1 Title: Real-time Hearing Threshold Determination of Auditory Brainstem 2 Responses by Cross-correlation Analysis

3

4 Authors: Haoyu Wang^{1-4,7}, Bei Li^{1,7}, Yan Lu⁴, Kun Han¹, Haibin Sheng¹⁻³, Jialei
5 Zhou⁵, Yumeng Qi⁴, Xueling Wang¹⁻³, Zhiwu Huang¹⁻³, Lei Song^{1-3,*} and
6 Yunfeng Hua^{1-4,6,*}

7

8 Affiliations:

9 ¹Department of Otolaryngology-Head and Neck Surgery, Shanghai Ninth
10 People's Hospital, Shanghai, China

11 ²Ear Institute, Shanghai Jiao Tong University School of Medicine, Shanghai,
12 China

13 ³Shanghai Key Laboratory of Translational Medicine on Ear and Nose
14 Diseases, Shanghai, China

15 ⁴Shanghai Institute of Precision Medicine, Ninth People's Hospital, Shanghai
16 Jiao Tong University School of Medicine, Shanghai, China

17 ⁵Department of Otorhinolaryngology-Head & Neck Surgery, Shanghai
18 Children's Hospital, Shanghai Jiao Tong University, Shanghai, China

19 ⁶Lead Contact

20 ⁷Equal contributions

21 *Correspondence: lei.song@yale.edu (L.S.) and yunfeng.hua@shsmu.edu.cn
22 (Y.H.)

23

24 **Summary**

25 Auditory brainstem response (ABR) serves as an objective indication of
 26 auditory perception at given sound level and is nowadays widely used in
 27 hearing function assessment. Despite efforts for automation over decades,
 28 hearing threshold determination by machine algorithm remains unreliable and
 29 thereby still rely on visual identification by trained personnel. Here, we
 30 described a procedure for automatic threshold determination that can be used
 31 in both animal and human ABR tests. The method terminates level averaging
 32 of ABR recordings upon detection of time-locked waveform through cross-
 33 correlation analysis. The threshold level was then indicated by a dramatic
 34 increase in the sweep numbers required to produce “qualified” level
 35 averaging. A good match was obtained between the algorithm outcome and
 36 the human readouts. Moreover, the method varies the level averaging based
 37 on the cross-correlation, thereby adapting to the signal-to-noise ratio of single
 38 sweep recordings. These features empower a robust and fully automated
 39 ABR test.

40

41 Introduction

42 The auditory brainstem responses (ABRs) are brain electrical potential
 43 changes due to synchronous neuronal activities evoked by suprathreshold
 44 acoustic stimuli (Jewett et al., 1970). These responses are detectable using
 45 non-invasive surface electrodes placed on the scalp of test subject and
 46 thereby widely used for hearing function assessment. In rodent and cat,
 47 typical ABR waveform is composed of initial five peaks in the early onset of
 48 sound evoked potentials, representing synchronous activities arising from
 49 projections along the auditory ascending pathway including auditory nerve,
 50 cochlear nucleus, superior olivary complex, lateral lemniscus and inferior
 51 colliculus, respectively (Henry, 1979; Melcher et al., 1996), whereas human
 52 has slightly different peak generators as demonstrated with intracranial
 53 recordings (Moller and Jannetta, 1983) and neuromagnetic responses
 54 (Parkkonen et al., 2009). Thus, ABR wave latencies and amplitudes provide
 55 clinical-significant information, for instance site of lesions or tumors in the
 56 auditory system (Lewis et al., 2015; Roeser et al., 2007) based on how the
 57 properties of ABR waveform are altered.

58 Although the ABR test itself is an objective measurement, the
 59 determination of threshold involves human interpretation of the ABR
 60 waveform. The readout of ABR threshold requires a trained personnel to
 61 supervise waveform recognition, which is labor-demanding. Besides, such
 62 interpretations oftentimes are subjective and may introduce errors that vary
 63 from person to person depending on his/her skill and experience, especially in
 64 cases with atypical waveform or with high background noise (Vidler and
 65 Parkert, 2004). In order to accurately detect mild hearing threshold elevation

66 in the diagnosis of progressive hearing loss (Barreira-Nielsen et al., 2016),
67 hidden hearing loss (Kujawa and Liberman, 2009; Mehraei et al., 2016; Ridley
68 et al., 2018; Sergeyenko et al., 2013), age-related hearing loss (Gates and
69 Mills, 2005; Sergeyenko et al., 2013) and tinnitus (Bramhall et al., 2018;
70 Castaneda et al., 2019), unbiased automatic approaches with high precision
71 and reliability are essential, particularly when screening is involved. Over
72 decades, many attempts were made to automate the procedure, including (1)
73 quantification of the waveform similarity by comparing either existing
74 templates (Davey et al., 2007; Elberling, 1979; Valderrama et al., 2014) or
75 based on features learned by artificial neural network from human annotation
76 (Alpsan and Ozdamar, 1991; McKearney and MacKinnon, 2019); (2)
77 quantification of the waveform stability by cross-correlation function between
78 single-sweeps (Bershad and Rockmore, 1974; Weber and Fletcher, 1980),
79 interleaved responses (Berninger et al., 2014; Xu et al., 1995) or responses at
80 adjacent stimulus levels (Suthakar and Liberman, 2019); (3) the 'signal
81 quality' through scoring procedures like F-ratios (Cebulla et al., 2000; Don and
82 Elberling, 1994; Elberling and Don, 1984; Sininger, 1993). However, due to
83 heterogeneity in inter-subject waveform and signal-to-noise-ratio (SNR)
84 introduced by variations in test subject conditions, electrode
85 placement/impedance, as well as acquisition settings, the accurate threshold
86 determination is only possible under a narrow range of experimental settings,
87 which hampers direct comparisons of ABR results across laboratories.

88 In this study, we proposed an automated approach for real-time ABR
89 threshold determination. Instead of using a prefixed sweep number, level
90 averaging was instructed based on the outcomes of cross-correlation analysis

91 during ongoing sweep acquisition. Near-threshold stimuli feature a sharp
92 increase in the end sweep number upon the detection of time-locked
93 response. We further explored the potential of using this feature for the
94 threshold determination in human subjects and the algorithm outcomes were
95 validated by the human readouts of the same ABR level representations.
96

97 **Results**

98 **Termination of On-going Averaging by Cross-correlation Analysis**

99 Auditory brainstem responses (ABRs) registered by surface electrodes are
100 embedded in high-level background activities and system noise. Smooth
101 baseline and clear waveform, if present, are obtained after averaging over
102 hundreds of sweeps. The number for averaging, however, depends on the
103 amplitude of the evoked response, which varies between test subjects due to
104 variations in, for instance, skull sizes, electrode impedances and placement
105 that determine the distance from the peak generator and the vector projection
106 to the electrodes. Within one recording session, these experimental
107 parameters are usually fixed and the baseline activities at different stimulus
108 levels are comparable, allowing signal-to-noise-ratio (SNR) quantification. It is
109 expected that strong responses from high level stimuli quickly reach confident
110 SNR level, while weak response evoked by low level stimulus requires more
111 averaging and for subthreshold recordings the SNR cannot be improved by
112 level averaging (**Figure 1A**).

113 Based on above notions, we designed a procedure to test whether the
114 change in sweep number is required for the average response at different
115 stimulus levels to reach a certain SNR level. Such change can produce
116 unbiased confident threshold reading. In detail, at a given stimulus level
117 recorded sweeps are loaded into three memory buffers (**Figure 1B**, yellow
118 boxes) and cross-correlation coefficients (CCs) are computed between two
119 out of three group averages (green boxes). If time-locked response,
120 irrespective of wave latencies and shapes, is present, two group averages
121 overlap with each other at time zero, and thereby the obtained CC peak is

found within a small time shift (L) from zero (magenta boxes). Three parallel runs (red box) can effectively reject false positives caused by overlapped random noise with similar peak latencies. Next, the correlation analysis is iterated with increasing sweep numbers (the inner loop). Once the correlation analysis identifies a time-locked response, the iteration will end with sweep number noted. Upon the absence of a time-locked response, the upper limit for iteration count (N) is used to terminate nonproductive attempts. Finally, the outer loop is implemented to scan the threshold response with decreasing stimulus levels and the stop command is triggered upon two consecutive levels that reached the iteration limit.

132

133 **ABR Threshold Determination in Mouse**

To test whether the proposed algorithm could determine the ABR threshold reliably, we recorded from ten mice (three wild-type adult C57BL/6 mice of normal hearing, two wild-type adult CBA mice experienced noise exposure and five telomerase knock-out mice with early-onset hearing loss; data are pooled in this study) single-sweep ABR sets starting from 90 dB (sound pressure level, SPL) to 0 dB with a step size of 5 dB. The raw data were corrected for baseline fluctuations through a smoothing spline fit before being processed by the algorithm.

Figure 2 presents a walkthrough of the procedure. First, an example of level series is plotted (**Figure 2A**) and the visually identified threshold denoted at 30 dB SPL with an asterisk. In the algorithm, three level averages (**Figure 2B**) were used to compute the correlation coefficient (CC). The changes in CC peak amplitude (**Figure 2C**) and the corresponding signal lag

(**Figure 2D**) upon different stimulus levels were then obtained to further explore whether the two derivatives can predict the threshold. Upon reducing stimulus strength, a monotonic decrease was observed in the CC peak amplitudes, whereas near-threshold levels feature a sudden jump from small absolute lags (≤ 2 data points or less than ± 0.082 ms for suprathreshold levels) to big absolute lags with large variation at subthreshold levels. This result suggested that the signal lag is a better response detector at near-threshold levels, thus justified its use in the algorithm to demarcate the threshold boundary.

Next, to test whether the algorithm could find the hearing threshold in real time, we grouped 60 sweeps as a batch then increment the group averaging by iterations (multiply by 60 for end sweep number) until positive results, which were scored by triple cross-correlation analysis, were returned by the algorithm. As shown in **Figure 2E**, the normalized count was small at the suprathreshold levels, while increased rapidly at near-threshold levels and reached its upper limit at subthreshold levels. Therefore, the estimated threshold was defined above the stimulus level at which the iteration upper limit was reached (the highest subthreshold level). Further attempt was made to model the change in iteration counts at near-threshold levels. We acquired an ABR dataset with 1-dB step size (**Figure S1A**) and fitted with both exponential and sigmoidal functions (**Figure S1B**). The visually identified threshold was found approximately at the stimulus level which yield 1.0 and 0.9 on the best-fitted exponential or sigmoidal function. In addition, we found the algorithm outcome was not strongly influenced by either the allowed lag for response detection (**Figure S2A**) or the applied iteration upper limit

(**Figure S2B**) in the cross-correlation analysis, suggesting that the threshold determination is robust and does not require fine parameter tuning.

174

175 **ABR Threshold Determination in Human**

176 To test whether the automated approach applied to human ABR, we acquired
177 ABR datasets from eight human participants. Because intermediate single
178 sweeps are not available on the commercial device, alternatively we provided
179 the algorithms with recorded average responses over different sweep
180 numbers (see **METHOD DETAILS**). Example average responses (**Figure**
181 **3A**), as well as group averages (**Figure 3B**), are shown at tested stimulus
182 levels (60 dB SPL to 0 dB with a step size of 5 dB). The obtained CC peak
183 amplitudes and the signal lags are plotted as a function of the stimulus levels
184 (**Figure 3C and 3D**). Note that 500 sweeps were added per iteration; the
185 upper limit of iteration count was set to seven (3500 sweeps) and a larger
186 decision boundary of lag (≤ 6 data points or less than ± 0.3 ms) for positive
187 responses was implemented due to broader waveforms in human ABR
188 evoked by click-sound. As shown in **Figure 3E**, the iteration count increases
189 fast to reach the upper limit near the visually identified threshold level.

190

191 **Comparison between Expert and Algorithm Determined Thresholds**

192 In order to evaluate the performance of our method, we recruited five human
193 experts to assess the same ABR sets independently and compared their
194 readouts of the thresholds to the algorithm outcomes (**Table 1**). Both
195 approaches reached similar conclusions for both mouse and human ABR
196 (**Figure 4A and 4B**), validating reliability of the algorithm. Moreover, averaging

197 over varied sweep numbers as used in the algorithm does not affect the
 198 threshold determination by either machine or human (**Figure S3A and S3B**).
 199 By further quantifying the minimum sweep number required by the algorithm
 200 (comparing results from the fixed number for all stimulus levels, 420 for
 201 mouse and 3500 for human ABR), we concluded that the algorithm avoided
 202 $66.72 \pm 4.98\%$ and $43.19 \pm 12.48\%$ of nonproductive sweeps, respectively
 203 (**Figure 4C and 4D**).

204

205 Discussion

206 Over decades several statistical approaches have been proposed to
 207 automatically detect ABR waveforms. Cross-correlation is one of the most
 208 favorite methods, as it can detect temporally stable waveforms with high
 209 sensitivity and robustness in a template-free fashion. This is crucial for
 210 recognizing ABRs, which oftentimes present large intersubject variability in
 211 terms of waveform shape and latency, from different levels of background
 212 noise. However, prior attempts using this approach relied on arbitrary decision
 213 boundary for response detection, for instance, minimum required CC
 214 (Berninger et al., 2014; Bershad and Rockmore, 1974; Suthakar and
 215 Liberman, 2019; Weber and Fletcher, 1980) or maximum allowed latency shift
 216 (Galbraith and Brown, 1990; Xu et al., 1995). Preselection of such criteria can
 217 be problematic in practice because even for the same test subject,
 218 comparable SNR across recording sessions is not guaranteed due to
 219 variabilities in electrode impedance and placement, as well as level averaging
 220 settings preferred by individual experimenter. Thus, it is unlikely that a
 221 universal response decision boundary can be applied on all ABR sets without
 222 introducing detection error, barring its usage in cases like cross-institution
 223 collaboration efforts where data pooling is needed.

224 Our approach, which is fundamentally different from the existing
 225 approaches, determines the hearing threshold based on the relative change in
 226 the sweep numbers required for positive ABR detection at different stimulus
 227 levels by the cross-correlation. The resulted signal lag of CC peak has proven
 228 to be a reliable criterion for detecting the time-locked responses with high
 229 sensitivity, but in principle, other quantifications like CC peak amplitude

(**Figure 2C**) or single-point F-distribution (data not shown) can also be used in the algorithm. At near-threshold levels, a rapid increase in the end sweep number was observed (**Figure 2E** and **3E**). It is not surprising because in order to reach confident SNR, small responses require increased averaging for baseline noise reduction, which means that the end sweep numbers quantitatively noted by the algorithm reflect the relative change in the SNR at different recording levels. Besides, this approach does not heavily rely on the selection of detection parameters. First, sub- and supra-threshold level representations feature large differences in the obtained signal lags (**Figure 2D**), which offers a wide range of decision boundary selection of lag without affecting detection accuracy (**Figure S2A**). Second, raising upper limit of the end sweep number only leads to a small shift from the estimated threshold (**Figure S2B**) due to the exponential increase of averaging at near-threshold levels (**Figure S1B**),

So far, the precision of threshold determination is constrained by the step size of level sampling. Modeling the obtained end sweep number change (**Figure S1**) and interpolating the level representation may produce higher precision. Further development of this approach is to combine with level sampling strategy including progressively reduced step size (Cebulla and Sturzebecher, 2015) and increased batch size of sweeps per iteration at near-threshold levels so that the model fitting can be improved by more effective data points in the transition.

In mouse ABR, the proposed method was proven reliable in threshold determination with a mean difference of 1.84 dB ($n = 10$) and maximum discrepancy of ± 5 dB between the algorithm outcomes and human readouts

(**Figure 4A**). Although the detection accuracy is comparable to that of the most up-to-date approach (Suthakar and Liberman, 2019), our algorithm has the advantage of not requiring a calibrated preset criterion and has more robust outcome. Validation with human ABR resulted in only slightly worse detection accuracy (mean and maximum difference, 4.58 and ± 10 dB; **Figure 4C**) than that of mouse ABR. Moreover, the ABR sets with average responses over varying sweep numbers as in the algorithm do not seem to introduce additional difficulty for visual threshold determination (**Figure S3A and S3B**). This is not surprising because excessive level averaging at suprathreshold levels is not expected to improve the threshold determination. This feature is extremely attractive for two reasons. First, it provides minimal quality control for unambiguous waveform recognition for both human and machine. Such standardized data collection will benefit artificial-intelligence based approaches by improving the quality of training data (McKearney and MacKinnon, 2019). Second, when to terminate averaging without compromising the quality of the recording is an important decision making during ABR recording (Don and Elberling, 1996; Madsen et al., 2018), this method has the potential to instruct the efficient test by avoiding nonproductive recordings at supra-threshold levels.

Limitations of the study

This is a proof-of-principle study for detecting time-locked responses in cumulatively averaged level representations using cross-correlation analysis. As level averaging is terminated earlier for supra-threshold levels, the algorithm should be able to shorten the test duration. However, quantification

280 of the exact saved time in a real setting requires implementation of the
281 algorithm on an optimized hardware, which is currently under engineering.

282

283 **Acknowledgments**

284 We thank Dr. Hao Wu for the support of this study. We thank Drs. Guangming
285 Chen and Lin Liu for contributing $terc^{-/-}$ mice. We thank Eric Song for
286 comments on the manuscript. This study was supported by Shanghai
287 Huangpu District Industry Support Fund (XK2019011 to Y.H.), the National
288 Science Foundation of China (81770995 to L.S., 81700903 to B.L. and
289 81800901 to Y.H.) and the Shanghai Key Laboratory of Translational Medicine
290 on Ear and Nose diseases (14DZ2260300).

291

292 **Author contributions**

293 Y.H. designed the study; Y.H. and L.S. supervised the study; B.L., K.H., X.W.
294 and Z.W. contributed the human ABR datasets; Y.L. and Y.Q. performed the
295 mouse ABR recording; H.W. and Y.H. wrote the algorithm and analyzed the
296 data with the help of Y.L., Y.Q., H.S., J.Z. and L.S.; Y.H. wrote the manuscript
297 with the help of H.W. and L.S.

298

299 **Declaration of interests**

300 The authors declare no competing interests.

301

302 **STAR METHODS**

303 **RESOURCE AVAILABILITY**

304 **Lead Contact.** Further information and requests for resources and reagents
305 should be directed to and will be fulfilled by the Lead Contact, Y.H.
306 (yunfeng.hua@shsmu.edu.cn).

307 **Materials Availability Statement.** This study did not generate new unique
308 reagents.

Data and Code Availability Statement. The codes were written in MATLAB scripts and included in the supplementary materials. All single sweep ABR datasets supporting the current study will be made available upon publication: <https://data.mendeley.com/datasets/4yb9772dff/draft?a=d2d509bc-2a09-426b-a1c8-25ef9c817455>

314

315

316 **EXPERIMENTAL MODEL AND SUBJECT DETAILS**

317 C57BL/6 and CBA mice were purchased from Sino-British SIPPR/BK Lab
318 Animal Ltd. (Shanghai, China). The telomerase-knock-out mice (*terc*^{-/-}) were
319 gifts from Prof. Lin Liu (Nankai University, China) and bred in house. Human
320 participants were recruited by Hearing and Speech Center, Shanghai Ninth
321 People's Hospital and consent forms were signed before the experiment. This
322 study was conducted at the Ear Institute and the Hospital Hearing and
323 Speech Center. All procedures were reviewed and approved by the
324 Institutional Authority for Laboratory Animal Care (HKDL2018503) and the
325 Hospital Ethics Committee for Medical Research (SH9H-2019-T79-1).

326

327 **METHOD DETAILS**

328 **ABR Recording**

329 Mouse ABRs were recorded via a TDT RZ6/BioSigRZ system (Tuck-Davis
330 Tech. Inc., US) in a sound-proof chamber as previously described ([Lin et al., 2019](#)). In brief, 7-week-old animals were anesthetized through intraperitoneal
331 injection of Chloral hydrate (500 mg/kg). During the recording, animal body
332 temperature was maintained at 37°C using a regulated heating pad (Harvard
333

334 Apparatus, US) with a rectal thermal probe placed under the body. Evoked
335 potentials were registered via subdermal needle electrodes (Rochester
336 Electro-Med. Inc., US) placed at the animal's vertex (active electrode), left
337 infra-auricular mastoid (reference electrode) and right shoulder region (ground
338 electrode). 3-ms tone pips at 16 kHz were delivered via an MF1 speaker
339 (Tuck-Davis Tech. Inc., US) positioned in the front of the animal 10 cm away
340 from the vertex. Acoustic stimuli were presented 20 stimuli per second and the
341 evoked potentials were collected at 24 kHz sampling rate. Artifact rejection
342 level was set to < 35% (mean rejection voltage 20.5 μ V). Sound level series
343 started from 90 to 0 dB sound pressure level (SPL) with 5-dB step size. For
344 one animal, the stimulus level series were repeated from +10 to -10 dB SPL
345 around the estimated threshold with 1-dB step (**Figure S1**).

346 Human ABRs were recorded by a commercial ABR device (Intelligent
347 Hearing Systems, US) with Smart EP software from eight volunteers aged 23-
348 73 years without the knowledge of their medical conditions. Sound stimulation
349 (100 μ s duration, rectangular envelopes) was generated and presented
350 monaurally through ER3 insert earphones with foam tips. Stimuli were
351 presented at a rate of 37.1/sec with alternating polarity. Electrode impedance
352 was < 5 k Ω and inter-electrode impedance was within $\pm$ 1 k Ω . The artifact
353 rejection level was < 31% (rejection voltage 31 μ V) to exclude contaminations
354 from EEG and myogenic potentials. The evoked potentials were collected with
355 20 kHz sampling rate and $\times$ 100,000 amplification. The bandpass filter was set
356 at 100-3000 Hz. Average responses over 500, 1000, and 2000 sweeps were
357 acquired and repeated three times for the level series starting from 60 to 0 dB
358 SPL with 5-dB step size.

359

360 **Cross-correlation Analysis of Mouse ABR**

361 Sweeps were loaded in three memory buffers. Cross-correlation (MATLAB
362 Central File Exchange Function *xcorr*, MathWorks, US) was computed from
363 two out of three group averages. This resulted in correlation coefficients as a
364 function of signal lag between two group averages). As the ABRs were time-
365 locked to the stimulus onset, they could be detected based on a neglectable
366 signal lag where the maximal coefficient was found (Xu et al., 1995). In this
367 study, maximum allowed lag (L) for a true ABR signal was set to ± 2 data point
368 from time zero (equivalent to ± 0.082 ms or 1% of the analyzed temporal
369 window). Three such tests were implemented in parallel in order to minimize
370 false positives caused by coincidently overlapped background activities
371 through rejecting inconsistent lag values. Moreover, the obtained correlation
372 coefficient peak amplitude was included as an independent variable (**Figure**
373 **2C**).

374 The test started with the loudest stimulus level (90 dB SPL) and was
375 reduced by a step size of 5 dB. At each given stimulus level, the sweep
376 number used to produce the group average was increased by iterations and
377 the end sweep number was noted upon detected response determined by the
378 cross-correlation analysis. For each iteration, sweeps were added in batches
379 (60 sweeps) and an upper limit was set (420 sweeps) to avoid nonproductive
380 attempts at subthreshold levels. Upon detected responses, the iteration was
381 then switched to lowered stimulus level, until the upper limit was reached for
382 two consecutive nonproductive trails. The estimated threshold was the lowest
383 applied stimulus level at which the upper iteration limit was not yet reached.

The transition of the end sweep numbers between supra- and subthreshold levels was modeled (**Figure S1**). Both sigmoidal (1) and exponential functions (2) were employed to fit the relationship between the normalized iteration count C' (equivalent to the sweep number) and the stimulus level S using a nonlinear least square method in MATLAB (Curve Fitting Toolbox, MathWorks, US). In the functions, $\alpha_1 = 0.6$ and $\alpha_2 = 0.2$ were fixed for calibrated lag criterion ($L = 2$), whereas β_1 and β_2 were obtained by fitting.

$$\text{Sigmoid model: } C'(S) = \frac{1}{1 + e^{\alpha_1(S - \beta_1)}} \quad (1)$$

$$\text{Exponential model: } C'(S) = e^{-\alpha_2(S - \beta_2)} \quad (2)$$

Cross-correlation Analysis of Human ABR

For human ABR, group average responses were recorded sequentially and used directly as inputs of the algorithm. Group averages over 500, 1000 and 2000 sweeps were recorded, whereas averages over 1500, 2500, 3000 and 3500 sweeps could be obtained by linear combination (3) where $E\{m\}$, $E\{n\}$ and $E\{m + n\}$ denote the time averages over m , n and $m + n$ sweeps, respectively.

$$E\{m + n\} = \frac{m \cdot E\{m\} + n \cdot E\{n\}}{m + n} \quad (3)$$

The maximum allowed lag for a true response was set to ≤ 6 data points from time zero (equivalent to ± 0.3 ms or 2% of the analyzed temporal window). The iteration upper limit was seven, corresponding to 3500 sweeps. The estimated threshold was the lowest stimulus level with a detectable response.

409 **Threshold Determination by Human Readout**

410 To estimate the ground-truth thresholds of the recorded mouse and human
 411 ABRs, average responses of all level series were provided to five clinicians to
 412 report the visually identified thresholds independently. The test subject
 413 identities were blinded to the judges. Either the fixed sweep number (the
 414 conventional averaging) or the varying numbers used in the algorithm (the
 415 algorithm averaging) was applied to compute the level averaging. The
 416 thresholds were determined by three out of five execution judges (with the
 417 highest and the lowest value excluded). The readouts were used to evaluate
 418 the accuracy of the algorithm outcomes (**Table 1**).

419 **Figure titles and legends**

420 **Figure 1.** Principle and algorithm design for automated threshold
 421 determination. (A) Typical level representation of ABRs. To obtain stable
 422 waveform, one needs less level averaging of sweep recordings at strong
 423 stimulus (green) than that of weak response (magenta), whereas level
 424 averaging cannot improve the SNR of subthreshold recordings containing no
 425 response (blue). (B) Algorithm flowchart. The test starts with the highest
 426 stimulus level (e.g., 90 dB SPL for mouse ABR) and a reset iteration count.
 427 Recorded sweep batches (e.g., 60 sweeps for mouse ABR) are cumulatively
 428 averaged in three data buffers (yellow boxes, $E\{A\}$, $E\{B\}$ and $E\{C\}$). Cross-
 429 correlation operation ($xcorr$) is carried out in each two out of three group
 430 averages (green boxes), which yields three CC peaks and their corresponding
 431 signal lags. If the absolute lag is less than the allowed value ($L \leq 2$ data points
 432 for mouse ABR), the outcome (magenta boxes, C1–C3) returns true,
 433 suggesting a positive response. In cases of all three positive outcomes (red
 434 box), the procedure descends to lower stimulus level (red line, the outer loop),
 435 otherwise the same stimulus level will be repeated with more sweeps (the
 436 inner loop, blue line). An iteration upper limit (blue box, e.g., $N = 7$ for mouse
 437 ABR) is implemented to jump out the inner loop upon subthreshold stimulus
 438 and flag no response.

439

440 **Figure 2.** Cross-correlation analysis of mouse ABR. (A) Example level
 441 representation (averaging over 420 sweeps) of a mouse. The visually
 442 identified threshold level was 20 dB SPL (asterisk). (B) Group averages
 443 generated for cross-correlation analysis. (C) Plot of the obtained CC peak

444 amplitude versus the level series. (D) Plot of signal lag at the CC peak versus
 445 the level series. At suprathreshold levels (dots, ≥ 20 dB), small signal lags
 446 were consistent across three parallel runs (0.33 ± 0.56 data points, mean $\pm$
 447 s.d.), while at subthreshold levels (cycles, < 30 dB) large absolute values and
 448 variability were observed (54.92 ± 39.36 data points, two-sample t-test: ****p
 449 $< 10^{-5}$). (E) Result from a test run of the algorithm. The iteration counts were
 450 plotted as a function of the level series. Detection of time-locked responses (L
 451 ≤ 2 data points) required more iterations of averaging at suprathreshold (black
 452 dots, 1.40 ± 1.06 , mean $\pm$ s.d.) than subthreshold stimulus levels (cycles, $7 \pm$
 453 0 , mean $\pm$ s.d.; two-sample t-test, ****p $< 10^{-5}$). After two consecutive hits at
 454 the iteration upper limit ($N = 7$, dash line), the algorithm flags no response for
 455 the applied stimulus level (cycles) and triggered a stop command to avoid
 456 nonproductive attempts with weaker stimuli. The algorithm-determined
 457 threshold matches visual identification.

458

459 **Figure 3.** Threshold determination of human ABR by cross-correlation
 460 analysis. (A) Example level representations (averaging over 3500 sweeps) of
 461 a human participant with the visually identified threshold at 15 dB (asterisk).
 462 (B) Group averages that used in the algorithm for computing the cross-
 463 correlation. (C) Plot of the obtained CC peak amplitudes versus the level
 464 series. (D) Plot of signal lag at the CC peak versus the level series. At
 465 suprathreshold levels (dots, ≥ 15 dB), small mean value of the signal lags was
 466 obtained (1.30 ± 1.84 data points, mean $\pm$ s.d.), whereas the subthreshold
 467 level (< 15 dB, cycles) features a significantly large and variable lag value
 468 (23.00 ± 24.93 data points, two-sample t-test, ****p $< 10^{-5}$). (E) Result from an

algorithm test run. The iteration counts were plotted as a function of the level series. Detection of the true responses by the cross-correlation analysis ($L \leq 6$ data points) within the iteration upper limit ($N = 7$, dash line) flags the suprathreshold levels (dots), which is consistent with those identified visually.

Figure 4. Comparisons between the thresholds determined by machine and human. (A) For mouse datasets, a close match was found between the algorithm-determined thresholds and those averaged from three out of five independent human readouts (maximum and mean discrepancies, 5 dB and 1.83 ± 2.45 dB, mean $\pm$ s.d.). Linear fit: adjust $R^2 = 0.99$. (B) Comparison between the total sweeps used in the conventional level averaging (420 sweeps at each level, left bar) and in the algorithm (with varying numbers, right bar). The latter requires $66.72 \pm 4.98\%$ (mean $\pm$ s.d.) fewer sweeps. Note that the sweeps were counted from all suprathreshold and two successive sub-threshold levels. (C) As for human datasets, similar thresholds were reported by both the algorithm and human readout (maximum and mean discrepancies, 10 dB and 4.58 ± 3.88 dB, mean $\pm$ s.d.). Linear fit: adjust $R^2 = 0.93$. (D) The total number of sweeps used in the conventional level averaging (left bar) versus in the algorithm (right bar). The latter requires $43.19 \pm 12.48\%$ (mean $\pm$ s.d.) fewer sweeps.

Table titles and legends

Table 1. Comparison between algorithm outcome and human readout

493 REFERENCES

- 494 Alpsan, D., and Ozdamar, O. (1991). Brain-Stem Auditory Evoked-Potential
495 Classification by Backpropagation Networks. 1991 Ieee International Joint
496 Conference on Neural Networks, Vols 1-3, 1266-1271.
- 497 Barreira-Nielsen, C., Fitzpatrick, E., Hashem, S., Whittingham, J., Barrowman, N.,
498 and Aglipay, M. (2016). Progressive Hearing Loss in Early Childhood. *Ear Hear* 37,
499 e311-321.
- 500 Berninger, E., Olofsson, A., and Leijon, A. (2014). Analysis of click-evoked auditory
501 brainstem responses using time domain cross-correlations between interleaved
502 responses. *Ear Hear* 35, 318-329.
- 503 Bershad, N.J., and Rockmore, A.J. (1974). On estimating signal-to-noise ratio using
504 the sample correlation coefficient. *IEEE Trans Inf Theory IT-20*, 112-113.
- 505 Bramhall, N.F., Konrad-Martin, D., and McMillan, G.P. (2018). Tinnitus and Auditory
506 Perception After a History of Noise Exposure: Relationship to Auditory Brainstem
507 Response Measures. *Ear Hear* 39, 881-894.
- 508 Castaneda, R., Natarajan, S., Yule Jeong, S., Na Hong, B., and Ho Kang, T. (2019).
509 Electrophysiological changes in auditory evoked potentials in rats with salicylate-
510 induced tinnitus. *Brain Res*.
- 511 Cebulla, M., and Sturzebecher, E. (2015). Automated auditory response detection:
512 Further improvement of the statistical test strategy by using progressive test steps of
513 iteration. *Int J Audiol* 54, 568-572.
- 514 Cebulla, M., Sturzebecher, E., and Wernecke, K.D. (2000). Objective detection of
515 auditory brainstem potentials: comparison of statistical tests in the time and frequency
516 domains. *Scand Audiol* 29, 44-51.
- 517 Davey, R., McCullagh, P., Lightbody, G., and McAllister, G. (2007). Auditory
518 brainstem response classification: a hybrid model using time and frequency features.
519 *Artif Intell Med* 40, 1-14.
- 520 Don, M., and Elberling, C. (1994). Evaluating residual background noise in human
521 auditory brain-stem responses. *J Acoust Soc Am* 96, 2746-2757.
- 522 Don, M., and Elberling, C. (1996). Use of quantitative measures of auditory brain-
523 stem response peak amplitude and residual background noise in the decision to stop
524 averaging. *The Journal of the Acoustical Society of America* 99, 491-499.
- 525 Elberling, C. (1979). Auditory electrophysiology. The use of templates and cross
526 correlation functions in the analysis of brain stem potentials. *Scand Audiol* 8, 187-
527 190.
- 528 Elberling, C., and Don, M. (1984). Quality estimation of averaged auditory brainstem
529 responses. *Scand Audiol* 13, 187-197.
- 530 Galbraith, G.C., and Brown, W.S. (1990). Cross-correlation and latency compensation
531 analysis of click-evoked and frequency-following brain-stem responses in man.
532 *Electroencephalogr Clin Neurophysiol* 77, 295-308.
- 533 Gates, G.A., and Mills, J.H. (2005). Presbycusis. *Lancet* 366, 1111-1120.
- 534 Henry, K.R. (1979). Auditory brainstem volume-conducted responses: origins in the
535 laboratory mouse. *J Am Aud Soc* 4, 173-178.
- 536 Jewett, D.L., Romano, M.N., and Williston, J.S. (1970). Human auditory evoked
537 potentials: possible brain stem components detected on the scalp. *Science* 167, 1517-
538 1518.
- 539 Kujawa, S.G., and Liberman, M.C. (2009). Adding insult to injury: cochlear nerve
540 degeneration after "temporary" noise-induced hearing loss. *J Neurosci* 29, 14077-
541 14085.

542 Lewis, J.D., Kopun, J., Neely, S.T., Schmid, K.K., and Gorga, M.P. (2015). Tone-burst
543 auditory brainstem response wave V latencies in normal-hearing and hearing-
544 impaired ears. *J Acoust Soc Am* 138, 3210-3219.

545 Lin, X., Li, G., Zhang, Y., Zhao, J., Lu, J., Gao, Y., Liu, H., Li, G.L., Yang, T., Song,
546 L., *et al.* (2019). Hearing consequences in Gjb2 knock-in mice: implications for
547 human p.V37I mutation. *Aging (Albany NY)* 11, 7416-7441.

548 Madsen, S.M.K., Harte, J.M., Elberling, C., and Dau, T. (2018). Accuracy of averaged
549 auditory brainstem response amplitude and latency estimates. *Int J Audiol* 57, 345-
550 353.

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

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

556 Melcher, J.R., Guinan, J.J., Jr., Knudson, I.M., and Kiang, N.Y. (1996). Generators of
557 the brainstem auditory evoked potential in cat. II. Correlating lesion sites with
558 waveform changes. *Hear Res* 93, 28-51.

559 Moller, A.R., and Jannetta, P.J. (1983). Interpretation of brainstem auditory evoked
560 potentials: results from intracranial recordings in humans. *Scand Audiol* 12, 125-133.

561 Parkkonen, L., Fujiki, N., and Makela, J.P. (2009). Sources of auditory brainstem
562 responses revisited: contribution by magnetoencephalography. *Hum Brain Mapp* 30,
563 1772-1782.

564 Ridley, C.L., Kopun, J.G., Neely, S.T., Gorga, M.P., and Rasetshwane, D.M. (2018).
565 Using Thresholds in Noise to Identify Hidden Hearing Loss in Humans. *Ear Hear* 39,
566 829-844.

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

569 Sergeyenko, Y., Lall, K., Liberman, M.C., and Kujawa, S.G. (2013). Age-related
570 cochlear synaptopathy: an early-onset contributor to auditory functional decline. *J*
571 *Neurosci* 33, 13686-13694.

572 Sininger, Y.S. (1993). Auditory brain stem response for objective measures of hearing.
573 *Ear Hear* 14, 23-30.

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

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

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

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

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

588

Figure S1. Modeling of the change in the algorithm iteration counts. (A) Average level representations acquired from a mouse by stimulations with a step size of 1 dB. The visually identified threshold was 24 dB (asterisk). (B) The normalized iteration counts versus level series is fit by both sigmoid (solid line) and exponential functions (dash line). The visually identified threshold is approximately at the level corresponding to 1.0 and 0.9 of the best-fit exponential and sigmoid growth, respectively.

Figure S2. Selection of detection parameters has limited influence on the threshold determination. (A) The algorithm outcomes of allowing different signal lags for ABR detection on the same mouse dataset. Despite of changes in the iteration counts, the resulted threshold did not drift. (B) The algorithm outcomes of using different iteration count upper limit. The sub-threshold levels flagged by reaching the upper limit were not altered.

Figure S3. Varying level averaging used in the algorithm does not affect the threshold determination by visual identification. (A) Level averages of mouse ABR datasets based on the algorithm iteration counts were provided to human experts to readout the threshold. The results were consistent to those determined by the algorithm (maximum and mean discrepancy, 5.00 dB and 3.50 ± 2.33 dB, mean $\pm$ s.d.). Linear fit: adjust $R^2 = 0.97$. (B) Similar results were obtained from both human readouts and the algorithm on the level averages of human ABR dataset over varying sweep number (maximum and mean discrepancy, 15 dB and 5.21 ± 5.00 dB, mean $\pm$ s.d.). Linear fit: adjust $R^2 = 0.78$.