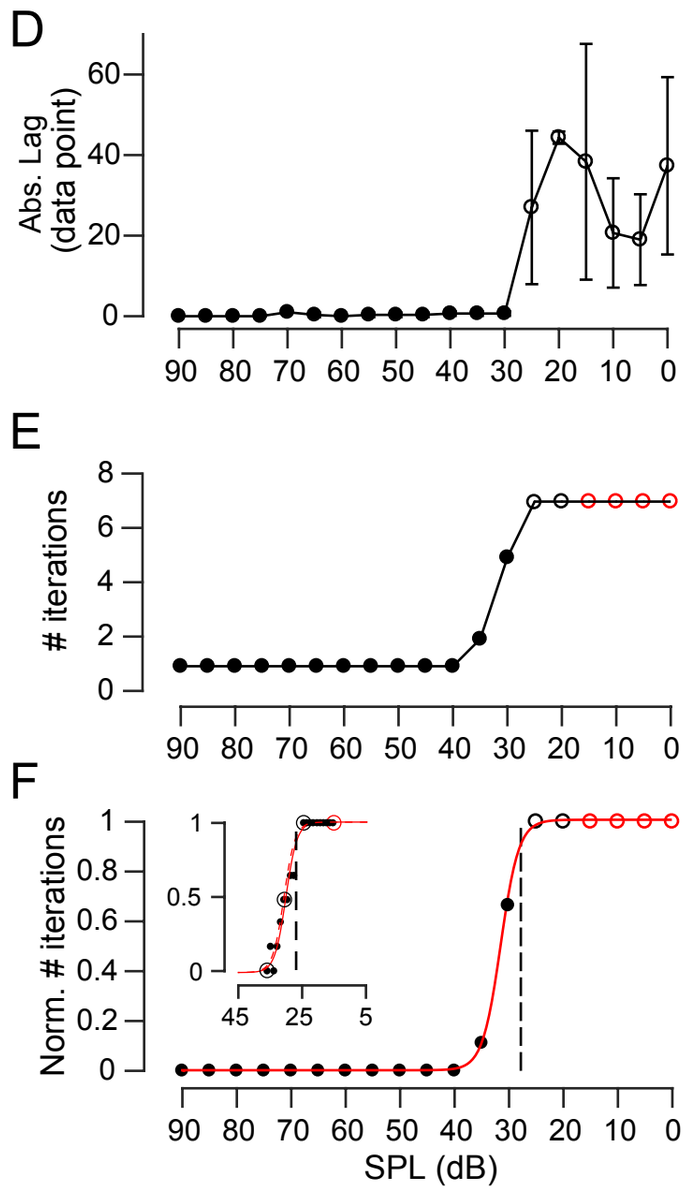


Wang et al., Fig 2

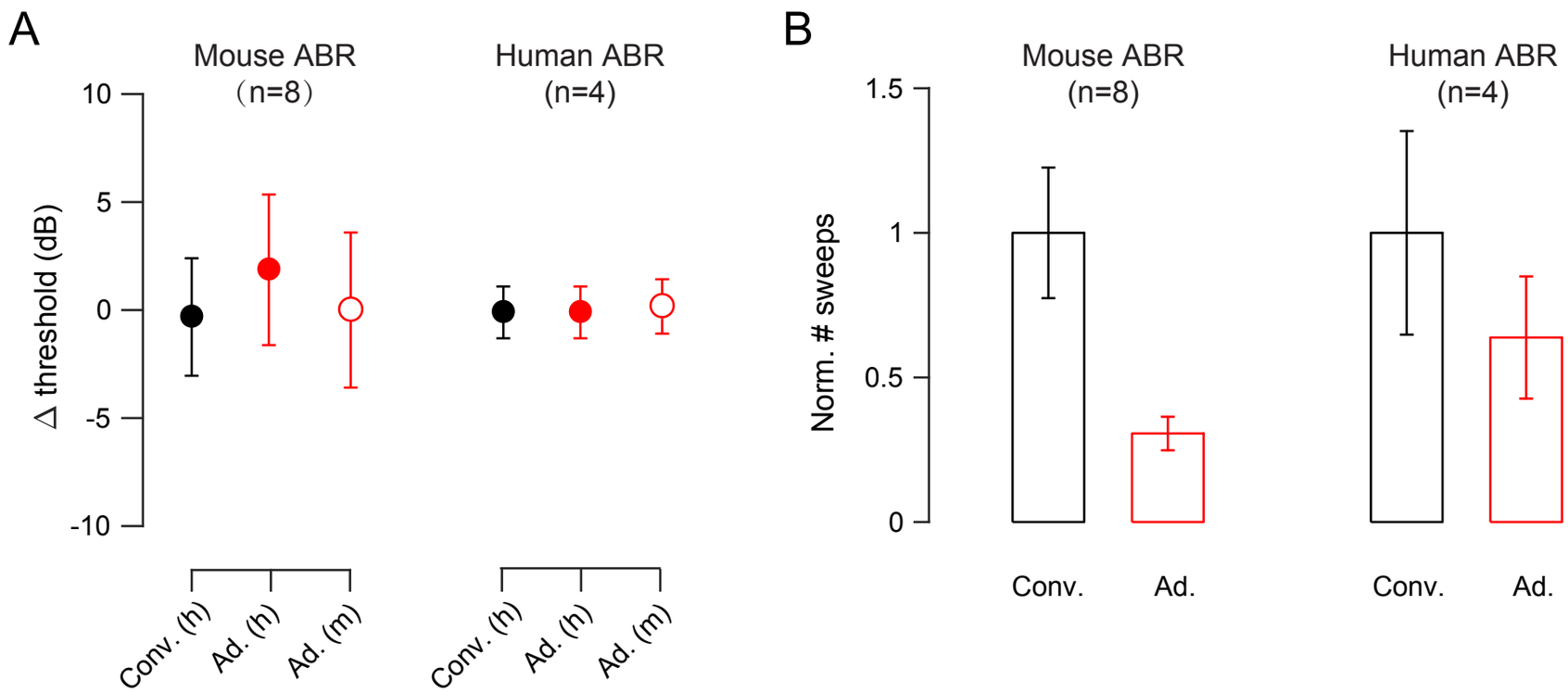








Wang et al., Fig 3



Wang et al., Fig 4