Effects of High-Intensity Airborne Ultrasound Exposure on Behavioural and Electrophysiological Measures of Auditory Function

Samuele Carcagno[1)], Andrew Di Battista[2)], Christopher J. Plack[1,3)]
[1)] Department of Psychology, Lancaster University, Lancaster, LA1 4YF, United Kingdom
[2)] Ultrahaptics Ltd., The West Wing, Glass Wharf, Bristol, BS2 0EL, United Kingdom
[3)] Manchester Centre for Audiology and Deafness, School of Health Sciences, University of Manchester, Manchester, M13 9PL, United Kingdom

Summary

Regulations on safe ultrasound exposure limits are based on a very limited number of studies, which have only considered audiometric threshold shifts as indicators of hearing deficits. The purpose of the current study was to assess the effects of exposure to high-intensity ultrasound on a range of measures of hearing function, which included audiometric thresholds, as well as subclinical measures of hearing deficits: speech-in-noise understanding, supra-threshold auditory brainstem response wave I amplitude and latency, and frequency following response levels to amplitude modulated (AM) tones. Changes in these measures were assessed before and after exposure of the left ear to high-intensity ultrasound in a group of nine young listeners. These changes were compared to those observed in a control group of nine young listeners. Exposure consisted in the presentation of a 40-kHz AM tone at levels of 105, 110, 115, and 120 dB SPL for 10 minutes at each level, plus an exposure to a 40-kHz unmodulated tone during an ultrasound detection task, for a total duration of 50 seconds. None of the measures of hearing function was found to change significantly more for the left compared to the right ear, for participants of the exposure group compared to control participants. Electroencephalographic recordings obtained during exposure to the AM tone did not show significant phase-locked activity at the modulation frequency or at low-frequency subharmonics of the ultrasound tone. One out of nine participants was able to perform the ultrasound detection task above chance level, although due to limitations of the experimental setup the mechanism by which she could detect the presentation of the tone remains unclear.

1 Introduction

There are several sources of airborne ultrasound (US - sound with frequencies > 20 kHz) to which the general public may be exposed, such as public address voice alarm systems, and pest deterrents [1, 2, 3]. There is also an increasing interest in the use of airborne US for the development of virtual haptic displays that can deliver tactile sensations in mid-air. These haptic displays can be used to augment the interaction with touchscreen interfaces, for example by creating virtual buttons or sliders above a touchscreen interface [4, 5]. The Ultrahaptics system consists of an array of transducers, positioned on a flat board, that generate virtual haptic displays by projecting high-intensity US at focal points in mid air; interaction of the user's unadorned hands with these focal points generates tactile sensations [4]. In order to generate tactile sensations airborne US needs to be projected on focal points on the skin at levels of around 145 dB SPL. Ambient levels at the ear will vary considerably depending on the distance and the orientation of the head of the user. With the US speakers placed at arms-length distance from the ears, the Leq obtained while rotating and translating the head across several positions has been estimated to be $\sim$ 120 dB SPL in the absence of hand interaction with the speakers [6]. Actual user-case exposure, with hand interaction, would be expected to be lower.

There are a number of international standards and guidelines setting maximum permissible levels (MPLs) for US exposure to prevent potential adverse effects (reviewed in [7]). However, several shortcomings of the existing standards and guidelines have been pointed out recently [1, 8]. These include, but are not limited to, 1) the fact that they are almost exclusively restricted to occupational exposures, 2) the fact that they are based on sparse datasets, 3) the fact that they do not take into consideration the higher high-frequency sensitivity of some subsets of the population, such as young adults and children, who may thus not be sufficiently protected by the existing guidelines, 4) the fact that some of these standards were developed to prevent hearing threshold shifts, but not other adverse effects, such as annoyance or inability to concentrate. Moreover, there are currently no international standards for measuring US exposure in the work environment [9].

An additional limitation of the current guidelines for hearing protection is that they are based on studies which measured only audiometric threshold shifts as an indicator of hearing loss. In several rodent species it has been shown that noise exposure can cause a permanent loss of synapses between the inner hair cells and auditory nerve fibres. This deafferentiation of the auditory nerve can occur in the absence of a permanent threshold shift (PTS) [10, 11, 12]. This syndrome has been referred to as "cochlear synaptopathy", and is associated in animal models with a reduction of wave I of the auditory brainstem response (ABR) at high stimulus levels, as well as with a reduction of the frequency following response (FFR) to high-frequency (∼1-kHz) amplitude modulation [13]. In humans, however, the results of a number of observational studies have not found a clear association between noise exposure (measured with either retrospective questionnaires or presumed on the basis of occupational status), and neural or behavioural measures of cochlear synaptopathy [14, 15]. In any case, it would be desirable to check that levels of US exposure that do not cause audiometric threshold shifts, do not also cause subclinical hearing losses that cannot be measured by the audiogram.

There are different mechanisms by which airborne US could generate auditory sensations. Some of these mechanisms may operate only at certain sound frequencies and/or levels. At the lower range of the US spectrum, up to about 28 kHz, it is possible that US directly excites the most basal cochlear filters [16, 17]. At levels exceeding about 120 dB SPL, audible subharmonics of the US frequency may be generated by the tympanic membrane or by the cochlea [18, 19, 20, 21, 22]. US with frequencies ranging from ∼25–60 kHz could also be transmitted from the eye to the inner ear via intracranical fluid conduction at levels as low as ∼ 100 dB SPL [23]. Excessive exposure to US could therefore damage cochlear structures, including inner and outer hair cells, and the synaptic connections between the inner hair cells and auditory nerve fibres, in a way similar to low-frequency noise [24], but at the cochlear places excited by the US stimulation.

The main aim of the current study was to test the hypothesis that short exposures to US, at typical levels that may reach the ear of a user while interacting with the Ultrahaptics system, cause subclinical hearing deficits in young normal-hearing listeners. This hypothesis was tested by measuring auditory function, before and after exposure of the left ear to US, with a test battery that included, besides audiometric thresholds in the clinical frequency range, wave I of the ABR, the FFR to amplitude modulated (AM) tones, speech perception in noise (SPiN) thresholds, and extended high-frequency audiometry. Differential left-right ear post-exposure changes in these measures were compared to those of a control group of participants who were not exposed to US. Two post-exposure assessments were made, one on the day immediately after the exposure, and one about a week after the exposure, to check for either temporary or permanent changes in the hearing measures. Additionally, we attempted to measure behavioural detection thresholds for the 40-kHz tone produced by the Ultrahaptics system, and analysed electroencephalographic (EEG) recordings obtained during exposure to a 40-kHz AM tone, to look for traces of phase-locked neural activity at the modulation frequency, and at subharmonics of the tone. Finally, we collected subjective reports of nausea, headaches, or other possible adverse subjective symptoms immediately after the US exposure.

2 Methods

This study was approved by the Lancaster University Faculty of Science and Technology Research Ethics Committee. The methods for this study were pre-registered on OSF: https://osf.io/pgvdj/. A few deviations and additions from the pre-registered protocol have been noted in the supplementary materials (SM).

A diagrammatic timeline of the experimental sessions is shown in Figure 1. Each session was performed on a different day. The average delay between the first (S1) and the second (S2) assessment session was similar between the exposure (6.33 days, sd=1.94) and control (5.89 days, sd=3.52) groups. The average delay between S2 and the third assessment session (S3) was also similar between the exposure (7.12 days, sd=0.83) and control (7.89 days, sd=1.17) groups. The delay between the US exposure (S-US) and the S2 session was, for all participants of the exposure group, of one day.

Each of the assessment sessions lasted about 2 hours, including short breaks between the tests. The S-US session lasted about 1.5 hours. The order of the tests in the assessment sessions was always the same, starting with the measurement of audiometric thresholds (including the extended high-frequency region), SPiN thresholds, ABR recording, and FFR recording. The S-US session, which was attended only by participants of the exposure group, started always with the behavioural US detection test, followed by the EEG recording during US exposure. All testing took place in double-walled IAC (IAC Acoustics, Winchester, UK) soundproof booths. Details of all the tests will be given in the sections below.

2.1 Participants

A total of 24 native British English participants were recruited for the study from the student population at Lancaster University. An otoscopic examination was performed prior to the beginning of the tests, and six

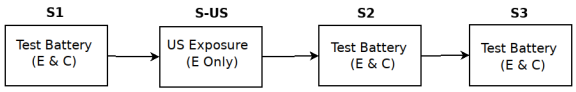

Figure 1: Timeline of the experimental sessions. The headings on top of each box indicate the session label. The text inside the box indicates the type of tests run in each session, and the group of participants tested (experimental: E; control: C). Sessions S1, S2, and S3 included the test battery for the evaluation of auditory function, and were performed by participants of both the experimental, and control group. US exposure took place during the S-US session, which was performed by participants of the experimental group only.

participants had to be excluded from the study due to the presence of wax occlusion in one or both ears. The remaining 18 participants (all females) were randomly assigned to either the exposure group (n=9, mean age=21 years, sd=1.5), or to the control group (n=9, mean age=21 years, sd=1.7). One participant of the exposure group was unable to attend the second post-exposure assessment session. Her data from the other sessions were nonetheless included in the analyses.

Participants were asked to limit exposure to loud noise on the day prior to each session by avoiding attendance to concerts or other loud venues. Although participants were not asked about previous exposure to US, given the fact that they were all students, and none spontaneously reported occupational US exposure, it is unlikely that their previous exposure to US would be different from that of the general population.

2.2 Assessment sessions - Behavioural tests

2.2.1 Audiometry

Audiometric thresholds were measured for pure tones at octave frequencies from 0.125 to 8 kHz (clinical frequency range) as well as for pure tones at 12 and 16 kHz (extended high-frequency range). The tones had a duration of 200 ms, including 10-ms cosine-raised onset and offset ramps. Thresholds were measured with a two-interval two-alternative forced-choice (2I-2AFC) paradigm. The presentation level of each tone was varied adaptively using a two-down one-up transformed up-down procedure tracking the 70.7% correct point on the psychometric function [25] to determine its detection threshold. On each trial the tone was randomly presented during one of two observation intervals marked by flashing lights on the computer screen, and separated by a 500-ms silent interval. Participants were asked to indicate the interval in which the sound occurred by pressing the corresponding button on a numeric keypad. Feedback was provided at the end of each trial by means of a coloured light on the computer screen.

A single block of trials was run for each combination of ear and frequency (in random order). Each block was terminated after 16 turnpoints of the adaptive track. The level was varied in 4-dB steps for the first four turnpoints, and by 2 dB for the remaining turnpoints. Threshold was estimated as the average of the last 12 turnpoints. The pure tones were synthesized with a sampling rate of 48 kHz, and 32-bit depth, were played through a E-MU 0204 USB sound card (E-MU Systems, Scotts Valley, U.S.A.), and presented via Sennheiser HDA300 headphones (Sennheiser electronic GmbH & Co. KG, Hanover, Germany)

2.2.2 Speech-in-noise reception

Speech-in-noise understanding was assessed using the digit triplets test (DTT) [26]. On each trial the listener was presented with three digits in the 1–9 range, but excluding 7 (the only digit consisting of two syllables). No repetitions of the same digit were allowed in a trial. The digits were voice recordings of a male speaker taken from McShefferty *et al.* [27]. A speech-shaped-noise with a root mean square (RMS) level of 65 dB SPL was presented throughout the duration of the trial. The level of the speaker's voice was varied adaptively using a one-down one-up transformed up-down procedure to determine the speech-reception threshold at the 50% correct point on the psychometric function [25]. Each trial started with the recording of a female voice saying the phrase "the digits", and was followed by the presentation of the digits spoken by the male voice. Participants were asked to input the three digits they heard, or give their best guess if they could not hear them clearly, using a numeric keypad. Responses with repeated digits within the same sequence were not allowed. Feedback was provided at the end of each trial by means of a coloured light on the computer screen.

A block of trials was terminated after 16 turnpoints. The target level was changed in 2-dB steps for the first four turnpoints, and by 1 dB for the remaining turnpoints. Threshold estimates for each block of trials were based on the average of the last 12 turnpoints. Participants completed two blocks of trials for each ear (first one block for each ear, in random order; then a second block for each ear, in random order). The participants' thresholds were estimated as the average of the threshold estimates obtained in each of the two blocks of trials for each ear. The recordings of the digits had a 48-kHz sampling rate and 16-bit depth. They were digitally mixed with the speech shaped noise, played through a E-MU 0204 USB sound card, and presented via Sennheiser HD650 headphones.

2.3 Assessment sessions - EEG tests

For these tests the EEG was recorded with the Biosemi ActiveTwo system (BioSemi B.V., Amsterdam, The Netherlands). Gold-plated active electrodes were used. One electrode was attached on the middle of the forehead, just below the hairline, one on the neck, at the level of the 7^{th} cervical vertebrae, and one on each earlobe. The common mode sense and driven right leg electrodes were attached on the forehead. The EEG signal was acquired at a sampling rate of 16.384 kHz with 24-bit resolution. Stimuli were generated with a sampling rate of 48 kHz and 32-bit resolution, were played through a 24-bit RME Hammerfall DSP multiface DAC (RME Intelligent Audio Solutions, Germany), and presented via mu-metal shielded ER3A Etymotic insert earphones (Etymotic Research Inc., Elk Grove, U.S.A.). Triggers marking the start of a stimulus were sent to the Biosemi receiver from additional channels of the soundcard after being transformed to discrete pulses by a custom-built device.

2.3.1 Auditory brainstem response

The ABR was recorded in response to 100-μs, 100-dB ppeSPL clicks in rarefaction polarity. The clicks were presented at a rate of 14.1 per second, with alternate presentation between the left and right ear. A total of 10,000 clicks were presented (5,000 to each ear).

The EEG was bandpass filtered offline between 0.1 and 1.5 kHz [28] with a 256-taps zero-phase-shift finite impulse response (FIR) filter. The triggers marking click onsets were adjusted to compensate for the 0.9-ms delay introduced by the earphones tubing, and the EEG was then segmented into discrete epochs relative to the onset of the clicks using a -2 to 12 ms time window. The forehead channel was re-referenced to the ipsilateral earlobe channel. All the analyses were performed using this montage. The segments were baseline corrected by subtracting the mean amplitude during the 2-ms pre-stimulus window, and averaged using the iterative weighted averaging algorithm [29]. The ABR wave I peaks were identified using an automatic peak-picking procedure which is described in the SM. Log-transformed peak amplitudes [30, 31] were used in the statistical analyses.

2.4 Frequency following response

The FFR was recorded in response to two simultaneous AM tones with carrier frequencies of 0.59 and 2 kHz, and modulation frequencies of 93.3, and 124.4 Hz, respectively. Each tone was presented at a level of 75 dB SPL. The tones were embedded in pink noise to reduce the contribution of high-spontaneous rate fibres to the recorded FFRs. The pink noise was presented at a spectrum level of 40 dB SPL *re.* 100 Hz, in a frequency region from 20 to 3000 Hz, with notches two equivalent rectangular bandwidths [32] wide around the carrier frequencies so as to form three noise bands (20–506, 683–1773, and 2253–3000 Hz). The stimuli had a duration of 450 ms, including 10-ms onset and offset raised-cosine ramps. Two-thousand stimuli were generated (1,000 for each ear; half with the tones in condensation, and half with the tones in rarefaction polarity), each with a fresh noise sample, and saved on disk. FFRs were collected in a single block of trials with the 2,000 stimuli presented in a random order. The inter-stimulus interval was jittered between 25 and 75 ms.

The EEG was bandpass filtered offline between 0.06 and 1 kHz with a 256-taps zero-phase-shift FIR filter. The triggers marking stimulus onsets were adjusted to compensate for the 0.9-ms delay introduced by the earphones tubing, and the EEG was then segmented into discrete epochs relative to the onset of the stimuli using a -5 to 450 ms time window. The forehead channel was re-referenced to the neck channel. All the analyses were performed using this vertical montage. The segments were baseline corrected by subtracting the mean amplitude during the 5-ms pre-stimulus window, and averaged using the iterative weighted averaging algorithm [29].

Spectral analyses were used to determine the level, in dB, of the FFR at each modulation frequency. The waveforms were windowed using a hamming window, and the waveform spectra were computed via fast Fourier transforms (FFTs). The signal level was estimated by the power at the FFT bin closest to the signal frequency.

2.5 Ultrasound Tests

US tones were presented using the Ultrahaptics array as a loudspeaker source pointing straight towards the left ear from a position to the left, and slightly to the front (angle $\sim$ 25 degrees), of the participant, at a distance of $\sim$ 112 cm from the left ear.

There are particular challenges associated with measuring SPL at ultrasonic frequencies. Complex field patterns are formed meaning deviations in microphone position of a few centimetres can have extreme effects [33]. Moreover, the complex interaction of the sound field with the head, torso and pinna mean that free-field measurements do not provide a full and meaningful picture of individual exposure levels [34]. For these reasons we adopted a calibration procedure that would provide meaningful output with the potential to be replicated by other researchers in different labs. Moreover, in the interest of safety, any deviation from accepted free-field measurements should result in an overestimate of SPL.

The level of the US tones was calibrated with a Brüel & Kjær (Nærum, Denmark) type 4191 microphone fitted in the ear of a Brüel & Kjær type 4100 head and torso simulator (HATS). The microphone

grid formed a flush boundary at the entrance to the ear canal of the HATS pinna. Therefore all US SPLs presented in the article are estimates of SPLs at the eardrum of the participant. The HATS was positioned on the chair where the participants would be performing the tests. The positions of the chair, and of the US speaker arrays were fixed throughout the experiment (they were the same during the calibration procedure and when participants were tested). Microphone data were acquired through a Picoscope (DrDAQ, Pico Technology) which was programmed for real-time SPL measurement with a 1/3 octave band filter centred at 40 kHz. A digital equalization filter was implemented to convert the free-field response of the microphone to a pressure-field response. Thus, all SPLs represent an estimate of actual incident pressure at the interface. SPL values were exponentially time weighted using a 1s time constant; equivalent to the 'SLOW' setting on a standard SPL meter.

Various 40-kHz US tones were presented, varying the signal voltage in order to find the voltage values that would result in SPLs of 100, 105, 110, 115, and 120 dB. This procedure was repeated with the HATS placed in three slightly different orientations: looking straight ahead, with the head slightly tilted towards the right, and straight ahead with the torso propped up by about 15 cm. The measurements for the first position (looking straight ahead) were repeated twice after repositioning the HATS. Each of these four datasets was fitted with a function to estimate the relation between voltage and output level. The RMS error between the recorded SPLs and the ones predicted by the estimated functions across the four function fits was 1.7 dB. The difference in the SPLs predicted by the function fits for the two datasets obtained with the HATS looking straight ahead was 2.44 dB. The maximum difference in the SPLs predicted by the function fits across the four datasets was 4.77 dB. These data indicate that slight changes to the position of the head of the participants would result in level changes of around 5 dB, or less. Because we intended to present US tones close to the MPLs set by the International Labour Office [35] we chose to calibrate on the fits obtained with the HATS position that predicted the highest SPLs (head slightly tilted towards the right), so that deviations in the position of the head from this reference position would result in slightly lower SPLs rather than in higher SPLs.

During all of the US tests, the right ear of the participant was plugged with a 3M E-A-R classic soft foam earplug (3M Company, Maplewood, U.S.A.), so that only the left ear would be exposed to high-intensity US. Tests conducted at Ultrahaptics indicate that, properly fitted, these earplugs provide about 30 dB of attenuation at 40 kHz. To investigate the possibility that US exposure could elicit adverse subjective effects, at the end of the US tests participants were presented with the following written question: "Have you experienced dizziness, loss of balance, feeling sick, headaches, or a feeling of pressure/fullness in the ears during the test? If yes, please specify which symptoms you have experienced".

2.5.1 Behavioural ultrasound detection

The ability to detect a 500-ms 40-kHz tone was assessed using a 2I-2AFC paradigm. On each trial the tone was randomly presented during one of two observation intervals marked by flashing lights on the computer screen, and separated by a 500-ms silent interval. Participants were asked to indicate the interval in which the tone occurred by pressing the corresponding button on a numeric keypad. Feedback was provided at the end of each trial by means of a coloured light on the computer screen. A hybrid adaptive/constant procedure [36, 37] was used: The presentation level of the tone was initially varied adaptively using a two-down one-up transformed up-down procedure to determine its detection threshold. However, the presentation level was limited to a maximum of 120 dB SPL, if the adaptive track reached this level at any time (including the initial turnpoints) the adaptive track was terminated early, and the procedure switched to a constant one to estimate the proportion of correct responses at the maximum level of 120 dB SPL until a total of 50 trials at this level had been completed. Otherwise the block was terminated after 16 turnpoints. The tone level was initially set at 110 dB, and was changed in 4-dB steps for the first four turnpoints, and by 2 dB for the remaining turnpoints. If the track converged, the threshold was estimated as the average of the last 12 turnpoints. Each participant completed two blocks of trials.

The electronic board of the US speakers generated a noise below the ultrasound frequency range when the speakers were playing. This noise was clearly audible, and its level increased when the output level of the US tone increased. In an attempt to prevent listeners from responding to this noise rather than to the US tone, a 34-sec sample of the noise generated by the speakers was recorded with a Zoom Q3HD (Zoom, Tokyo, Japan) portable recorder and played back to mask the noise generated by the US speakers. The masking noise was lowpass filtered at 16 kHz, and 100, 2.5-sec samples drawn at random starting points from the 34 sec recording were extracted. The spectra of the noise recording and the masker are shown in Figure S1 in the SM. The masker samples were played back through two JBL 305P MkII speakers (JBL Professional, Northridge, U.S.A.) symmetrically placed around the US speakers, at a level at the listener's left ear of 71 dB C-weighted during each trial. The masking noise started 0.5 seconds before the start of each trial, ended 0.5 seconds after the end of each trial, and was gated on and off with 50-ms raised-cosine onset and offset ramps.

The level of the masking noise was established by preliminary tests during which the first author (SC) ran several blocks of the US detection task varying the masker level across blocks in 10-dB steps to find the level at which he was performing the task at chance level. The level of the masker for the main experiment was set 30 dB above the level at which SC was performing at chance level. SC had normal hearing in the clinical frequency range up to 8 kHz for both ears at the time of the tests. No data are available on SC's hearing sensitivity above 8 kHz, but it should be noted that, given that he was 37 years old at the time of the tests, it is unlikely that he would have been able to hear not only the 40-kHz tone, but also its first subharmonic.

2.5.2 EEG recordings during ultrasound presentation

The EEG was acquired in response to a 40-kHz US tone amplitude modulated at a rate of 124.4 Hz. Four blocks of trials, in which the level of the tone was either 105, 110, 115, or 120 dB SPL were run. It was not possible to send a trigger with sub-millisecond accuracy from the US speaker array to the EEG system. For this reason, during each block a single US tone was presented continuously for 10 minutes. The four blocks were randomly ordered. The EEG acquisition settings and electrode configurations used for this test were the same as the ones used for the EEG tests in the assessment sessions, and described in Section 2.3.

The FFR is largest for tones with frequencies around 500 Hz, and can only be recorded for tones with frequencies below about 2,000 Hz [38]. For this reason we limited the analysis to subharmonics 6 to 8 of the 40-kHz carrier (corresponding to frequencies of 625, 312.5, and 156.25 Hz), as well as to the modulation frequency of 124.4 Hz. The EEG was bandpass filtered offline between 60 and 1,000 Hz with a 256-taps zero-phase-shift FIR filter. The forehead channel was re-referenced to the neck channel. All the analyses were performed using this vertical montage.

To improve the signal-to-noise ratio (SNR) the continuous recordings were split into shorter segments which were then averaged. To ensure that the phase of the signal of interest was coherent across segments, the 10-min recording was split into consecutive segments. Four different segmentations were performed with segment durations of 1, 2, 4, or 5 seconds, so that an integer number of cycles would fit into a segment for signal frequencies of 625, 312.5, 156.25, and 124.4 Hz, respectively. For each segmentation the segments were then separately averaged. The resulting waveforms were windowed using a hamming window, and the waveform spectra were computed via FFTs. For each of the target signal frequencies the level of the signal and of the noise were estimated from the FFT obtained from the corresponding segmentation procedure. The signal level was estimated by the power at the FFT bin closest to the signal frequency. The noise level was estimated by summing the power of 1-Hz bands above and below the signal bin, but excluding a 2-Hz band above and below the signal frequency to minimize the effects of spectral leakage on the noise estimate.

2.6 Statistical analyses

Statistical analyses were run in R [39] by means of Welch two-sample t-tests. The tests were specified in the pre-registered protocol. There are four main families of tests corresponding to the research questions described in the introduction:

- Does US exposure (at the levels, frequencies, and durations used in this study) have any *temporary* effects on hearing function as assessed by behavioural and psychophysical measures?
- Does US exposure (at the levels, frequencies, and durations used in this study) have any *permanent* effects on hearing function as assessed by behavioural and psychophysical measures?
- Are 40-kHz US tones detectable behaviourally?
- Are 40-kHz AM US tones detectable from FFR recordings?

For the first research question, for each measure of interest we first computed the difference between values obtained for the left and the right ears, then the difference of the resulting values between S2 and S1. These between-session changes of between-ear differences were the dependent variables that were compared between the exposure and control groups by means of t tests. The measures of interest were:

- Average audiometric threshold across the clinical audiometric range (0.125–8 kHz; $\mathrm{PTA}_{0.125-8}$).
- Average audiometric threshold across the extended high-frequency range (12 and 16 kHz; PTA_{12-16}).
- SPiN threshold.
- Log-transformed wave I ABR amplitude.
- Wave I ABR latency.
- FFR level at the modulation frequency for the 0.59-kHz carrier ($\mathrm{FFR}_{0.59}$).
- FFR level at the modulation frequency for the 2-kHz carrier (FFR_{2}).

We thus performed a total of seven tests, and set the α level to $0.05/7 \simeq 0.007$ using a Bonferroni correction for multiple comparisons, and one-tailed tests because in each case the hypothesis was directional. For

simplicity, for these, and all other tests, uncorrected p values will be reported in the paper; significance should be assessed with respect to the specified α level.

The second research question is analogous to the first one, but involves differences between S3 and S1 (instead of differences between S2 and S1) to test for permanent changes in auditory function. For this research question we thus also performed a total of seven one-tailed t tests, and set the α level to $0.05/7 \simeq 0.007$.

For the third research question, previous studies [16, 17] have shown that there are large interindividual variations in the detectability of ultrasound even within the population of young normal hearing listeners. For some listeners detection thresholds have been measured up to a frequency of 28 kHz, while for other listeners thresholds are already unmeasurable at a frequency of 20 kHz. For this reason we did not run group-level statistical analyses. Instead the detectability of the tone was assessed separately for each listener. Given that this required a total of nine tests across listeners, the α level was set at 0.05/9. Plans on how to assess if detectability was above chance in case one or both blocks of the hybrid adaptive/constant procedure converged to a threshold were given in the pre-registration protocol. However, for none of the listeners did any of the blocks converge to a threshold, so we will only consider the case in which the proportion of correct responses at 120 dB SPL is available for both blocks of trials. Following binomial probability, a listener should get at least 63 correct out of 100 responses to provide evidence of detection at greater than chance level at an α level of 0.05/9.

For the fourth research question, it is also likely that there may be large interindividual differences, hence we did not run group-level statistical analyses. The presence of a signal for each participant, level, and frequency tested can be detected using an $F_{2,2m}$ test [40] where m is the number of bins used to compute the noise power. This test is based on the fact that FFT power estimates have a χ^2 distribution both at the signal frequency, and at neighbouring frequencies. Therefore their ratio can be tested using an F statistic. Because power at the signal frequency is the sum of two independent squared variables (the real and imaginary parts) the signal power estimate is distributed as a χ^2 variable with 2 degrees of freedom, while the noise power estimate, which is obtained by averaging m bins, is a χ^2 variables with $2m$ degrees of freedom. Running a test at each of the four levels, for each of the four target frequencies, and for each of the nine participant required a total of 144 tests, so the α level was set at 0.05/144. The number of noise bins falling in the two 1-Hz bands above and below the target bin varied according to the segment duration, and was 2, 4, 8, or 10 bins, respectively, for segment durations of 1, 2, 4, or 5 seconds. Following the equations in [40] the criterion SNR to detect a significant signal at the α level of 0.05/144 was therefore set at 20.18, 13.86, 11.02, and 10.48 dB for the 625, 312.5, 156.25, and 124.4 Hz signals, respectively.

3 Results

3.1 Audiometry

The audiograms for each participant, session, and ear are shown in Figure S2 in the SM. The average audiograms for each combination of ear, session, and group were close to 0 dB HL, although thresholds tended to be slightly higher at 16 kHz.

Overall the audiograms across S1 and S2 appeared relatively stable for both the control (SM Figure S3), and the exposure (SM Figure S4) group. A few listeners from either group showed apparent losses or gains of sensitivity > 10 dB at one or more frequencies. However, in these cases the standard deviation of the turnpoints of the adaptive track of the block of trials with the highest threshold was almost invariably high, suggesting that these changes were due to a high lapse rate in that block of trials. One listener from the exposure group showed an apparent threshold shift of more than 30 dB at 1 kHz for the exposed ear. However, this listener did not show large threshold shifts for the exposed ear at the other test frequencies. Out of concern for the participant her 1-kHz thresholds were immediately re-tested twice for both the left, and the right ear. The participant was asked to pay full attention to the task before re-testing. Her 1-kHz threshold for the left ear went down from 34 to -20 dB HL in the first repetition, and then to -10 dB HL in the second repetition. Her 1-kHz thresholds for the right ear were quite stable across the three repetitions, around -4.5 dB HL. Given that her thresholds went back to normal in the re-tests it is clear that the apparent threshold shift was a false alarm most likely caused by a high lapse rate. Nonetheless, to avoid bias the data from the two re-tests were not used further. Only the original data with the threshold shift have been analysed and used for the figures and statistical tests reported in the manuscript.

Figure 2 shows, for each group, the difference in average audiometric thresholds across the clinical, and the extended high frequency ranges, between the left and right ear, between session 2 and session 1. A loss of sensitivity for the exposure group in the left (exposed) ear relative to the right ear would manifest as an increase in the threshold difference shown in the figure. Averaged across participants, the threshold differences were small, less than 1.1 dB in absolute value, for both groups in either frequency range. The t-tests comparing the threshold differences between the exposure and the control group did not reveal significant differences either in the clinical range ($t_{(11.987)} = 0.944$, $p = 0.182$), or in the extended high frequency range ($t_{15.57} = -0.331$, $p = 0.628$).

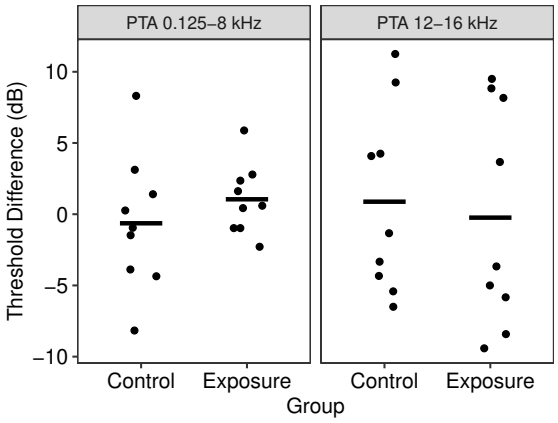

Figure 2: Difference in audiometric thresholds between the left and right ear between session 2 and session 1: $(T_{L2} - T_{L1}) - (T_{R2} - T_{R1})$, where T refers to the threshold, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the threshold difference in the exposure group would indicate a relative post-exposure loss of sensitivity in the left (exposed) ear compared to the right ear. Points plot individual listeners' data. Segments plot group averages.

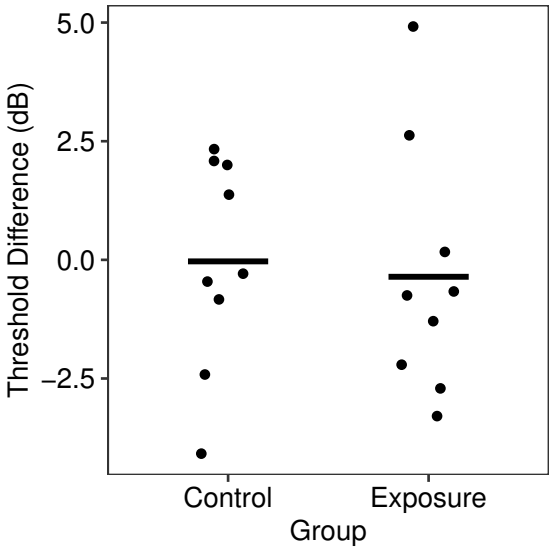

Figure 3: Difference in DTT thresholds between the left and right ear between session 2 and session 1: $(T_{L2} - T_{L1}) - (T_{R2} - T_{R1})$, where T refers to the threshold, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the threshold difference in the exposure group would indicate a relative post-exposure performance drop for the left (exposed) ear compared to the right ear.

The difference in average audiometric thresholds across the clinical and the extended high frequency ranges, between the left and right ear, between session 3 and session 1, are shown for each group in Figure S5. Threshold differences were small, less than 1.5 dB in absolute value, for both groups in either frequency range. The t-tests comparing the threshold differences between the exposure, and the control group did not reveal significant differences either in the clinical range ($t_{(12.247)} = 0.85$, $p = 0.206$), or in the extended high frequency range ($t_{13.761} = 0.248$, $p = 0.404$).

3.2 Speech in noise reception

The DTT thresholds for each participant, session, and ear are shown in Figure S6 in the SM. DTT thresholds were relatively stable across sessions for both the exposure and the control group. Figure 3 shows, for each group, the difference in DTT thresholds between the left and right ear, between session 2 and session 1. A decrement in SPiN for the exposure group in the left (exposed) ear relative to the right ear would manifest as an increase in the threshold difference shown in the figure. Average threshold differences were small, less than 0.5 dB in absolute value, for both groups. The t-test comparing the threshold differences between the exposure and the control group did not reveal a significant difference ($t_{(15.529)} = -0.283$, $p = 0.609$).

The difference in DTT thresholds between the left and right ear, between session 3 and session 1, are shown for each group in Figure S7. Threshold differences were small, less than 1 dB in absolute value, for both groups. The t-test comparing the threshold differences between the exposure and the control group did not reveal a significant difference ($t_{(13.799)} = -0.693$, $p = 0.75$).

3.3 Auditory brainstem response

3.3.1 Wave I ABR amplitude

Figures S8 and S9 in the SM show the ABR waveforms for each participant of the control and exposure group, respectively. ABR grand averages for each group are shown in Figure 4. The wave I ABR amplitudes for each participant, session, and ear are shown in Figure S10 in the SM. ABR amplitudes were remarkably stable across sessions for both the exposure, and the control group. Figure 5 shows, for each group, the average geometric ratio in wave I amplitude between the right and left ear, between session 2 and session 1. A decrement in wave I amplitude for the exposure group in the left (exposed) ear relative to the right ear would manifest as an increase in the amplitude ratio shown in the figure. Average amplitude ratios were close to one, with average amplitudes changing by less than 10% in either direction, for both groups. The t-test comparing the log-transformed amplitude (log-amplitude) differences between the exposure, and the control group did not reveal a significant difference ($t_{(15.887)} = -1.315$, $p = 0.896$).

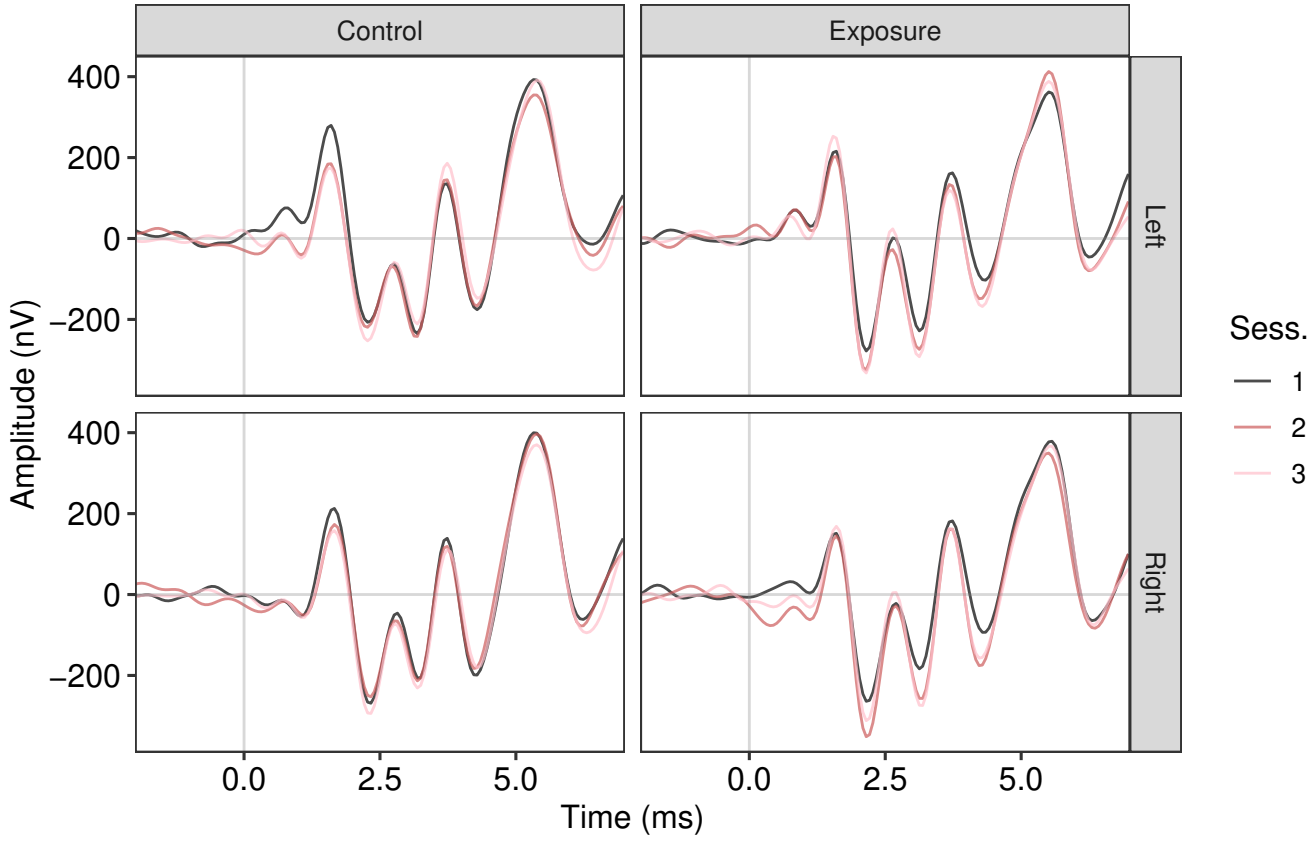

Figure 4: (Colour online) ABR grand averages.

Geometric average wave I amplitude ratios between the right and left ear, between session 3 and session 1 are shown in Figure S11. Average amplitude ratios were close to one. The t-test comparing the log-amplitude differences between the exposure, and the control group did not reveal a significant difference ($t_{(11.165)} = -1.982$, $p = 0.964$).

3.3.2 Wave I ABR Latency

The wave I ABR latencies for each participant, session, and ear are shown in Figure S12 in the SM. ABR latencies were quite stable across sessions for both the exposure, and the control group. Figure 6 shows, for each group, the average difference in wave I latency between the left and right ear, between session 2 and session 1. An increase in wave I latency for the exposure group in the left (exposed) ear relative to the right ear would manifest as an increase in the latency difference shown in the figure. Average latency differences were close to zero for both groups. The t-test comparing the latency differences between the exposure, and the control group did not reveal a significant difference ($t_{(15.436)} = 1.201$, $p = 0.124$).

Average wave I latency differences between the left and right ear, between session 3 and session 1 are shown in Figure S13. Average latency differences were close to zero for both groups. The t-test comparing the latency differences between the exposure, and the control group did not reveal a significant difference ($t_{(11.804)} = 0.321$, $p = 0.377$).

3.4 Frequency following response

The FFR levels for each participant, session, ear, and carrier frequency are shown in Figure S14 in the SM. FFR levels were less stable across sessions than ABR amplitudes for participants of both groups. Figure 7 shows, for each group and carrier frequency, the average differences in FFR levels between the right and left ear, between session 2 and session 1. A decrement in FFR level for the exposure group in the left (exposed) ear relative to the right ear would manifest as an increase in the level difference shown in the figure. Average FFR level differences were in the range of a few dBs, but the variability was large. The t-tests comparing the threshold differences between the exposure and the control group did not reveal a significant difference either at the low ($t_{(13.844)} = 1.208$, $p = 0.124$), or at the high ($t_{(15.885)} = -0.535$, $p = 0.7$) carrier frequency.

Average differences in FFR levels between the right and left ear, between session 3 and session 1 are shown in Figure S15. Average FFR level differences were in the range of a few dBs, but the variability was

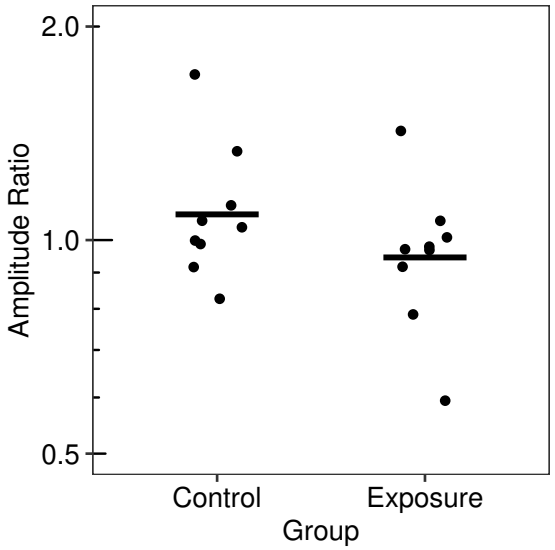

Figure 5: ABR wave I amplitude ratio between the right and left ear, between session 2 and session 1: $(A_{R2}/A_{R1})/(A_{L2}/A_{L1})$, where A refers to the amplitude, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the amplitude ratio in the exposure group would indicate a relative post-exposure wave I amplitude decrease in the left (exposed) ear compared to the right ear.

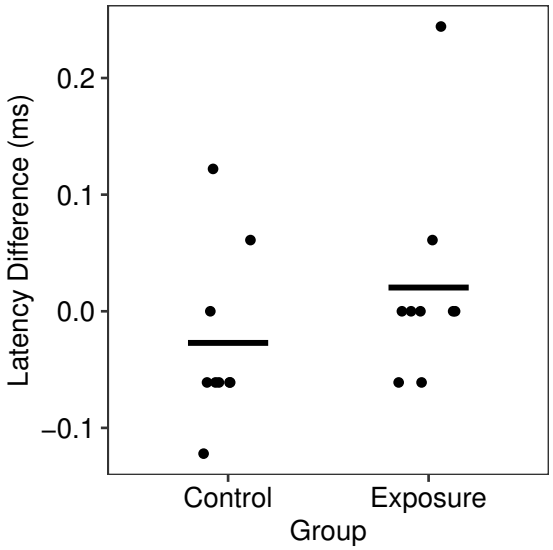

Figure 6: ABR wave I latency difference between the left and right ear, between session 2 and session 1: $(T_{L2} - T_{L1}) - (T_{R2} - T_{R1})$, where T refers to the latency, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the latency difference in the exposure group would indicate a relative post-exposure wave I latency increase in the left (exposed) ear compared to the right ear.

large. The t-tests comparing the threshold differences between the exposure and the control group did not reveal a significant difference either at the low ($t_{(12.844)} = 1.208$, $p = 0.124$), or at the high ($t_{(14.371)} = 0.612$, $p = 0.275$) carrier frequency.

3.5 Confidence intervals

Although the lack of significant differences in the dependent variables measured in this study does not provide evidence of either temporary or permanent effects of US exposure on hearing function, they should not be taken on their own as evidence against this hypothesis. It is useful to look at interval estimates to understand the range of possible effects that the results of the experiment could support. Confidence intervals (CIs) do not necessarily reflect measurement precision, and cannot be generally interpreted as Bayesian credibility intervals covering the $X\%$ most probable values of a parameter of interest [41], although under some assumptions, for simple normal models CIs and credibility intervals are often quite similar [42, 43]. For this reason, besides computing CIs, we also computed Bayesian credibility intervals. Credibility intervals were computed as 99% highest density intervals (HDIs) of the posterior distribution of the parameter of interest [44]. Posterior distributions were obtained by means of Markov Chain Monte Carlo sampling using JAGS [45] and R [39]. The JAGS model code is provided in the SM. The dependent variables (between-session changes of between-ear differences were) were modeled with a normal likelihood function and heterogeneous variances between groups. Priors were vague on the scale of the data. CIs and HDIs for all the tests involving differences between S1 and S2 are shown in Table 1. CIs and HDIs for all the tests involving differences between S1 and S3 are shown in Table S1 in the SM. To be consistent with the one-tailed tests performed in this study, the CIs need to be one sided, and corrected for multiple comparisons. These are provided in the first column of the tables. However, one-sided CIs are unbound on one side; two-sided CIs provide a more intuitive understanding of the uncertainty of the parameters of interest. The second column of the tables provides 99% CIs uncorrected for multiple comparisons (note that an uncorrected two-sided 99% CI is practically quite close to a two-sided 95% CI corrected for seven multiple comparisons using the Bonferroni method). The third column of the tables provides 99% HDIs.

3.6 Behavioural ultrasound detection

For all of the participants the adaptive track reached a level of 120 dB in both blocks of trials, hence the procedure switched in each case to a constant one estimating the proportion of correct responses at 120 dB SPL. This proportion is shown for each participant in Figure S16 of the SM. For eight of the nine participants performance in the task was at chance

Variable	Corrected 95% CI (one-tailed)	Uncorrected 99% CI (two-tailed)	Bayesian HDI
$PTA_{0.125-8}$	-3.42–Inf	-3.77–7.14	-4.76–8.11
PTA_{12-16}	-10.44–Inf	-11.03–8.79	-13.21–11.07
DTT	-3.49–Inf	-3.69–3.04	-4.44–3.75
ABR WI Log-Amp.	-0.43–Inf	-0.45–0.17	-0.52–0.23
ABR WI Lat.	-0.06–Inf	-0.07–0.16	-0.09–0.19
$FFR_{0.59}$	-6.41–Inf	-7.14–16.85	-9.49–19.43
FFR_2	-8.79–Inf	-9.25–6.39	-11.06–8

Table 1: Interval estimates for the changes between S1 and S2 for the dependent measures analyzed in the study. The first column shows 95% one-sided CIs corrected for multiple comparisons. The second column shows uncorrected 99% CIs. The third column shows 99% Bayesian HDIs.

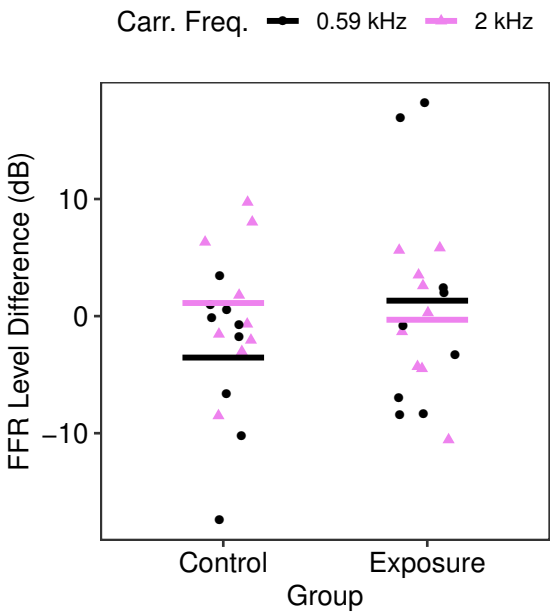

Figure 7: (Colour online) Difference in FFR level between the right and left ear, between session 2 and session 1 $(M_{R2}-M_{R1})-(M_{L2}-M_{L1})$, where M refers to the level, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the level difference in the exposure group would indicate a relative post-exposure decrease in FFR level for the left (exposed) ear compared to the right ear.

level. However, one participant performed clearly above chance level, with 94 out of 100 correct responses.

3.7 EEG recordings to ultrasound

The FFR SNR at the target subharmonic frequencies, and at the modulation frequency of the 40-kHz AM tone for each participant of the exposure group is shown in Figure S17. For none of the participants did the FFR SNR reach the criterion for statistical significance in any condition. Figure S18 shows the across-participant average SNR for each condition. The average SNR was in each case close to zero.

3.8 Subjective effects

Seven participants reported no subjective effects or symptoms after the US exposure session. Two participants reported generic effects likely unrelated to US exposure. The first one (P17) reported feeling "a bit fidgety" during the session, but told the experimenter that this was probably related to having to sit still for the entire duration of the session. The second one (P23) reported a "slight feeling of pressure/fullness in ears – initially when earplugs inserted and then more towards the end". It should be noted that towards the end of the session this participant was exposed to the lowest US levels (105, and 110 dB SPL).

4 Discussion

In this study we assessed the performance of a group of young listeners on a series of behavioural and electrophysiological hearing tests before and after their left ear was exposed to high-intensity US. Their performance changes were compared to those of a control group of listeners who were not exposed to US. Additionally, participants of the exposure group performed behavioural, and electrophysiological tests to assess the detectability of the US to which they were exposed. The results can be summarized as follows:

- We did not find evidence that US exposure, at the levels, frequencies, and durations used in the current study has any temporary, or permanent effects on hearing function as assessed by several psychophysical, and electrophysiological measures.

- Only one out of nine listeners was able to detect the presentation of a 40-kHz 120 dB SPL US tone. Due to limitations of the experimental setup, however, it is unclear whether this listener was able to hear the tone itself, one of its subharmonics, or extraneous level/spatial cues asso-

ciated with the low-frequency noise made by the US speakers when they were playing US.

- We did not find evidence that low-frequency subharmonics (< 1 kHz), or the modulation frequency, of an AM 40-kHz US tone presented at levels ranging from 105 to 120 dB SPL could be detected electrophysiologically using the FFR.

A discussion of each of these points will be presented in the following sections.

4.1 Effects of ultrasound exposure

We did not find evidence of either temporary, or permanent audiometric threshold shifts as a result of exposure to US, which included the presentation of 40-kHz US tones at levels of 105, 110, 115, and 120 dB SPL for 10 minutes at each level. The 99% HDIs for the threshold difference in the clinical and extended high frequency range suggest that, even if US exposure at the levels and durations used in the current study would lead to a temporary threshold shift (TTS), the shift could not be larger than about 11 dB. We are aware of only three published studies, reviewed by Lawton [46], that have investigated the presence of temporary or permanent threshold shifts after exposure to US at similar, or higher levels than those used in the current study. Parrack [47] found that 5-min exposures to US tones between 21 and 37 kHz at levels ranging from 148 to 154 dB SPL caused TTSs at subharmonics of the US frequencies; these TTSs subsided rapidly and did not lead to PTSs. Grigor'Eva [48] failed to find TTSs after one-hour long exposures to a 20-kHz US tone of either 110, or 115 dB SPL. The reports of these two studies do not provide the number of participants tested, nor demographic information. Acton and Carson [49] measured the audiograms of 16 workers before and after a working day which involved exposure to various drills and washers that produced sounds with one-third octave band levels sometimes in excess of 100 dB SPL at ultrasonic frequencies, and below about 90 dB SPL at lower frequencies. Although they found a few large TTSs at individual frequencies for some of the ears tested (6% of the datapoints), because of their random pattern, and the fact that some of the shifts were positive and some were negative, the authors attributed the shifts to measurement variability and did not attach any particular significance to them. No detailed information on the age of the participants tested is provided, except for the fact that most of the men tested had some degree of presbycusis and were older than the women. In an additional study Di Battista [6] did not find evidence of TTSs in a group of 10 participants ranging in age from 24 to 64 years, after 5-min exposures to 40-kHz US tones ranging in level from 100 to 120 dB SPL. Overall, the results of our study are consistent with those of these previous investigations that did not find significant audiometric TTSs after exposure to US up to levels of 120 dB SPL.

In addition to the lack of significant changes in audiometric thresholds, we did not find evidence of effects of US exposure on subclinical measures of hearing function that included DTT thresholds, wave I amplitude and latency measurements, and the level of the FFR to AM tones. Average DTT threshold changes were close to zero, but caution should be exercised in interpreting this result because the 99% HDI for DTT threshold differences is compatible with the possibility of threshold increases after US exposure of up to 3.75 dB. For comparison, the average difference in DTT thresholds between normal hearing and hearing-impaired listeners is about 4 dB [50]. In any case, given that the most important frequency region for speech perception lies below ~ 5 kHz, and that TTSs have been found at most for the third subharmonic of a US tone [47], it seems unlikely that DTT thresholds could be affected as a result of exposure to a 40-kHz US tone. The lack of significant effects of US on the ABR and FFR is potentially more informative, because at high stimulus levels large sections of the cochlea contribute to these responses, and both are greatly affected by the contribution of basal (high-frequency) cochlear sites [51, 52]. The 99% HDI for wave I ABR amplitude suggests that our results would be compatible with potential relatively small wave I log-amplitude reductions of at most 0.23, which corresponds a decrease in amplitude of $\sim 20\%$. For comparison, wave I amplitude reductions as a function of age in a 40-years span have been estimated to be around 38%, after accounting for concomitant reductions due to hearing loss in the 2–4 kHz frequency range [53]. The 99% HDI for ABR wave I latency suggests that our results could be compatible with modest latency changes of at most 0.19 ms. For comparison, wave I latency increases as a function of age in a 40-years span have been estimated to be around 0.25 ms, after accounting for concomitant latency changes due to hearing loss in the 2–4 kHz frequency range [53]. Due to the relatively high variability of FFR levels obtained in this study the 99% HDIs for potential US-exposure related reductions in FFR level are large, and compatible with changes of up to 19.4 dB for the 0.59-kHz carrier, and of up to 8 dB for the 2-kHz carrier.

One limitation of the current study is that the previous history of US exposure of the participants was not known. If potential negative effects of US do not increase linearly with the historical amount of exposure but plateau after a certain threshold, and the exposure history of our participants had reached this threshold, any negative effects would have been missed in our study. Because our participants were recruited from the student population and did not spontaneously report a history of occupational US exposure, it is unlikely that their exposure would be

different from that of the general public. However, quantifying non-occupational US exposure would be very challenging given the increasing number of US sources in public places [1, 2, 3], and given that these US sources are generally inaudible.

4.2 Behavioural detection of ultrasound

Eight listeners were unable to detect a 500-ms 40-kHz tone, presented at a level of 120 dB SPL, but one listener (P14) performed the 2I-2AFC task clearly above chance level with 94% correct responses. Previous studies [16, 17] have shown that some listeners were able to detect US tones up to a frequency of 28 kHz, while none of the listeners tested were able to hear US tones of 30 kHz. A pink noise was used in these previous studies to ensure that participants could not perform the detection task by listening to subharmonics of the US tones. The maximum presentation level in these studies was 110 dB SPL. The presentation level of the US tone used in our study was 10 dB higher. Given that the frequency of our tone was more than 10 kHz higher than the highest detectable frequency in previous studies, it seems unlikely that the higher SPL used in our study would have been sufficient to make a 40 kHz tone detectable, although we cannot rule out this possibility. Two alternative possibilities remain to explain the results of P14. The first one is that for this listener the masker was not sufficiently intense to mask the low-frequency noise produced by the speakers, or the spatial cues arising from the different positions of the US speaker array, and the speakers playing the masker. However, given that the level of the masker was set 30 dB above the level at which the first author, who is highly experienced in psychoacoustics tasks, was performing at chance level, this possibility seems somewhat unlikely. The second possibility is that this listener was able to perform the task by detecting the first subharmonic of the US tone, which would have fallen at a frequency of 20 kHz. Although no 20-kHz component is visible in the spectrum of the recording of the US tone (see Figure S1), at high SPLs subharmonics have been detected in physiological recordings from non-human animals [20]. These subharmonics are thought to be generated mainly by the tympanic membrane in the middle ear, although some may be also generated by the cochlea [21]. Although in humans, subharmonics radiated from the eardrum have only been recorded at levels of at least 140 dB SPL [18], theoretical models predict that levels of ∼120 dB SPL could be sufficient to generate them [19, 22]. Given that the masker used in the current experiment was lowpass filtered at 16 kHz, a subharmonic at 20 kHz would not have been masked and may have been detectable by the listener who performed the detection task above chance level. Her 16-kHz threshold for the left ear, averaged across sessions, was -2.7 dB HL, the second best, and one of the only three <10 dB HL among participants of the exposure group. Hence, this listener would have been more sensitive to the presence of a 20-kHz subharmonic than most other listeners of the exposure group.

4.3 Electrophysiological detection of ultrasound

A number of studies have investigated the effects of ultrasonic stimulation on neurophysiological responses in humans using EEG, magnetoencephalography or neuroimaging techniques. The results have been mixed; some studies have failed to detect cortical activity evoked by US stimuli [54], while other studies, comparing stimuli with and without ultrasonic components, have found differences in the power of certain EEG frequency bands or detected a greater activation of some brain regions in response to stimuli with ultrasonic components using neuroimaging methods [55, 56]. Our study differs from the previous ones because we investigated the detectability of US using the FFR, a steady-state evoked potential response that, if present, contains energy at frequencies harmonically related to those of the stimulus, or generated by nonlinear interactions in the auditory system [38, 57].

Because the FFR can only be detected for frequencies below ∼2 kHz, and for stimuli ∼40–45 dB above perceptual threshold [38], we had a priori low expectations of finding FFRs to the AM US tone employed in this experiment. TTSs have been detected only up the third subharmonic of a US tone, and at levels much higher than those used in the current study. Thus it was unlikely that subharmonics of a 40-kHz tone could be detected in the frequency region below 2 kHz where the FFR can be recorded. Although the 124.4 Hz modulation frequency falls into this frequency region, given that the highest frequency at which US has been detected (while subharmonics were masked) is 28 kHz [16, 17], it is unlikely that even the most basal cochlear filters could be responding to the 40-kHz AM tone components to generate a response at the modulation frequency. Acoustic recordings of the AM US tone showed the presence of a component at the modulation frequency of 124.4 Hz, probably generated by modulation distortions in the air [58]. Although it was not possible to establish the level of this component, its level was likely too low to be detected via the FFR. Overall, the absence of FFRs to the US tone found in our study is not surprising.

4.4 Subjective effects

Only two participants reported minor subjective effects after US exposure, but these were vague and possibly unrelated to US presentation. Sensitivity to US may be limited to a sensitive subset of the population,

and various research reports, reviewed by Leighton [1], indicate that only some people manifest negative symptoms when they are nearby US sources. In our study we did not specifically recruit participants with a history of negative reactions to US sources, and given that our sample size was small it is possible that none of our participants belonged to a subset of the population who may have a heightened sensitivity to US. Adverse reactions to the presence of a US source may be partly psychogenic, and it is unclear to what extent interindividual differences in reactions to US reflect actual differences in hearing sensitivity or psychological differences [59]. It is possible that both play a role depending on the specific frequencies and levels of the US components, that in turn determine their audibility.

All the participants of the exposure group had normal hearing for the exposed ear up to 12 kHz, and only two of them had thresholds slightly above 20 dB HL for the exposed ear at 16 kHz. For this reason we can exclude that the lack of major reactions to US in our study was due to poor high-frequency hearing. Given the high interindividual variability of thresholds for sounds in the ultrasonic frequency range even for young normal hearing listeners [16, 17, 54], it is nonetheless possible that our sample did not include enough participants with sufficient sensitivity to observe major negative reactions to US exposure. Indeed only one of our participants was able to detect the presentation of the US tone, but this participant did not show any negative subjective reactions.

4.5 Conclusions

We did not find evidence of either audiometric threshold shifts or changes of behavioural or electrophysiological subclinical measures of hearing function in a group of young participants exposed to US up to levels of 120 dB SPL, compared to a control group. Our results are consistent with previous studies that did not find audiometric threshold shifts after exposure to US at similar levels. Our sample size was relatively small, consisting of nine participants per group, and caution should be exercised in interpreting the null results. However, analyses of the credibility intervals for the dependent measures suggest that any effects if they existed, would not be large, with the exception of the FFR measures, which were quite variable and did not yield precise estimates.

Acknowledgement

This study was funded by Ultrahaptics Ltd, a company developing ultrahaptic technologies. Author ADB works for Ultrahaptics Ltd. Author CP was supported by the NIHR Manchester Biomedical Research Centre. We would like to thank the Associate Editor and two anonymous reviewers for valuable comments on a previous version of this manuscript. The underlying data in this paper is available from https://osf.io/pgvdj/.

References

[1] T. G. Leighton. Are some people suffering as a result of increasing mass exposure of the public to ultrasound in air? *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 472(2185):20150624, 2016.

[2] M. D. Fletcher, S. Lloyd Jones, P. R. White, C. N. Dolder, B. Lineton, and T. G. Leighton. Public exposure to ultrasound and very high-frequency sound in air. *J. Acoust. Soc. Am.*, 144(4):2554, 2018.

[3] C. N. Dolder, M. D. Fletcher, S. Lloyd Jones, B. Lineton, S. R. Dennison, M. Symmonds, P. R. White, and T. G. Leighton. Measurements of ultrasonic deterrents and an acoustically branded hairdryer: Ambiguities in guideline compliance. *J. Acoust. Soc. Am.*, 144(4):2565, 2018.

[4] T. Carter, S. A. Seah, B. Long, B. Drinkwater, and S. Subramanian. Ultrahaptics: Multi-point mid-air haptic feedback for touch surfaces. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 505–514, New York, NY, USA, 2013. ACM.

[5] G. Wilson, T. Carter, S. Subramanian, and S. A. Brewster. Perception of ultrasonic haptic feedback on the hand: Localisation and apparent motion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1133–1142, New York, NY, USA, 2014. ACM.

[6] A. Di Battista. The effect of 40 khz ultrasonic noise exposure on human hearing. In *Proceedings of the 23rd International Congress on Acoustics*, pages 4873–4788, Berlin, Germany, 2019. Deutsche Gesellschaft für Akustik.

[7] B. W. Lawton. *Damage to human hearing by airborne sound of very high frequency or ultrasonic frequency*, volume Health and Safety Executive Contract Research Report 343/2001. HSE Books, Sudbury, 2001.

[8] T. G. Leighton. Ultrasound in air–Guidelines, applications, public exposures, and claims of attacks in Cuba and China. *J. Acoust. Soc. Am.*, 144(4):2473, 2018.

[9] J. Radosz and D. Pleban. Ultrasonic noise measurements in the work environment. *J. Acoust. Soc. Am.*, 144(4):2532, 2018.

[10] S. G. Kujawa and M. C. Liberman. Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss. *J. Neurosci.*, 29(45):14077–14085, 2009.

[11] H. W. Lin, A. C. Furman, S. G. Kujawa, and M. C. Liberman. Primary neural degeneration in the Guinea pig cochlea after reversible noise-induced threshold shift. *J. Assoc. Res. Otolaryngol.*, 12(5):605–616, 2011.

[12] A. E. Hickox, E. Larsen, M. G. Heinz, L. Shinobu, and J. P. Whitton. Translational issues in cochlear synaptopathy. *Hear. Res.*, 349:164–171, 2017.

[13] L. A. Shaheen, M. D. Valero, and M. C. Liberman. Towards a diagnosis of cochlear neuropathy with envelope following responses. *J. Assoc. Res. Otolaryngol.*, 16(6):727–745, 2015.

[14] N. Bramhall, E. F. Beach, B. Epp, C. G. Le Prell, E. A. Lopez-Poveda, C. J. Plack, R. Schaette, S. Verhulst, and B. Canlon. The search for noise-induced cochlear synaptopathy in humans: Mission impossible? *Hear. Res.*, 377:88–103, 2019.

[15] C. G. Le Prell. Effects of noise exposure on auditory brainstem response and speech-in-noise tasks: a review of the literature. *Int. J. Audiol.*, 58(sup1):S3–S32, 2019.

[16] K. Ashihara, K. Kurakata, T. Mizunami, and K. Matsushita. Hearing threshold for pure tones above 20 kHz. *Acoust. Sci. Technol.*, 27(1):12–19, 2006.

[17] K. Ashihara. Hearing thresholds for pure tones above 16 kHz. *J. Acoust. Soc. Am.*, 122(3):EL52–57, 2007.

[18] H. E. von Gierke. Subharmonics generated in human and anial ears by intense sound. *J. Acoust. Soc. Am.*, 22:675–675, 1950.

[19] W. Pong and W. Marcaccio. Nonlinearity of the middle ear as a possible source of subharmonics. *J. Acoust. Soc. Am.*, 35(5):679–681, 1963.

[20] P. Dallos and C. O. Linnell. Subharmonic Components in Cochlear-Microphonic Potentials. *J. Acoust. Soc. Am.*, 40(1):4–11, 1966.

[21] P. Dallos. On the Generation of Odd-Fractional Subharmonics. *J. Acoust. Soc. Am.*, 40(6):1381–1391, 1966.

[22] S. Huang, W. Dong, and E. S. Olson. Subharmonic distortion in ear canal pressure and intracochlear pressure and motion. *J. Assoc. Res. Otolaryngol.*, 13(4):461–471, 2012.

[23] Martin L Lenhardt. Eyes as fenestrations to the ears: a novel mechanism for high-frequency and ultrasonic hearing. *Int. Tinnitus J.*, 13(1):3–10, 2007.

[24] T. N. Le, L. V. Straatman, J. Lea, and B. Westerberg. Current insights in noise-induced hearing loss: a literature review of the underlying mechanism, pathophysiology, asymmetry, and management options. *J. Otolaryngol. Head Neck Surg.*, 46(1):41, 2017.

[25] H. Levitt. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49(2):467–477, 1971.

[26] C. Smits, T. S. Kapteyn, and T. Houtgast. Development and validation of an automatic speech-in-noise screening test by telephone. *Int. J. Audiol.*, 43(1):15–28, 2004.

[27] D. McShefferty, W. M. Whitmer, I. R. Swan, and M. A. Akeroyd. The effect of experience on the sensitivity and specificity of the whispered voice test: a diagnostic accuracy study. *BMJ Open*, 3(4):e002394, 2013.

[28] M. Granzow, H. Riedel, and B. Kollmeier. Single-sweep-based methods to improve the quality of auditory brain stem responses Part I: Optimized linear filtering. *Z. Audiol.*, 40(1):32–44, 2001.

[29] H. Riedel, M. Granzow, and B. Kollmeier. Single-sweep-based methods to improve the quality of auditory brain stem responses Part II: Averaging methods. *Z. Audiol.*, 40(2):62–85, 2001.

[30] T Sand. Statistical properties of ABR amplitudes and latencies. Implications for computation of reference limits and relation to click phase. *Scand. Audiol.*, 19(3):131–7, 1990.

[31] N. F. Bramhall, D. Konrad-Martin, G. P. McMillan, and S. E. Griest. Auditory brainstem response altered in humans with noise exposure despite normal outer hair cell function. *Ear Hear.*, 38(1):e1–e12, 2017.

[32] B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, 47(1-2):103–138, 1990.

[33] A. Wolff. Airborne ultrasound at german workplaces. In *InterNoise16*, INTER-NOISE and NOISE-CON Congress and Conference Proceedings, pages 5498–5502, Hamburg, Germany, 2016. Institute of Noise Control Engineering.

[34] C. Kling, R. Schöneweiß, C. Ullish-Nelken, and A. Wolff. Airborne ultrasound noise at workplaces. In *Proceedings of the 23rd International Congress on Acoustics*, pages 6333–6337,

Berlin, Germany, 2019. Deutsche Gesellschaft für Akustik.

[35] International Labour Office. *Protection of workers against noise and vibration in the working environment*. ILO Code Practice. International Labour Office, Geneva, 1977.

[36] K. Hopkins and B. C. Moore. Development of a fast method for measuring sensitivity to temporal fine structure information at low frequencies. *Int. J. Audiol.*, 49(12):940–946, 2010.

[37] A. King, K. Hopkins, and C. J. Plack. The effects of age and hearing loss on interaural phase difference discrimination. *J. Acoust. Soc. Am.*, 135(1):342–351, 2014.

[38] A. Krishnan. Frequency-following response. In Robert F Burkard, Jos J Eggermont, and Manuel Don, editors, *Auditory evoked potentials: Basic principles and clinical applications*, chapter 15, pages 313–333. Lippincott Williams, Philadelphia, 2007.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[40] R. A. Dobie and M. J. Wilson. A comparison of t test, F test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion-product otoacoustic emissions, or other sinusoids. *J. Acoust. Soc. Am.*, 100(4 Pt 1):2236–2246, 1996.

[41] R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E. J. Wagenmakers. The fallacy of placing confidence in confidence intervals. *Psychon. Bull. Rev.*, 23(1):103–123, 2016.

[42] S. Greenland and C. Poole. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, 24(1):62–68, 2013.

[43] C. J. Albers, H. A. L. Kiers, and Don van Ravenzwaaij. Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra Psychol.*, 4(1):31, 2018.

[44] John K. Kruschke. *Doing Bayesian data analysis, a tutorial with R, JAGS, and Stan*. Elsevier, London, 2^{nd} edition, 2014.

[45] Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Kurt Hornik, Friedrich Leisch, and Achim Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria, 2003.

[46] B. W. Lawton. Exposure limits for airborne sound of very high frequency and ultrasonic frequency. Technical Report 334, Institute of Sound & Vibration Research, Southampton, 2013.

[47] H. O. Parrack. Effect of air-borne ultrasound on humans. *Int. J. Audiol.*, 5(3):294–308, 1966.

[48] Grigor'Eva V.M. Effect of ultrasonic vibrations on personnel working with ultrasonic equipment. *Sov. Phys. Acoust.*, 11:426–427, 1966.

[49] W.I. Acton and M.B. Carson. Auditory and subjective effects of airborne noise from industrial ultrasonic sources. *Br. J. Ind. Med.*, 24(4):297–304, 1967.

[50] C. G. Le Prell and D. S. Brungart. Speech-in-noise tests and supra-threshold auditory evoked potentials as metrics for noise damage and clinical trial outcome measures. *Otol. Neurotol.*, 37(8):295–302, 2016.

[51] M. Don and J. J. Eggermont. Analysis of the click-evoked brainstem potentials in man unsing high-pass noise masking. *J. Acoust. Soc. Am.*, 63(4):1084–1092, 1978.

[52] T. Dau. The importance of cochlear processing for the formation of auditory brainstem and frequency following responses. *J. Acoust. Soc. Am.*, 113(2):936–950, 2003.

[53] D. Konrad-Martin, M. F. Dille, G. McMillan, S. Griest, D. McDermott, S. A. Fausti, and D. F. Austin. Age-related changes in the auditory brainstem response. *J. Am. Acad. Audiol.*, 23(1):18–35, 2012.

[54] R. Kühler, M. Weichenberger, M. Bauer, J. Hensel, R. Brühl, A. Ihlenfeld, B. Ittermann, T. Sander, S. Kuhn, and C. Koch. Does airborne ultrasound lead to activation of the auditory cortex? *Biomed. Tech. (Berl.)*, 2019.

[55] T. Oohashi, E. Nishina, M. Honda, Y. Yonekura, Y. Fuwamoto, N. Kawai, T. Maekawa, S. Nakamura, H. Fukuyama, and H. Shibasaki. Inaudible high-frequency sounds affect brain activity: hypersonic effect. *J. Neurophysiol.*, 83(6):3548–3558, 2000.

[56] R. Kuribayashi and H. Nittono. High-resolution audio with inaudible high-frequency components induces a relaxed attentional state without conscious awareness. *Front. Psychol.*, 8:93, 2017.

[57] H. E. Gockel, R. Farooq, L. Muhammed, C. J. Plack, and R. P. Carlyon. Differences between psychoacoustic and frequency following response measures of distortion tone level and masking. *J. Acoust. Soc. Am.*, 132(4):2524–2535, 2012.

[58] J. F. Pompei. *Sound from ultrasound: The parametric array as an audible sound source.* PhD Thesis, Massachussets Institute of Technology, 2002.

[59] Mark D Fletcher, Sian Lloyd Jones, Paul R White, Craig N Dolder, Timothy G Leighton, and Benjamin Lineton. Effects of very high-frequency sound and ultrasound on humans. Part II: A double-blind randomized provocation study of inaudible 20-kHz ultrasound. *J. Acoust. Soc. Am.*, 144(4):2521, 10 2018.

Supplementary figures referenced in the main manuscript

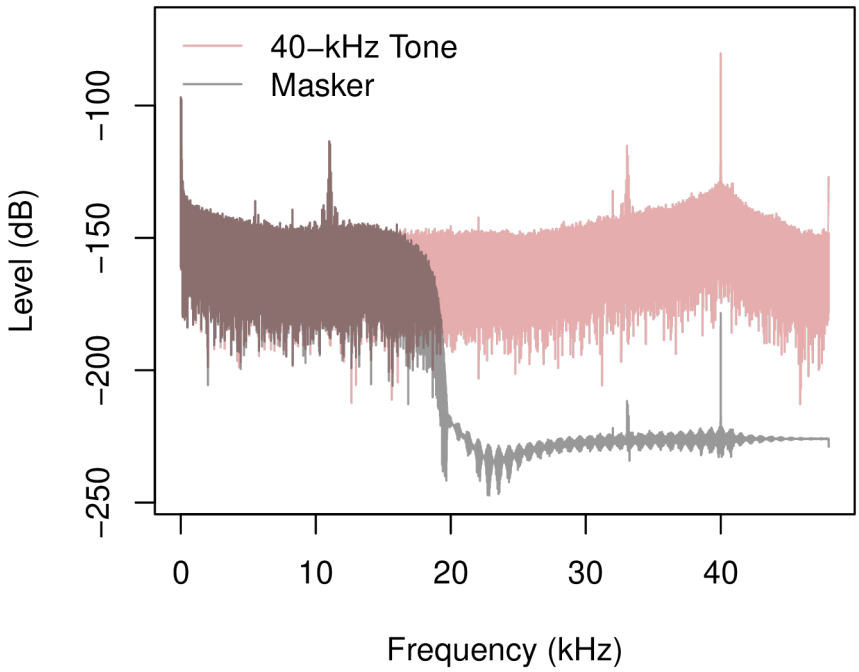

Figure S1: Spectrum of the 40-kHz US tone, and of the masker used in the behavioural US detection task. It should be noted that the two large peaks visible in the spectrum of the US recording around 11, and 33 kHz, as well as the smaller peaks around 5.5, 8.3, and 22 kHz were also present in recordings taken in the soundproof booth while the US speakers were not playing, so they are unrelated to the presentation of the US tone. Some of these peaks are also present in the masker, and may have been audible. However, given that the masker was presented during both the interval containing the US tone, and the interval without the US tone, their presence could not give a cue to the presence/absence of the US tone.

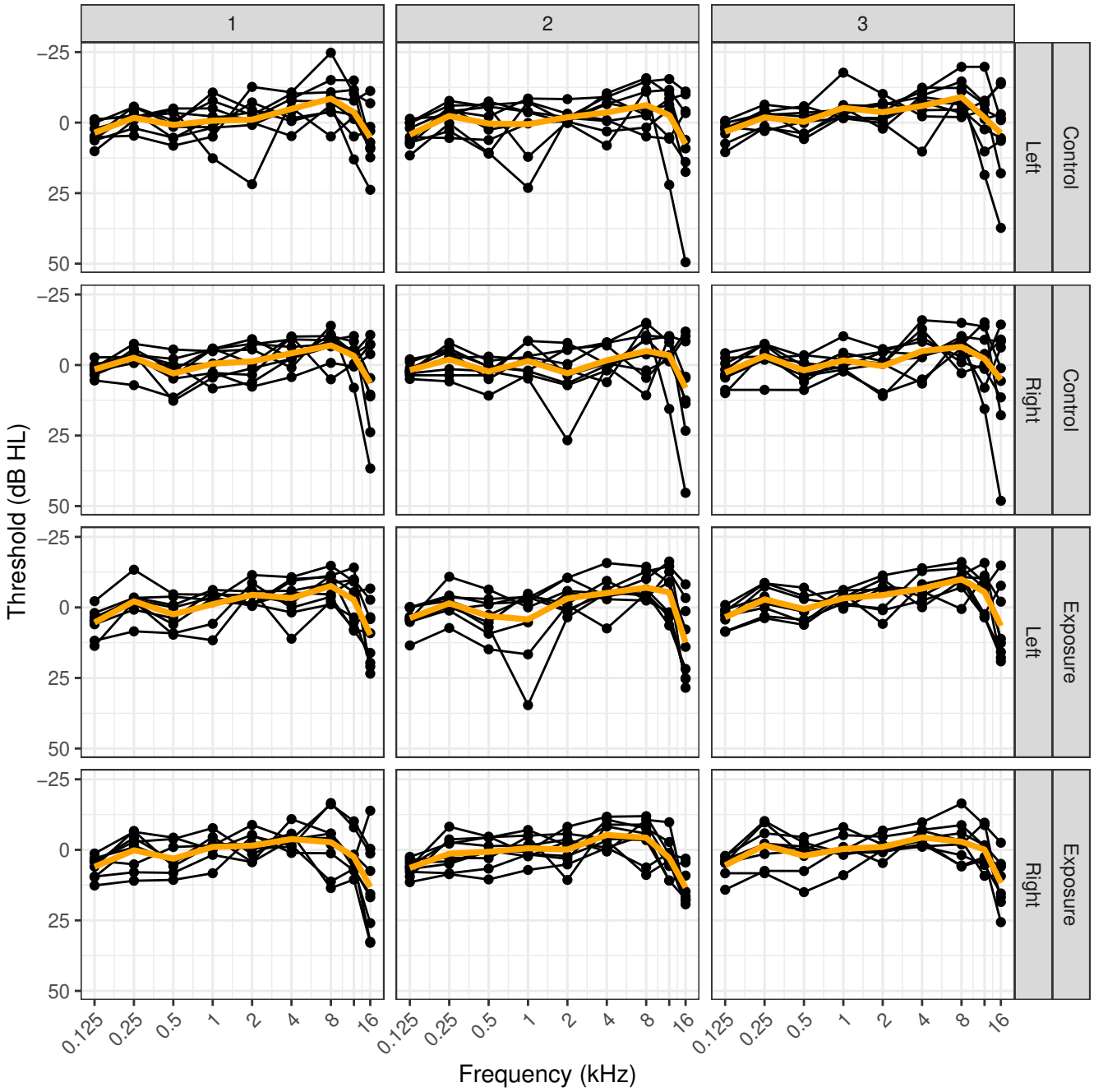

Figure S2: Audiograms for each participant as a function of session number (1, 2, or 3), ear (left, or right), and group (exposure, or control). The orange line shows the average for each panel.

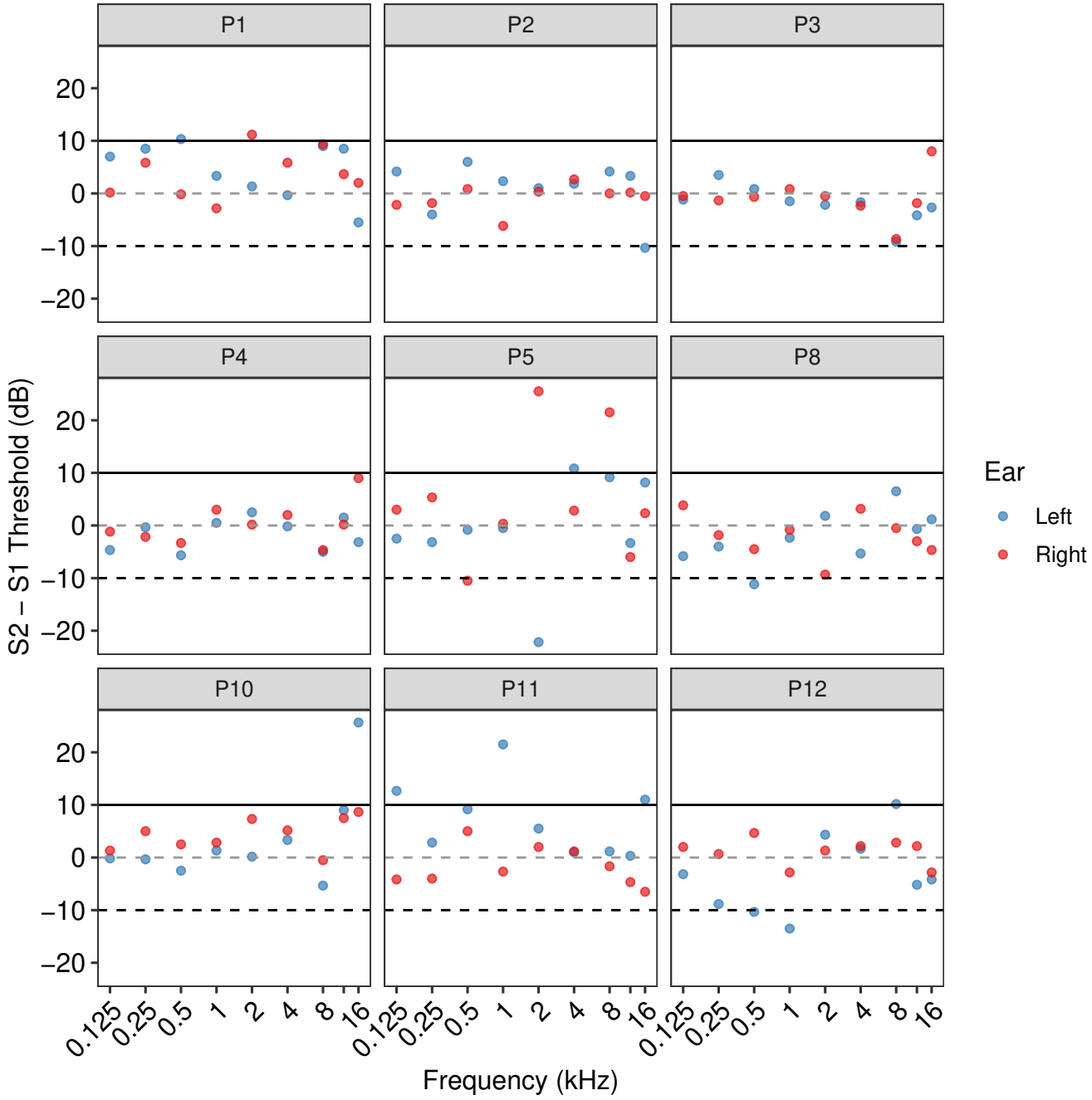

Figure S3: Differences in audiometric thresholds between S2 and S1 for each participant of the control group. Points above the solid line indicate estimated losses of sensitivity > 10 dB. Points below the dashed line indicate estimated gains of sensitivity > 10 dB.

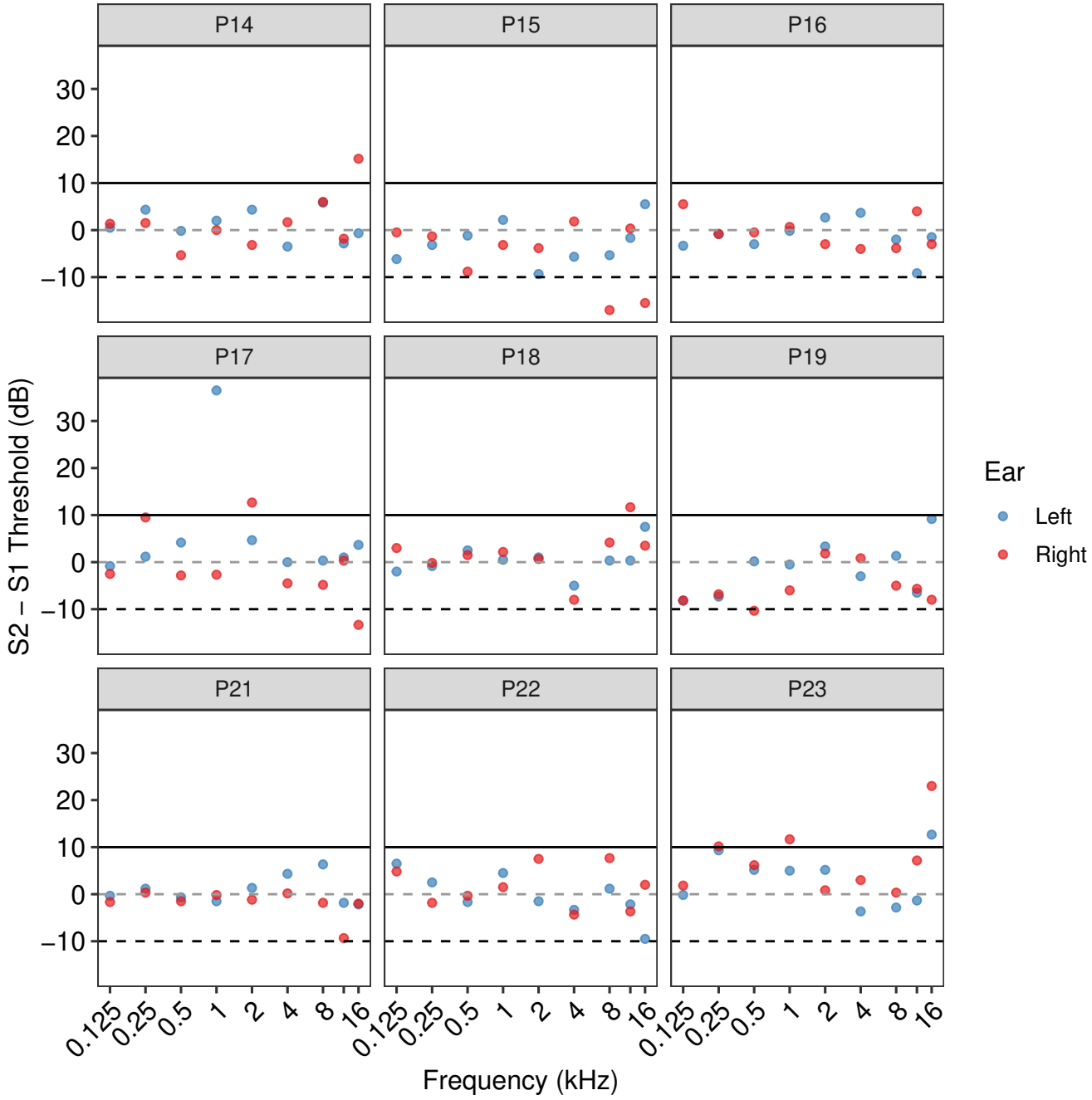

Figure S4: Differences in audiometric thresholds between S2 and S1 for each participant of the exposure group. Points above the solid line indicate estimated losses of sensitivity > 10 dB. Points below the dashed line indicate estimated gains of sensitivity > 10 dB.

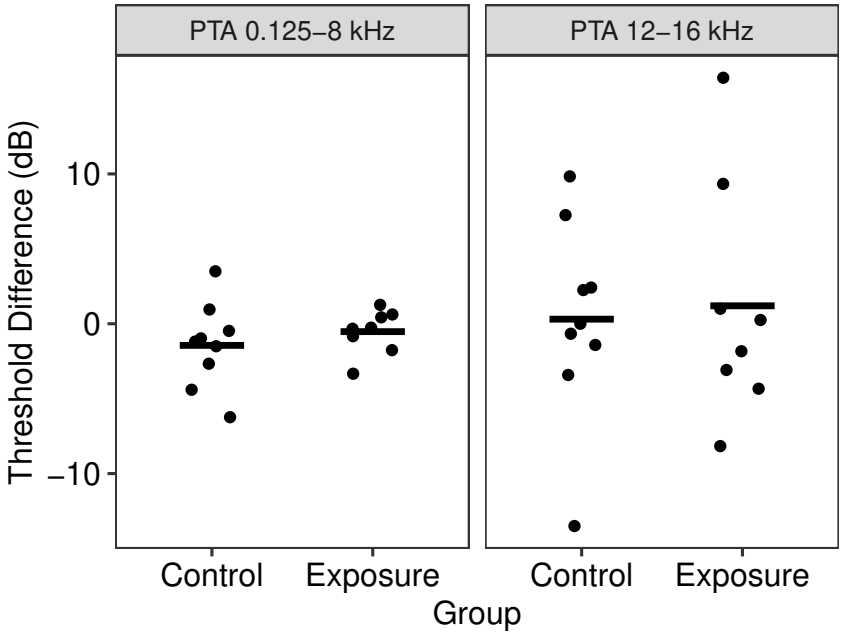

Figure S5: Difference in audiometric thresholds between the left and right ear, between session 3 and session 1: $(T_{L3} - T_{L1}) - (T_{R3} - T_{R1})$, where T refers to the threshold, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the threshold difference in the exposure group would indicate a relative post-exposure loss of sensitivity in the left (exposed) ear compared to the right ear. Points plot individual listeners' data. Segments plot group averages.

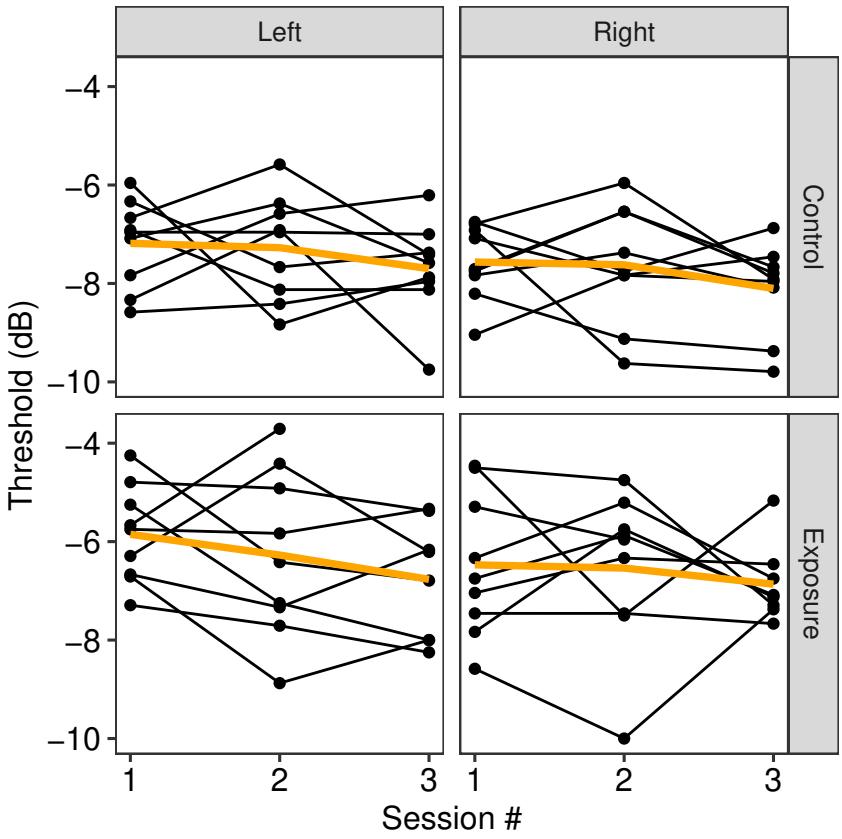

Figure S6: Threshold in the DTT task for each participant, as a function of session number. Results for each group and ear are shown in different panels. The orange line shows the average for each panel.

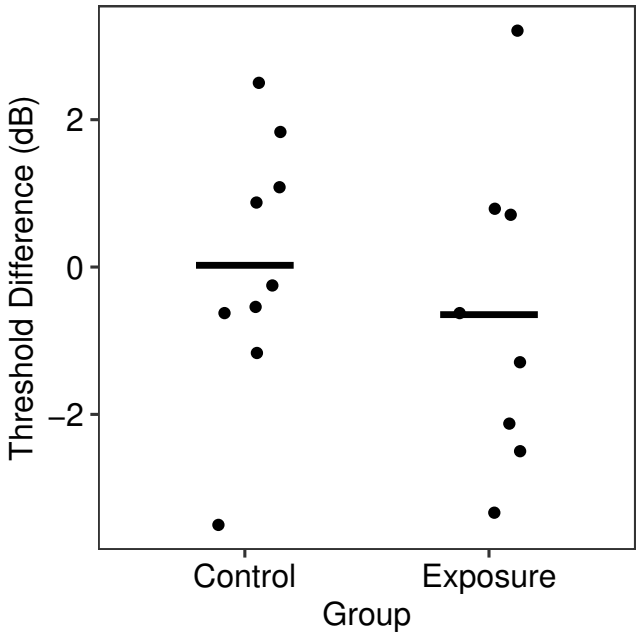

Figure S7: Difference in DTT thresholds between the left and right ear, between session 3 and session 1: $(T_{L3} - T_{L1}) - (T_{R3} - T_{R1})$, where T refers to the threshold, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the threshold difference in the exposure group would indicate a relative post-exposure performance drop for the left (exposed) ear compared to the right ear.

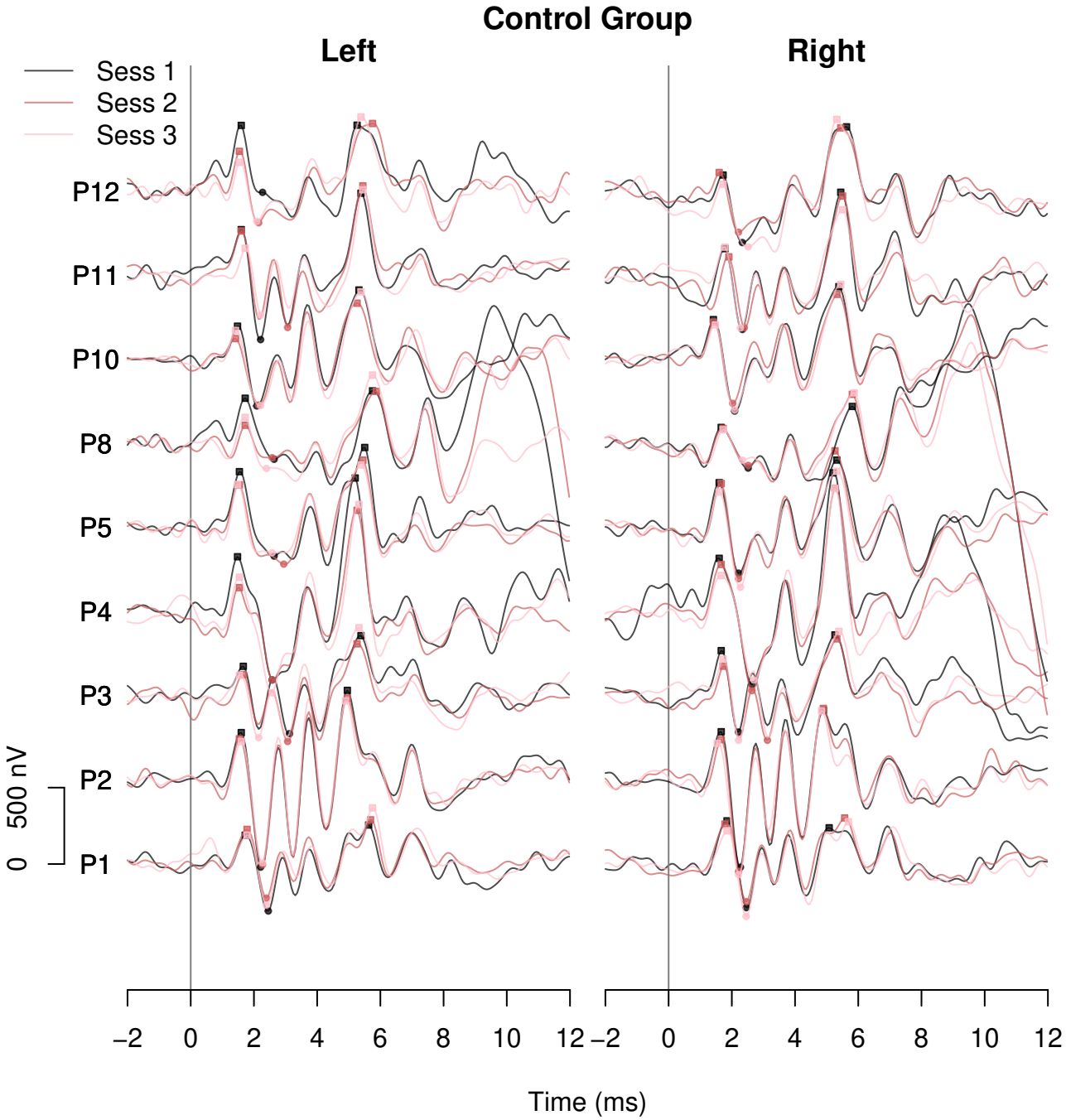

Figure S8: ABR waveforms for participants of the control group.

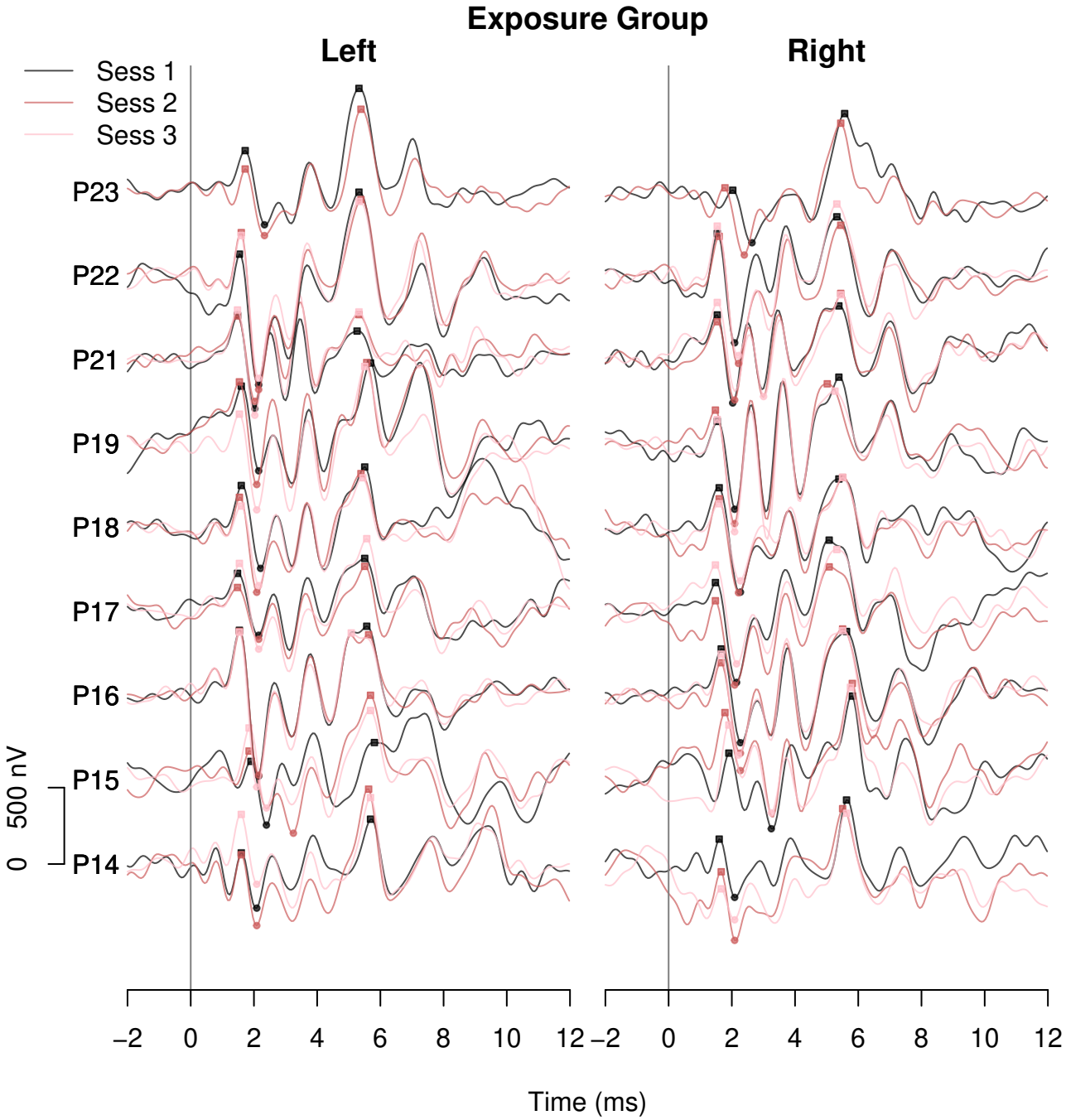

Figure S9: ABR waveforms for participants of the exposure group.

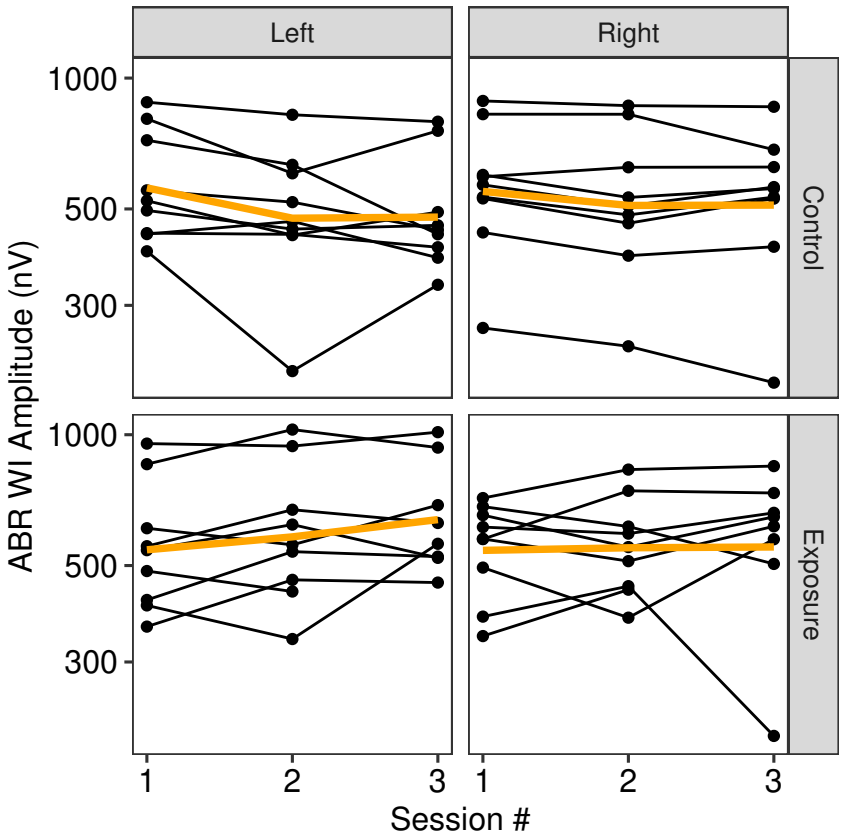

Figure S10: Wave I ABR amplitudes for each participant, as a function of session number. Results for each group and ear are shown in different panels. The orange line shows the geometric average for each panel.

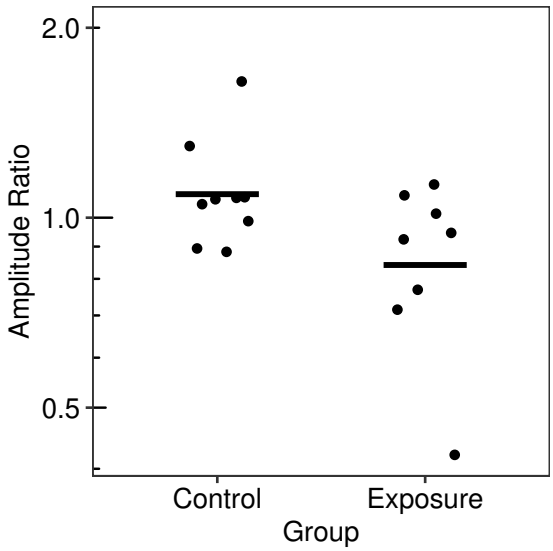

Figure S11: ABR wave I amplitude ratio between the right and left ear, between session 3 and session 1: $(A_{R3}/A_{R1})/(A_{L3}/A_{L1})$, where A refers to the amplitude, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the amplitude ratio in the exposure group would indicate a relative post-exposure wave I amplitude decrease in the left (exposed) ear compared to the right ear.

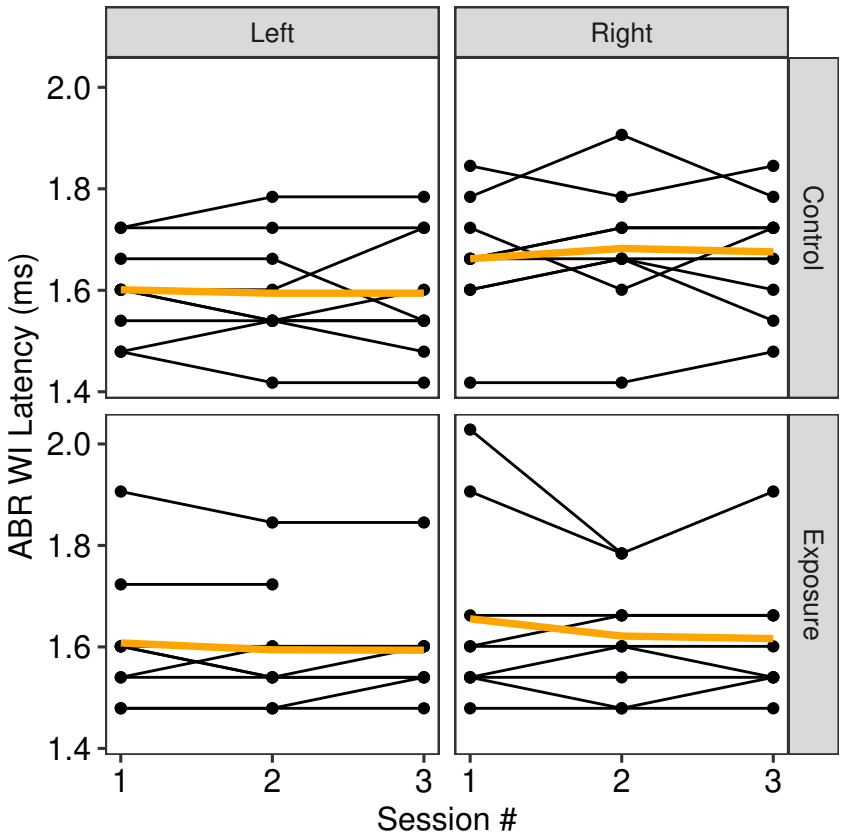

Figure S12: Wave I ABR latencies for each participant, as a function of session number. Results for each group and ear are shown in different panels. The orange line shows the average for each panel.

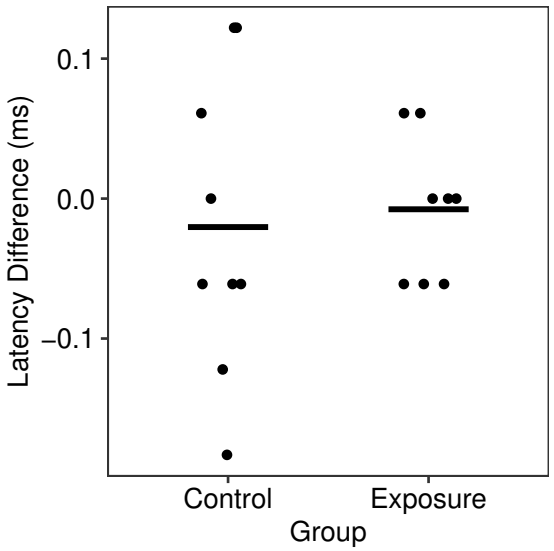

Figure S13: ABR wave I latency difference between the left and right ear, between session 3 and session 1: $(T_{L3} - T_{L1}) - (T_{R3} - T_{R1})$, where T refers to the latency, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the latency difference in the exposure group would indicate a relative post-exposure wave I latency increase in the left (exposed) ear compared to the right ear.

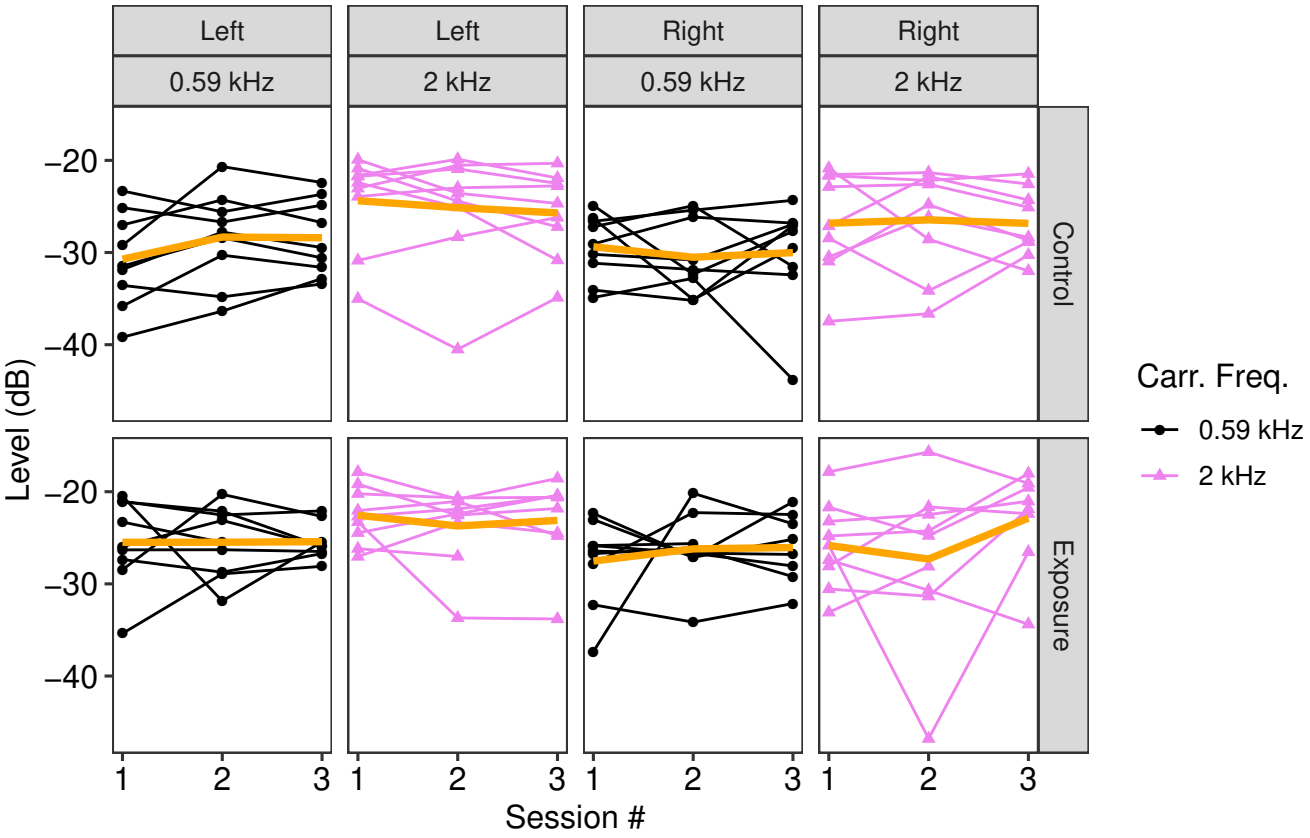

Figure S14: FFR levels for each participant at the modulation frequencies of the carriers, as a function of session number. Results for each group, ear, and carrier frequency are shown in different panels. The orange line shows the average for each panel.

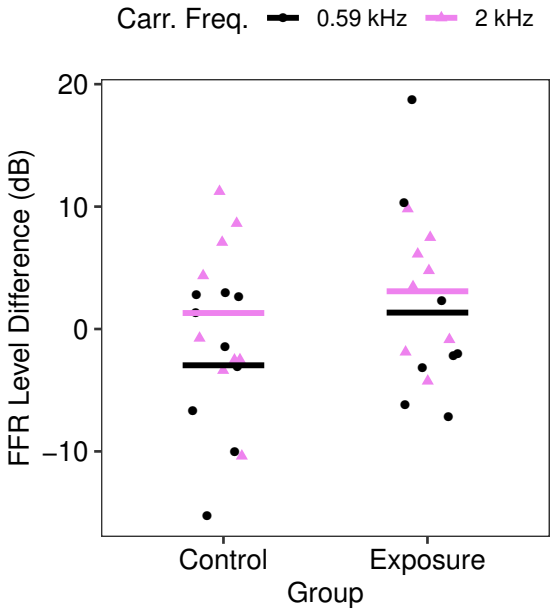

Figure S15: Difference in FFR level between the right and left ear, between session 3 and session 1 ($M_{R3} - M_{R1}) - (M_{L3} - M_{L1}$), where M refers to the level, the first subscript indicates the ear, and the second subscript the session number. An *increase* in the level difference in the exposure group would indicate a relative post-exposure decrease in FFR level for the left (exposed) ear compared to the right ear.

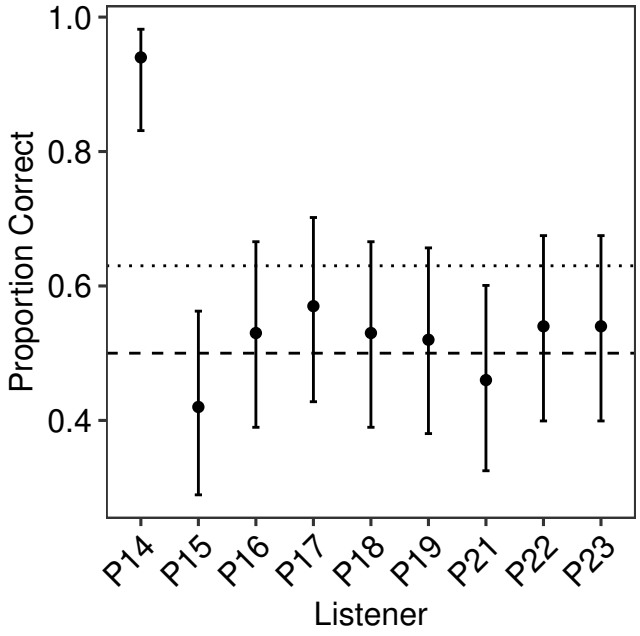

Figure S16: Proportion of correct responses in the detection of the 120 dB SPL ultrasound tone. The dashed line marks chance level. The dotted line marks the threshold for declaring significantly greater than chance level performance after correction for multiple comparisons. The error bars enclose 95% confidence intervals (corrected for multiple comparisons).

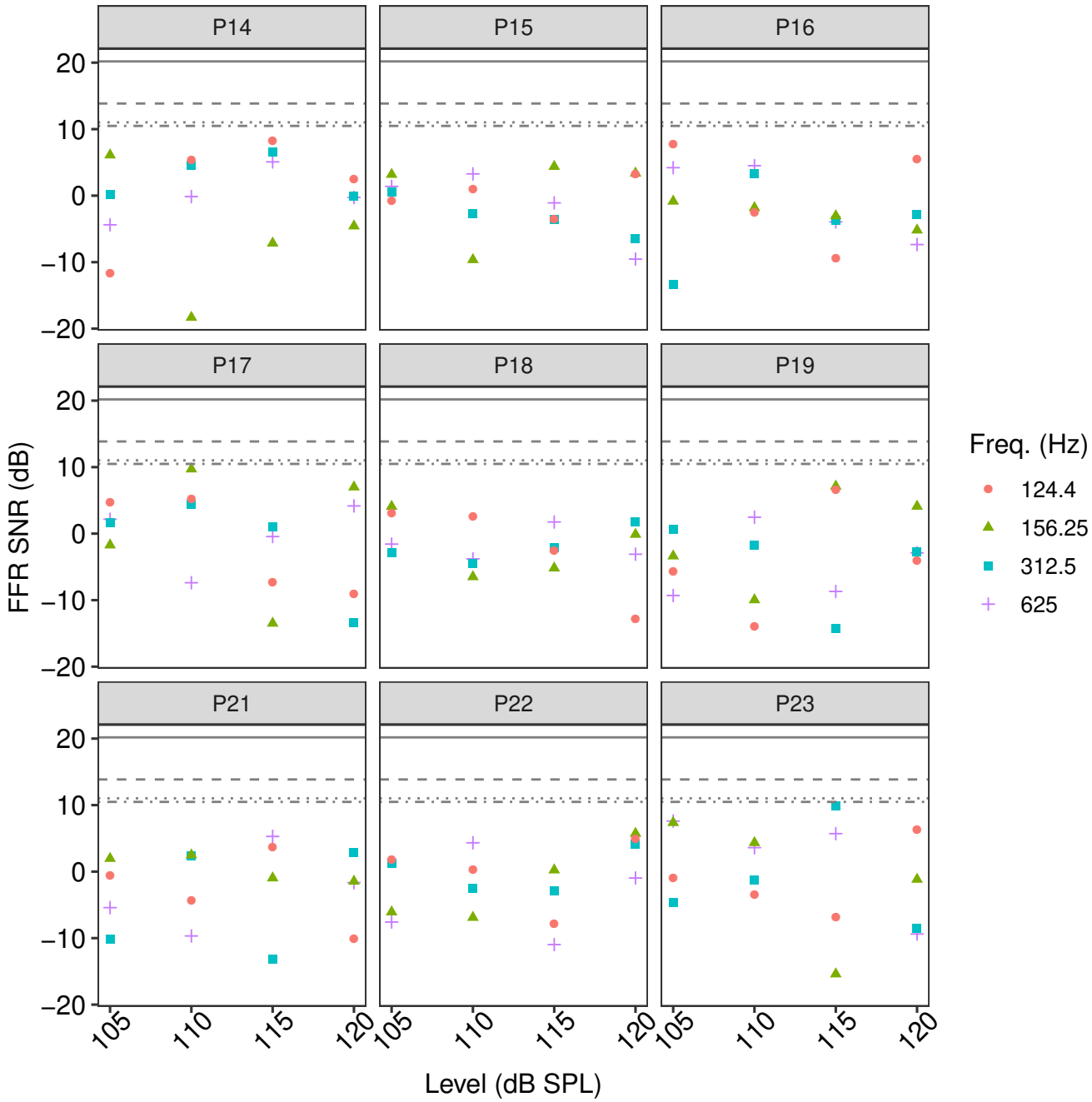

Figure S17: FFR SNR at subharmonics frequencies, and at the modulation frequency of the ultrasound tone for each participant of the exposure group. The lower dash-dotted line marks the SNR threshold for significant signal detection (after accounting for multiple comparisons) for the 124.4 Hz frequency (based on 5-seconds segments and 5 noise bins on each side). The dotted line marks the SNR threshold for significant signal detection for the 156.25 Hz frequency (based on 4-seconds segments and 4 noise bins on each side). The dashed line marks the SNR threshold for significant signal detection for the 312.5 Hz frequency (based on 2-seconds segments and 2 noise bin on each side). The upper solid line marks the SNR threshold for significant signal detection for the 625 Hz frequency (based on 1-seconds segments and 1 noise bin on each side).

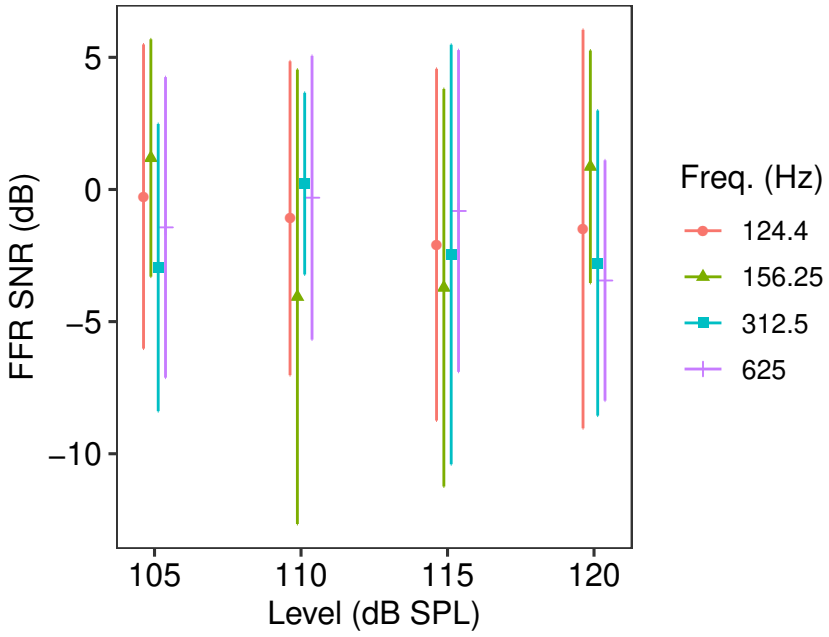

Figure S18: Across participant average FFR SNR at subharmonic and modulation frequencies of the ultrasound tone. The error bars mark ±1 s.d.

Supplementary tables referenced in the main manuscript

Variable	Corrected 95% CI (one-tailed)	Uncorrected 99% CI (two-tailed)	Bayesian HDI
$PTA_{0.125-8}$	-2.16–Inf	-2.37–4.2	-3.07–4.83
PTA_{12-16}	-9.17–Inf	-9.82–11.6	-12.47–14.33
DTT	-3.37–Inf	-3.55–2.21	-4.26–2.92
ABR WI Log-Amp.	-0.64–Inf	-0.66–0.15	-0.76–0.24
ABR WI Lat.	-0.1–Inf	-0.11–0.13	-0.13–0.16
$FFR_{0.59}$	-6.53–Inf	-7.25–15.85	-9.96–18.49
FFR_2	-6.31–Inf	-6.83–10.38	-8.43–12.64

Table S1: Interval estimates for the changes between S1 and S3 for the dependent measures analyzed in the study. The first column shows 95% one-sided CIs corrected for multiple comparisons. The second column show uncorrected 99% CIs. The third column shows 99% Bayesian HDIs.

Changes from the pre-registered protocol

- Although the pre-registered protocol specified testing 10 participants per group, due to time constraints it was not possible to continue data collection to achieve this goal.

- The protocol specified a delay of one to three days between S1 and S-US, and between S-US and S2, and a delay of one week between session S2 and S3. Because participants were occasionally unable to attend a scheduled session it was not possible to follow exactly the planned schedule for each participant.

- Exclusion criteria in the pre-registered protocol included audiometric thresholds > 20 dB HL at any octave frequency between 0.125 kHz and 8 kHz (inclusive) in either ear. One participant of the control group had an estimated threshold of 21.8 dB HL for the left ear at 2 kHz in her first session. However, the standard deviation of the turnpoints of the adaptive track used to estimate this threshold was high (5.14 dB). Because of this, and because the rest of the audiogram appeared normal we assumed that this high threshold was likely due to attentional lapses in the block of trials used to estimate it, and the participant was allowed to proceed onto the other sessions. Her threshold estimates for the left ear at 2 kHz in the remaining sessions were normal, confirming our suspicion that the high threshold estimated in the first session was indeed due to attentional lapses, and her data have been included in the analyses.

- The fact that the US speakers made an audible noise below the US frequency range was only discovered after submission of the pre-registration protocol, therefore the use of the masking noise in the behavioural US detection task is not mentioned there.

- The protocol for the behavioural US detection task specified that the US tone would be amplitude modulated at a frequency of 124.4 Hz. However, recordings of this amplitude modulated tone showed a component at the modulation frequency, possibly generated by modulation distortion in the air. This component was clearly audible. For this reason it was decided to use an unmodulated US tone instead.

- The pre-registration plan for the ultrasound EEG test specified performing an FFT on each 10-min block. However, to achieve a better signal-to-noise ratio shorter segments of the recording were averaged. The results obtained with this analysis were nonetheless qualitatively similar to the ones obtained with the pre-planned analysis, and did not change the study conclusions.

Supplementary methods

ABR wave I peak-peaking algorithm

The latency of the wave I peak was first identified in the grand-average waveform (obtained by averaging across participants from both groups) within a time window centred at a latency of 1.6 ms, and bounds set at ± 0.51 ms. These bounds correspond to ± 3 standard deviations of the ABR wave I latency reported by Issa and Ross [1]. The grad-average wave I peak was identified by selecting the highest local maximum in the search window. The wave I peaks were then searched in the individual subject waveforms within a search window centred at the grand-average wave I peak latency, and with bounds of ± 0.51 ms of the grand-average peak latency. Peaks were identified by selecting the highest local maximum in the search window, or the highest absolute point if no local maxima were present in the search window. Wave I amplitudes were measured from peak to trough. Troughs were identified by selecting the lowest local minimum in a search window going from 0.25 to 1.5 ms from the estimated peak latency, or the lowest absolute point if no local minima were present in the search window.

Supplementary results

Test-retest repeatability

Audiometry

Figure S19 shows the average absolute threshold difference between S1 and S2, and between S1 and S3, across participants from both groups, for each test frequency. Average absolute threshold differences were generally less than 5 dB, although they were higher at 16 kHz, where they reached 7 dB. The absolute threshold differences in this study were higher than those reported in a recent study by John *et al.* [2]. The higher absolute threshold differences found in this study may be partly due to the fact that test and retest were performed on different days, while in the John *et al.* study they were performed within the same day. Another difference between the two studies is that John *et al.* used a modified Hughson-Westlake clinical procedure to estimate thresholds, while in the current study a forced-choice procedure with a transformed up-down adaptive track was used.

Marshall *et al.* [3] measured detection thresholds for a pure tone in quiet using a forced-choice adaptive task, and a clinical procedure on nine listeners for ten blocks. They found that test-retest reliability, assessed by calculating the standard deviation across threshold estimates for each participant (intra-subject SD) was lower for the forced-choice adaptive task than for the clinical procedure. The average intra-subject SD for the forced-choice adaptive task in quiet was 2.2 dB. The average intra-subject SD for each condition of our study is shown in Figure S20. At the same test frequency used by Marshall *et al.* [3] the intra-subject SD was 2.3 dB for the right ear, but it was considerably higher (4.8 dB) for the left ear. It was also higher at most other frequencies in the clinical frequency range. The most likely reason for the higher intra-subject SDs observed in our study is that some listeners occasionally had high lapse rates. Listener's motivation is a factor known to affect psychophysical performance [4], but difficult to control, and we suspect that the occasionally high lapse rates may be due to this.

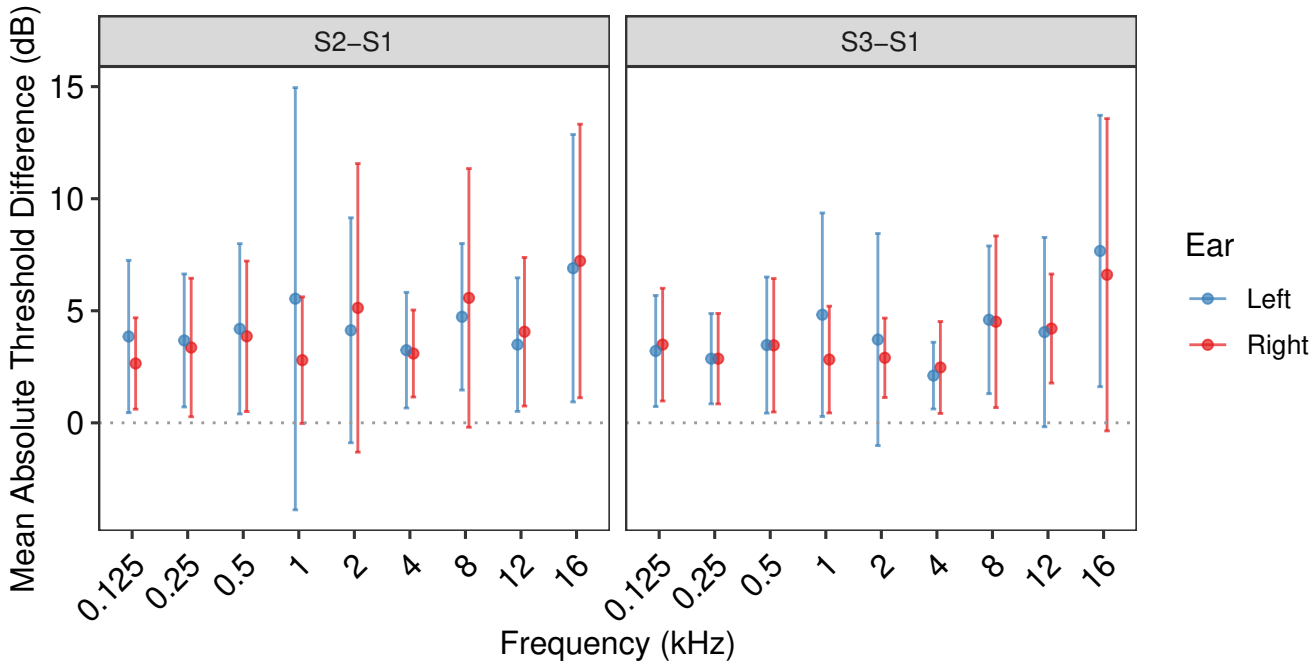

Figure S19: Average absolute test-retest differences in audiometric thresholds between session 1 and session 2 (left panel), and session 1 and session 3 (right panel). Averages were computed across participants from both the control and exposure groups. The error bars represent ±1 s.d.

We ran some Monte Carlo simulations of a virtual listener performing the forced-choice procedure with the adaptive track parameters (step size, number of turnpoints, etc...) used in the current study to investigate how reliability would be affected by varying the lapse rate. The virtual listener had a logistic psychometric function, with a 70.7% correct point of 0 dB HL, and a slope of 3.7 dB, which was typical of the slopes found by fitting psychometric functions to the data of this study. The results of the simulations showed that for a virtual listener with a 0% lapse rate the absolute threshold differences calculated on 1,000 random samples drawn with resampling from 1,000 Monte Carlo simulations was 1.1 dB, with an SD of 1.1. The absolute threshold difference

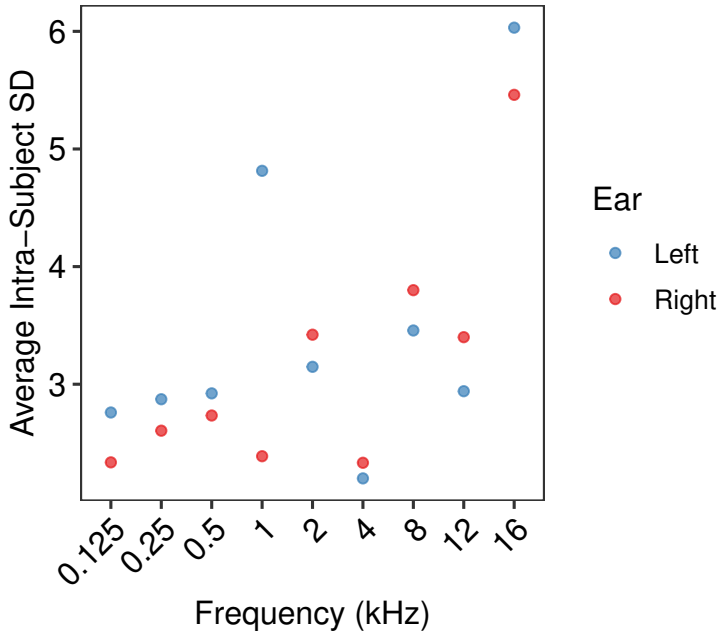

Figure S20: Average intra-subject standard deviations for audiometric threshold estimates measured in the three sessions. Averages were computed across participants from both the control and exposure groups.

was still below 1.5 dB for a virtual listener with 1 or 2% lapse rates, but increased to 4 dB for a virtual listener with a 5% lapse rate. Likewise the SD of the threshold estimates was only 1.3 dB for a virtual listener with a 0% lapse rate, increased to 1.49 dB for a virtual listener with a 2% lapse rate, and then reached 4.4 dB for a virtual listener with a 5% lapse rate.

It should be noted that while a few listeners showed apparent (positive or negative) large threshold shifts, $>$ 10 dB, at multiple test frequencies, the majority of the listeners either did not show large threshold shifts, or showed them only in 1 or 2 of the 18 different conditions (see Figures S3, and S4). Furthermore, when the data were averaged across the two frequency ranges of interest (clinical, and extended high frequency range), the absolute mean threshold differences were much lower. These can be seen in Table S2, which lists absolute threshold differences for all the dependent measures analysed in the study. Likewise the mean intra-subject SDs were much lower when the data were averaged across the two frequency ranges of interest. These can be seen in Table S3 which lists intra-subject SDs for all the dependent measures analysed in the study.

Another way to measure repeatability is the intraclass correlation coefficient (ICC). The ICC measures within-subject variability across sessions relative to between-subject variability, and it can be low if interindividual differences are small in the population or sample observed. ICCs calculated with the rptR package [5] for all the dependent measures analysed in the study are given in Table S4. The ICC for the average PTA in the extended high frequency range was higher than the ICC for the average PTA in the clinical frequency range, despite the fact that intrasubject SDs were lower in the clinical than in the extended high frequency range. The reason for this is that between-subject variability was higher in the extended than in the clinical frequency range.

Variable	S1-S2 Left	S1-S2 Right	S1-S2 Mean	S1-S3 Left	S1-S3 Right	S1-S3 Mean
$PTA_{0.125-8}$	2.63	2.3	2.46	2.15	1.25	1.7
PTA_{12-16}	3.87	4.87	4.37	4.46	4.5	4.48
DTT	1.19	1.13	1.16	1.07	0.88	0.98
ABR WI Log-Amp.	0.16	0.13	0.14	0.15	0.14	0.15
ABR WI Lat.	0.03	0.06	0.05	0.05	0.03	0.04
$FFR_{0.59}$	3.71	3.48	3.59	2.73	3.71	3.22
FFR_2	2.75	4.12	3.43	3.24	3.76	3.5

Table S2: Mean absolute differences between S1 and S2, and between S1 and S3 for the dependent measures analysed in the study. Values are given for the left ear, the right ear, and the mean of the left and right ear values.

Variable	Left	Right	Mean
$PTA_{0.125-8}$	2.03	1.56	1.79
PTA_{12-16}	3.5	3.69	3.6
DTT	0.83	0.82	0.82
ABR WI Log-Amp.	0.12	0.11	0.12
ABR WI Lat.	0.03	0.04	0.04
$FFR_{0.59}$	2.23	2.72	2.48
FFR_2	2.04	3.34	2.69

Table S3: Mean intra-subject SDs for the dependent measures analysed in the study. Values are given for the left ear, the right ear, and the mean of the left and right ear values.

Variable	ICC	95% CI
$PTA_{0.125-8}$	0.535	0.272–0.695
PTA_{12-16}	0.717	0.502–0.838
DTT	0.487	0.234–0.661
ABR WI Log-Amp.	0.742	0.526–0.849
ABR WI Lat.	0.813	0.638–0.896
$FFR_{0.59}$	0.536	0.291–0.701
FFR_2	0.554	0.308–0.719

Table S4: ICCs for the dependent measures analysed in the study. The second column shows the 95% ICC confidence intervals.

DTT

Both mean absolute across-session differences (Table S2), and mean intra-subject SDs (Table S3) were relatively small, indicating good reliability of the measure. The ICC (Table S4), however, was modest due to the fact that between-subject variability was low.

ABR

Both wave I ABR amplitudes and latencies were remarkably stable across sessions, as indexed by the mean absolute differences (Table S2), mean intra-subject SDs (Table S3), and ICCs (Table S4). The log-amplitude mean absolute differences are easier to interpret when converted to ratios by exponentiating. When converted to ratios they ranged from 1.14 to 1.17. Likewise average intra-subject SDs are easier to interpret when converted to ratios (or equivalently when calculated as geometric averages of the geometric intra-subject SDs). These ranged from 1.12 to 1.13. The ICCs were similar in size to those obtained in two recent study of supra-threshold ABR test-retest reliability [6, 7].

FFR

Reliability of the FFR measures was only moderate, as indexed by the mean asolute differences (Table S2), mean intra-subject SDs (Table S3), and ICCs (Table S4). A recent study by Guest *et al.* [7] reported high reliability for FFR level in response to AM tones. However, there were several differences in the stimuli and procedures used in this previous study, and the current study, which may explain the lower test-retest reliability observed in the current study. Unlike the previous study we presented the stimuli monaurally, we presented two stimuli simultaneously, and we used AM tones rather than transposed tones. Monaural stimulation leads to lower FFR amplitudes even when the monaural stimuli are presented at higher SPLs to compensate for level differences [8]. At stimulus levels of 75 dB, FFR amplitudes to multiple simultaneous stimuli have been found to be reduced in amplitude compared to when the stimuli are presented individually [9]; these amplitude reductions were largest for the stimulus with the lower carrier frequency. Transposed tones enhance phase locking to the envelope of modulated high-frequency carriers compared to AM tones [10, 11]. Overall these three factors are likely to explain at least in part the lower FFR amplitudes observed in the current study, which resulted in FFR levels being closer to the noise baseline, and likely reduced test-retest reliability. Other differences between the studies may also have played a role in the reduced test-retest reliability observed in the current study. For example, due to space limitations, in the current study it was not possible to recline the chair during the recordings, which may have led to increased myogenic artifacts.

Bayesian model

The JAGS code for the Bayesian model is provided below:

```
model {
   # likelihood
   for (i in 1:Ntotal) {
      y[i] ~ dnorm(mu[x[i]], 1/sigma[x[i]]^2)
   }

   #priors
   for (j in 1:2) {
     mu[j] ~ dnorm(meanY , 1/(100*sdY)^2)
     sigma[j] ~ dunif(sdY/1000 , sdY*1000)
   }
   muDiff = mu[1]-mu[2]
```

`y` is a vector with the dependent variable. `x` is a vector indicating the group (experimental or control). `meanY`, and `sdY` are respectively the mean, and the standard deviation of the dependent variable across groups; these values are used to set vague priors on the scale of the data.

References

[1] A. Issa and H. F. Ross. An improved procedure for assessing ABR latency in young subjects based on a new normative data set. *Int. J. Pediatr. Otorhinolaryngol.*, 32(1):35–47, 1995.

[2] A. B. John, Vinay, and B. M. Kreisman. Equivalence and test-retest reproducibility of conventional and extended-high-frequency audiometric thresholds obtained using pure-tone and narrow-band-noise stimuli. *Int. J. Audiol.*, 56(9):635–642, 2017.

[3] L. Marshall, T. E. Hanna, and R. H. Wilson. Effect of step size on clinical and adaptive 2IFC procedures in quiet and in a noise background. *J. Speech Hear. Res.*, 39(4):687–696, 1996.

[4] M. Florentine, S. Buus, and C. R. Mason. Level discrimination as a function of level for tones from 0.25 to 16 kHz. *J. Acoust. Soc. Am.*, 81(5):1528–1541, 1987.

[5] M. A. Stoffel, S. Nakagawa, and H. Schielzeth. rptr: repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 8(11):1639–1644, 2017.

[6] G. Prendergast, W. Tu, H. Guest, R. E. Millman, K. Kluk, S. Couth, K. J. Munro, and C. J. Plack. Supra-threshold auditory brainstem response amplitudes in humans: Test-retest reliability, electrode montage and noise exposure. *Hear. Res.*, 364:38–47, 2018.

[7] H. Guest, K. J. Munro, G. Prendergast, and C. J. Plack. Reliability and interrelations of seven proxy measures of cochlear synaptopathy. *Hear. Res.*, 375:34–43, 2019.

[8] J. L. Clark, G. Moushegian, and A. L. Rupert. Interaural time effects on the frequency-following response. *J. Am. Acad. Audiol.*, 8(5):308–313, 1997.

[9] M. S. John, O. G. Lins, B. L. Boucher, and T. W. Picton. Multiple auditory steady-state responses (MASTER): stimulus and recording parameters. *Audiology*, 37(2):59–82, 1998.

[10] L. R. Bernstein and C. Trahiotis. Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli". *J. Acoust. Soc. Am.*, 112(3 Pt 1):1026–1036, 2002.

[11] A. Dreyer and B. Delgutte. Phase locking of auditory-nerve fibers to the envelopes of high-frequency sounds: implications for sound localization. *J. Neurophysiol.*, 96(5):2327–2341, 2006.