Salience-dependent disruption of sustained auditory attention can be inferred from evoked pupil responses and neural tracking of task-irrelevant sounds

Lorenz Fiedler[1], Ingrid Johnsrude[1,2], Dorothea Wendt[1,3]

[1]Eriksholm Research Centre, Part of Oticon A/S, Snekkersten, Denmark

[2]Brain and Mind Institute, University of Western Ontario, London, Ontario N6A 5B7, Canada

[3]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

Corresponding Author: Lorenz Fiedler

Number of pages: 42

Number of figures: 7

Number of tables: 0

Number of words for abstract: 250

Number of words for significance statement: 115

Number of words for introduction: 682

Number of words for discussion: 1,425

Short title: Salience-dependent disruption of sustained auditory attention

Abstract

Stimulus-driven attention allows us to react to relevant stimuli (and imminent danger!) outside our current focus of attention. But irrelevant stimuli can disrupt attention as well, for example

during listening to speech. The degree to which sound captures attention is called salience, which can be estimated by existing, behaviorally validated, computational models (Huang & Elhilali, 2017). Here we examined whether neurophysiological responses to task-irrelevant sounds indicate the degree of distraction during a sustained listening task and how much this depends on individual hearing thresholds. To this end, we asked N = 47 Danish speaking adults (28/19 female/male, 24 hearing aid users) to listen to continuous speech while one-second-long, task-irrelevant natural sounds (distractors) of varying computed salience were presented at unpredictable times and locations. The task-irrelevant sounds evoked a consistent pupil dilation (PD), neural distractor-tracking (DT) and a drop of neural target-speech-tracking (ΔTT). Within-subject statistical modelling showed that PD and DT are enhanced for sounds with higher computed salience (salience sensitivity), independent of hearing thresholds. Participants with generally larger PD showed stronger ΔTT and performed worse in target speech comprehension. Participants who exhibited a stronger ΔTT for more salient sounds (salience sensitivity) understood less of the target speech. Beyond hearing loss, PD, target speech tracking (and its salience sensitivity) explained 43% of variance in target speech comprehension. We conclude that distraction can be inferred from neurophysiological responses to task-irrelevant stimuli. Our findings represent a step towards neurophysiological assessment of attention dynamics during continuous listening with potential applications in hearing care diagnostics.

Significance statement

Successful speech-in-noise comprehension in daily life does not only depend on the acuity of the auditory input, but also cognitive factors like attentional control. Being able to measure distraction-dependent neurophysiological responses to peripheral, task-irrelevant stimuli would enable monitoring how much the attentional focus is instantaneously captured away from a target under sustained attention. Here we show that both pupil dilation and neural tracking of distractor sounds, reflect the degree to which people with both normal and elevated hearing thresholds are distracted. Such a measure could be used to non-invasively track the focus of attention and thus could find application in hearing care diagnostics, where cognitive factors like attentional control are being increasingly recognized as important.

Introduction

Distractibility is a vital feature of our attentional system. For example, we still need to hear a car approaching while we focus our attention on the traffic light across the street. While voluntary attention (a.k.a. top-down attention, intentional attention) narrows down incoming information to a minimum, peripheral stimuli can still capture our attention in a stimulus-driven fashion (also called automatic or bottom-up attention). Unfortunately, some sounds may capture our attention for no relevant reason. For example, someone laughing at the table next to us in a restaurant may capture attention even though we are trying to focus on our current conversational partner. Hearing comes with

the vital blessing of helping us to detect potential danger from any direction, but with the curse of leading to distraction by irrelevant sounds.

The degree to which a sound captures stimulus-driven attention is called salience (Kaya & Elhilali, 2014). Voluntary selective auditory attention is established through the neural processing of the incoming sensory signals mainly at the auditory cortical level. How, and where, stimulus-driven attention is implemented along the auditory pathway and which stimulus features are driving it is as yet unclear (Huang & Elhilali, 2020; Serences & Kastner, 2014). For example, a sound can capture attention due to its sheer loudness (or other psychoacoustic features) independent of its meaning, whereas our own name may capture our attention even if it is spoken at a moderate level (Cherry, 1953). Here we first focus on objective, acoustic features that give rise to salience and second, the beyond-acoustic aspects of salience that may be captured in the statistical dependencies in the neurophysiological measures and their ability to predict behavior within-subjects.

There have been several attempts to develop and behaviorally validate models for auditory salience. Inspired by vision research, earlier attempts (Duangudom & Anderson, 2007; Kalinli & Narayanan, 2007; Kaya & Elhilali, 2012; Kayser et al., 2005) treated the spectrogram like a picture and thus have been criticized for ignoring the unfolding of a sound in time (Kaya & Elhilali, 2017). Later, the temporal unfolding of established psychoacoustic features such as loudness, harmonicity and

pitch were incorporated into the models (Huang & Elhilali, 2017; Kaya & Elhilali, 2014). Such models are meant to explain an automatic, stimulus-driven behavior, but have been validated behaviorally using explicit tasks, such as rating and detection. The question is whether these models hold in continuous listening tasks with participants naïve about the purpose of the study. Only recently, these models were found to explain some variability in neural phase-locking during continuous listening paradigms (Huang & Elhilali, 2020; Straetmans et al., 2021). While these studies investigated event-related potentials to discrete events, we here aim to generalize their findings to cortical tracking of continuous stimuli and pupillometry.

Qualitatively, listeners with hearing impairment (HI) complain about a lack of situational awareness on the one hand, but also about distraction by irrelevant sounds on the other (Nielsen & Henriksen, 2022). That means they are lacking selectivity in the processing of salient stimuli, which may have two causes: 1) hearing loss leads to a degraded input to the auditory pathway which corrupts features essential for stream segregation and selection (David et al., 2018; Oxenham, 2008); 2) Increased cognitive load and fatigue may limit the resources available for attentional control (Peelle, 2018). To better understand why selectivity in perception of salient stimuli is reduced in people with HI, we need methods that can assess the degree of distraction in scenarios of high ecological validity such as continuous speech.

We asked if established neurophysiological measures for attention (cortical tracking) and listening effort (pupil size) correlate with the output of a recent model for auditory salience (Huang & Elhilali, 2017, 2020). We examined the neurophysiological response to task-irrelevant sounds while both hearing aid users and non-hearing aid users listened to continuous speech. We hypothesize that: (1) there is a consistent neurophysiological response (i.e., neural tracking and pupil dilation) to task-irrelevant sounds; (2) these responses correlate positively with computed salience; and (3) participants with higher hearing thresholds (i.e., HI) exhibit stronger neurophysiological responses and salience sensitivity, even after compensation for elevated hearing thresholds.

Methods

Participants

A total of N = 47 participants (28 female) were enrolled in this study as members of two groups: 1) 24 hearing-aid users (HI) with sensorineural hearing loss, one aged 23 years and the rest aged between 59 and 82 years; 2) 23 non-hearing-aid users (NH) aged between 27 and 64 years of age. Figure 2A depicts hearing thresholds of the better ear and Figure 2B depicts the correlation between pure-tone average thresholds (PTA; 500-4000 Hz) and age. Due to the collinearity of age and PTA in our sample, we only dissociate between the groups for illustrative purposes. In the statistical modelling, we controlled for age effects before PTA was used as a predictor (see statistical modelling). The study was approved by the Copenhagen

regional ethics committee: (De Videnskabsetiske Komitéer, Region Hovedstaden).

Apparatus

The experiment took place in an acoustically and electro-magnetically shielded booth. Participants were seated in a comfortable office chair. The office chair was placed such that participants were seated at a reference point in the middle of the room and their ears were level with the loudspeakers. A computer keyboard was placed on a desk in front of the participants. Stickers were used to label the left and right Ctrl-button with “Ja” & “Nej” (Yes & No), respectively. The assignment of the labels to left and right was balanced across participants. The space bar was labelled with “Start”. Behind the keyboard, a Tobii Pro Spectrum eye tracker (Tobii AB, Stockholm, Sweden) was placed. The light conditions were kept at a moderate level and measured at 75 lux at the reference point.

Three loudspeakers (Genelec 8030A, Genelec Oy, Iisalmi, Finland) were placed at a radius of one meter from the reference point. One loudspeaker was positioned at 0 degrees azimuth (front, behind the desk, above the screen), and two were positioned to the left and right at ±90 degrees azimuth. The loudspeakers were connected to an RME Multiface II audio interface (AudioAG, Haimhausen, Germany). The audio interface was connected via USB to a personal computer running Microsoft Windows 10 (Microsoft Corp., Redmond, USA), where MATLAB 2020b (Mathworks Inc., Natick, USA)

and Tobii Eye Tracker Manager were installed. The audio was controlled from MATLAB using Soundmexpro toolbox (Hörtech gGmbH, Oldenburg, Germany). Underneath the frontal loudspeaker, a screen was placed to display task instructions and comprehension questions. The screen was controlled via MATLAB using Psychtoolbox version 3.0.17 (Brainard, 1997).

A BioSemi ActiveTwo EEG amplifier (BioSemi B.V., Amsterdam, The Netherlands) with 64 channels was placed behind the participant. A BioSemi USB Trigger Interface cable connected the computer with the BioSemi receiver. This connection was used to send timing markers, while precise timing was ensured by sending triggers to the EEG amplifier (via BrainProducts StimTrak, Brain Products GmbH, Gilching, Germany) from the audio interface. For EEG data collection, the BioSemi receiver was connected via USB to another, identical computer, which had BioSemi ActiView installed.

Stimulus material

Target stimuli: A collection of 320 continuous speech recordings, each with a length of 30 seconds, were used as target stimuli. The recordings were local radio newsclips (used in earlier studies Alickovic et al., 2020, 2021; Fiedler et al., 2021) spoken by four different Danish radio announcers (2 female, 2 male). The root-mean-square (RMS) of the recordings was equalized, such that each of the 30-second-long recordings had the same RMS.

Distractor sounds: The goal was to collect sounds that varied in spectro-temporal profiles. A collection of 690, one-second-long

natural sounds was compiled from freely available sources such as freesound.org, mixkit.co, as well as existing databases from former studies (Ananthabhotla et al., 2019; Gygi & Shafiro, 2019; Santoro et al., 2014; Zhao et al., 2019). The collection consisted of sounds from animals, tools, machines, music, human sounds like coughing and laughing, but no speech. The RMS of the distractor sounds was equalized, such that each of the one-second-long sounds had the same RMS. Subsequently, the overall level of the sounds was lowered by 10 dB, such that the RMS of the distractor sounds was 10 dB below the RMS of the target stimuli.

The distractor sounds were embedded in stationary noise (Figure 1B). White noise was created and filtered such that its long-term average spectrum (LTAS) fitted the LTAS of all distractor sounds following procedures of May et al., (2018). In brief, the LTAS of all concatenated distractor stimuli was calculated in bandwidths of 1/3 octaves. A finite-impulse response filter was designed (using the function fir2 in MATLAB) and applied to the white noise such that its LTAS was adjusted to match the LTAS of the distractor sounds.

Randomization: We randomized the selection of target stimuli such that, for each participant, 92 of the 320 30-sec target stimuli were selected and one was appended to the first to make a one-minute stimulus (Figure 1B). Each participant heard 46 such one-minute-long stimuli, one on each trial. The concatenated stimuli always consisted of two different target talkers and after 30 seconds there was a switch of target talker.

We randomized the selection of distractor stimuli by randomly selecting 280 out of the 690 different sounds for each participant. An inter-distractor interval for distractor onsets was randomly assigned to each of the distractor stimuli by drawing from a discrete uniform distribution between 5 and 10 seconds in steps of 0.2 seconds. Furthermore, the distractor location “left” or “right” was randomly assigned with the exception that the same location was never repeated more than 5 times in a row.

Based on the randomization of target and distractor stimuli, individual three-channel audio files were created for each trial (Figure 1B). Each sound file started with three seconds of silence, after which the noise started at the left and right loudspeakers (±90°). After 3s of noise, the target stimulus started at the loudspeaker in front (0°).

The distractor sounds were presented at the left and right loudspeakers embedded in the noise (Figure 1B). Within the one-minute-long trials, distractor stimuli could occur between 5 and 55 seconds relative to target onset (the ‘distractor period’). This was done to ensure that there was enough time for the participant to establish an attentional focus on the target talker at the beginning of the trial and that there was enough time to collect the physiological responses to the last distractor sound of the trial.

To uniformly distribute the distractor sounds across the distractor period, the inter-distractor interval was carried over from one trial to the next. For example, if the last distractor in

trial n-1 was presented at 51 seconds, and the next scheduled distractor had an associated inter-distractor interval of 8 seconds, 4 seconds of that were used at the end of trial n-1 (i.e., between 51 and 55 sec), and the remaining 4 seconds were appended at the beginning of trial n (i.e., next distractor started 5 sec plus 4 sec = 9 sec into trial n). At the beginning of each distractor sound, the noise was faded out, and was faded back in again at the end of the distractor using 20-msec half-sinusoidal ramps. Since the noise had the same LTAS as the distractors, the fade-out/fade-in ensured that, during the distractor sounds, the energetic masking was constant in the long-term average sense.

Study design

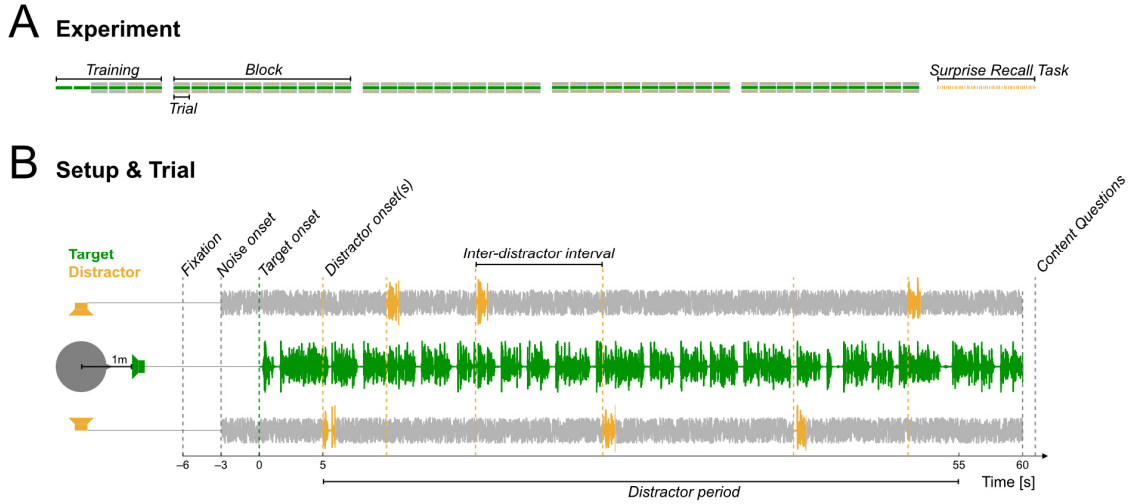


Figure 1: Study design. A) The main experiment consisted of 46 trials (the first 6 trials were training trials) followed by a surprise recall task. **B)** Three loudspeakers were set up at the respective angles of 0 and ±90 degrees. Each trial started with three seconds of silence, followed by the onset of stationary noise from the loudspeakers at ±90 degrees. Three seconds later, target speech was presented from the loudspeaker at 0 degrees. Between 5 seconds post target onset (earliest) and 55 seconds post target onset (latest), 1-second-long sounds were presented randomly at + or – 90 degrees and with a jittered onset interval between 5 and 10 seconds.

The experiment was blocked, and participants could take breaks between blocks. The first six trials were for training (Figure 1A). In the first two trials, only the target stimuli were presented. The stationary noise was introduced in trials 3 & 4 and the distractors were introduced in trials 5 & 6. Forty trials were presented in the main experiment. Together with the two last trials from the training, 42 trials were presented with distractor sounds; with largely different targets and distractors for each participant, as described above.

Hearing loss compensation and adaptation

Hearing-aid users were not wearing their hearing aids during the experiment. To make sure that all participants were presented with stimuli of comparable loudness, we individualized the amplification in two steps (as described in detail below): first, we used linear gain to spectrally equalize (i.e., flatten) local dips in their pure-tone audiograms and second, we estimated a stimulus-specific hearing threshold. The resulting amplification was applied to the stimuli before presentation.

The linear gain was applied by designing an individual finite-impulse response (FIR) filter for each participant, with the goal to compensate local dips in the audiogram. We aimed to provide gain to frequencies that exceeded 20 dB difference to the best hearing threshold. The presentation in free field only allowed for one, better-ear gain profile to make sure that we would not present at uncomfortable, excessively loud, levels. First, we measured the individual pure-tone audiograms to identify, for each person, for each frequency (125, 250, 750, 1000, 1500,

311 2000, 3000, 4000, 6000, 8000 Hz) the lower (i.e., better) hearing
312 threshold (HT) between the two ears (Figure 2A). Second, we
313 identified the lowest (i.e., best) HT across these measurements,
314 which can be seen as a reference peak in the audiogram. Third,
315 we calculated the to-be-added gain for frequencies with an HT
316 more than 20 dB higher (i.e., worse) than the reference. The
317 gain was calculated from the difference between the specific HT
318 and the best HT minus 20 dB. Hence, we filled up dips in the
319 audiogram that exceeded 20 dB relative to the reference peak.
320 To derive the gain for a 1/3-octave spaced FIR-filter, we took the
321 gain calculated at the frequencies of the audiogram as support
322 points and interpolated the gain values missing on a 1/3-octave-
323 spaced frequency axis. Finally, an FIR-filter was designed using
324 the function fir2 in MATLAB and applied to all stimuli.

325 Next, a stimulus-specific hearing threshold was estimated to
326 derive an individual broad-band amplification to be applied to
327 the equalized stimuli. We estimated this by presenting the
328 participants with a total of eight 30-second examples of the
329 target stimuli (two from each talker), to which the gain was
330 applied as derived in step one. The presentation started below
331 hearing level and the level was increased at a rate of 2 dB per
332 second. We asked the participants to press the space bar as
333 soon as they could hear the slightest whisper. The average
334 amplification at which they pressed the button was calculated
335 and applied as an additional broad-band amplification, which will
336 be referred to as individual hearing level (IHL). The overall
337 resulting gain is depicted in Figure 2C. The comparison of the

PTA and the applied gain suggests that the resulting amplification approximately fits the one-half-gain rule (Fig. 2D; Lybarger, 1944). As a sanity check, the adaptive procedure was repeated but this time we asked participants to press the space bar as soon as they could understand 95% of the words, which will be referred to as comprehension level (CL). This was to make sure that the presentation level was above the individual CL, or in other words, that participants could understand the target stimuli. We found that the individual difference between CL and IHL varied between 7 and 29 dB. For all participants, sounds were presented at 40 dB above IHL.

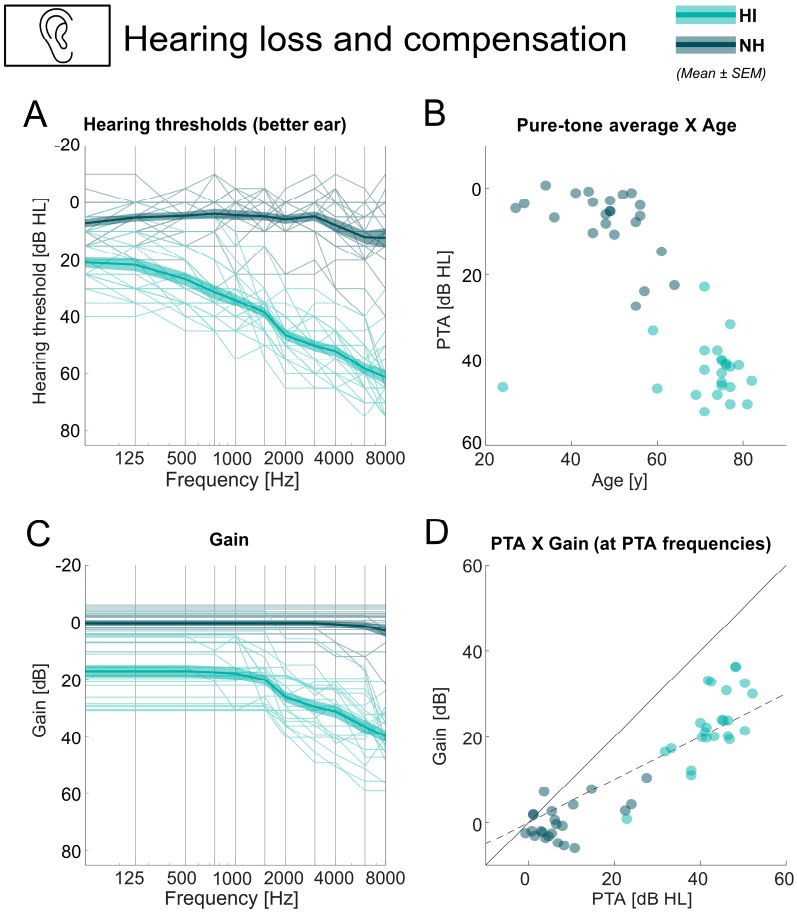


Figure 2: Hearing loss and compensation. Hearing aid users (HI) depicted in lighter green. Non-hearing-aid users (NH) depicted in darker green. **A**) Pure-tone air conduction thresholds of the frequency-wise better ear showed typical high-

frequency hearing loss for hearing aid users (HI), which was also found in some non-hearing-aid users (NH). **B)** Scatter plot of age and hearing loss depicts strong collinearity. **C)** The depicted individual linear gain was applied to compensate for elevated hearing thresholds. First, frequency-dependent gain was applied to equalize for hearing thresholds. Second, an adaptive procedure was used to estimate the stimulus specific hearing threshold after equalization. **D)** Scatter plot of PTA vs. gain (average at PTA frequencies) illustrates that the resulting gain reached approximately half of the PTA (dotted line).

Procedure and task

Experiments were conducted at Eriksholm Research Centre (Snekkersten, Denmark). Before the experiment, potential participants were asked to carefully read a document about the procedure and the related risks. At the beginning of their visit, the key points of this document were verbally repeated by the experimenter and participants gave written informed consent. The whole visit took between 2.5 and 3 hours.

A pure-tone audiogram was conducted at the beginning of the test session. Based on the audiogram, the individual gain was calculated (see Hearing loss compensation and adaptation). Then, the adaptive procedure for finding the IHL was conducted. After that, the participants were fitted with an EEG cap and two additional electrodes were secured to their mastoids. The electrodes were prepared outside the booth with electrolyte gel. Next, participants were seated inside the booth. The electrodes were connected to the EEG amplifier and the signals, as well as the offset-voltage, were checked. If there was no proper connection (indicated by an offset-voltage > 50 mV or noisy signals) additional gel was applied to the relevant electrode and the connection was checked again. As soon as all electrodes showed proper contact, the eye tracker was adjusted (such that

it reliably captured both eyes) and calibrated using the five-dot calibration procedure in Tobii Eye Tracker Manager.

The main experiment started with six training trials (see above), initiated by the participant pressing “Start” on the computer keyboard. After each trial, participants were prompted with two statements about the content of the newsclips. They were asked to indicate whether each statement was true using the computer keyboard buttons labelled with “Ja” (Yes) or “Nej” (No). The performance in this task will be referred to as ***target comprehension***. After the response to the second question, the next trial started automatically. After training, participants were asked if the loudness was at a comfortable level, which all participants confirmed. If needed, EEG electrodes or the eye tracker were re-adjusted. Next, participants underwent four blocks of testing, each consisting of 10 one-minute trials. Between each block, participants were offered a break. On average, the break lasted for around 3 minutes. During the break, the door of the booth was opened, and a fan was put in front to provide additional fresh air. During the breaks, participants were allowed to engage in a conversation with the experimenter. After the four experimental blocks, the EEG cap was removed from the participant’s head, and they were allowed to use the bathroom to wash their hair.

After participants returned from the bathroom, it was revealed that there was an additional surprise recall task of about 10 minutes. Participants were seated in the booth again and were presented with 200 distractor sounds. Half of the 200 sounds

399 were presented during the main experiment, whereas the other
400 half were not. After each sound, a question mark appeared on
401 the screen and participants were asked to indicate (using the
402 “Yes” and “No” keys) whether they recalled hearing that sound
403 in the main experiment. After they pressed a key, the next sound
404 was presented between 1 and 2 seconds later (jittered interval).
405 Behavioral performance on this task will be referred to as
406 ***distractor recall***.

407 Pupillometry

408 Task-evoked pupil dilation has been shown to depend on
409 cognitive effort (Hess & Polt, 1964; Kahneman & Beatty, 1966),
410 most likely due to phasic modulation of arousal, as controlled by
411 noradrenergic activity in the locus coeruleus (Aston-Jones &
412 Cohen, 2005; Joshi et al., 2016; Reimer et al., 2016). Effort and
413 attentional control have been closely linked in the literature
414 (Kahneman, 1973; for review: van der Wel & van Steenbergen,
415 2018). In animals it was shown that noradrenaline decreases
416 signal-to-noise ratio in firing patterns (Foote et al., 1975) and
417 sharpens frequency tuning (Manunta & Edeline, 2004) of
418 neurons in auditory cortex. Task-evoked pupil dilation has been
419 shown to reflect listening effort spent on the processing of to-be-
420 attended sentences in varying degrees of noise as well as
421 hearing loss (Ohlenforst et al., 2017, 2018). How the pupil
422 responds to task-irrelevant sounds is not as well documented.
423 There have been attempts to investigate how the pupil response
424 reflects salience, but only in response to stimuli that were
425 explicitly task relevant (Huang & Elhilali, 2017; Liao et al., 2016;

Zhao et al., 2019). While stimulus-driven attention is initially automatic and may affect phasic arousal (Rajkowski et al., 1994), a reorientation of voluntary attention back to the target requires cognitive control, hence additional effort. The more salient a distractor is, the more effort will be needed to reorient attention back to the target, hence, the greater the pupil dilation should be.

We recorded pupillometric data at a sampling rate of 1200 Hz. For each trial, the raw pupillometric data consisted of vectors, one for each eye, indexing pupil size in mm as well as missing values (Not a Number, NaN) at timepoints where the pupil could not be detected. Missing values could occur due to blinks, but also other noise in the image processing during the pupil detection. It is common practice to remove samples before and after consecutive missing values to remove artifacts related to blinks (Winn et al., 2018). We applied this rule only to consecutive missing values longer than 20 ms (i.e., 24 samples). For these instances, we set the samples within the time range of 35 ms before (i.e., 40 samples) and 100 ms after (i.e., 120 samples) to missing values (NaNs). After this procedure, we still found spikes in the data such that we applied another rule based on the magnitude of the difference between successive samples. For each trial, the standard deviation (SD) of the sample-to-sample difference was calculated and successive samples with an absolute difference that exceeded two SDs were both set to NaN. All NaNs were linearly interpolated and the pupillometric data were low-pass filtered

using a finite-impulse response filter with a one-second-wide Hamming window.

To extract the pupil size time course around the distractor sounds, the preprocessed pupil size data were epoched between -1 and 5 seconds relative to distractor onset. The mean of the pupil size in the baseline interval between -1 and 0 seconds (***baseline pupil size***) was subtracted from the pupil size time course. ***distractor-evoked mean pupil dilation (MPD)*** was defined as the mean within the interval between 1 and 3 seconds. only kept pupil responses in which less than 30% of the datapoints were missing before interpolation. Using this procedure, we obtained a *distractor-evoked MPD* in response to an average of 90.3% of the distractors. More than 80% of the distractor-evoked MPDs were available for 39 participants; between 50% and 80% were available for 7 participants; and only 21% were available for 1 participant. In sum, 11338 distractor-evoked MPDs were available for subsequent analysis.

Two covariates were extracted from pupil size to control for potential confounds, such as a subject with generally larger baseline pupil size also showing larger responses to distractors: ***trial baseline pupil size*** refers to the mean pupil size in the trial baseline period between -3 to 0 seconds relative to target speech onset. Trial mean pupil dilation (***trial MPD***) refers to the mean pupil size difference from Trial MPD during the trial between 0 and 60 seconds relative to target onset.

Electroencephalography

EEG data were recorded from 66 channels at a sampling rate of 8192 Hz. The preprocessing was mainly done with the FieldTrip-toolbox (version 20170321, Oostenveld et al., 2011), as well as custom-written code using MATLAB R2020b.

All the preprocessing was done on continuous data to avoid edge artifacts from filters. First, EEG data were re-referenced to the average of both mastoid electrodes. Second, the EEG data were resampled to 200 Hz. Next, a bidirectional band-stop filter (Butterworth, filter order: 6, cutoff frequencies: 47 & 53.2 Hz) and a bidirectional band-pass filter (FIR, cutoff frequencies: 1 Hz & 80 Hz, other parameters were FieldTrip default) were applied. The EEG data were epoched according to the markers sent at the beginning and the end of a trial. The EEG data were visually inspected, and EEG channels with obvious noise were identified, removed, and interpolated from neighboring channels. While 26 subjects had no bad channels at all, 20 subjects had between 1 and 3 bad channels and 1 subject had 5 bad channels.

Next, an independent component analysis (Makeig et al., 2004) was performed using the *runica* algorithm, as implemented in FieldTrip. The number of independent components (IC) was set to $N_{IC} = N_C – N_{BC} – 1$, where N_C is the number of EEG channels (i.e., 66) and N_{BC} is the number of bad channels. The IC were visually inspected both in the time and frequency domain, in addition to the topographic distribution of their weights. IC that could be clearly related to eyeblinks, eye movements, muscle

artifacts, heartbeat, or single channel noise were removed from the EEG data. On average, 16.62 (SD: 6.35) IC were removed.

Neural tracking of target speech and distractor sounds.

Neural tracking can be extracted from electrophysiological signals such as electroencephalography (EEG). It quantifies the strength of the overall neural response to a continuous stimulus feature such as the envelope. In numerous studies, it has been shown that attended auditory stimuli are neurally tracked more strongly than ignored stimuli (Ding & Simon, 2012; Fiedler et al., 2017, 2019; O'Sullivan et al., 2015; Tune et al., 2021). These studies mainly investigated the effect of sustained voluntary attention on neural tracking, but not the effect of brief, stimulus-driven disruption of attention. Only recently, researchers have begun to investigate how events in the to-be-ignored stream, such as changes in background noise (Khalighinejad et al., 2019), or sudden brief sounds (Holtze et al., 2021), affect the neural tracking of to-be-attended speech, also with regards to their computed salience (Straetmans et al., 2021). Here, we ask if the computed salience of to-be-ignored sounds correlates with the neural tracking of these sounds, as well as with the attention-capture-related perturbation of neural tracking of target speech.

We calculated neural tracking of target speech and distractor sounds to analyze how strongly the stimuli were represented in the EEG and how much this representation was modulated by attention. We followed the well-established approach of temporal response functions (TRF; Crosse et al., 2016) using the mTRF-toolbox version 2.3

(https://github.com/mickcrosse/mTRF-Toolbox). The approach involved training a linear forward model of the TRF between envelope onsets and the EEG. Envelope onsets were extracted by following procedures as in Fiedler et al., 2019. EEG data were low-pass filtered at 10 Hz using a bidirectional 16-order Butterworth filter.

We calculated neural tracking was the same for both target speech and distractor sounds the same way. We used a within-participant leave-one-out cross validation approach consisting of three steps. First, a forward model for the temporal response functions was trained on the envelope onsets and EEG data from all but one trial, second, the TRF was used to predict the EEG of the left-out trial, and lastly, the Pearson-correlation coefficient between the predicted and the measured EEG signal of the left-out trial was calculated. This correlation coefficient depends directly on the strength of the neural representation of the envelope onsets in the EEG, and hence, will be referred to as neural tracking.

Regarding the neural tracking of target speech (target tracking), first we calculated windowed neural tracking by predicting one-second-long windows of the EEG in steps of 0.1 seconds. Second, to estimate the effects of distractor presentation on target tracking, we predicted windows of target tracking two seconds pre- and post-distractor onset. The pre-distractor target tracking will be referred to as ***baseline target tracking***, and the pre-to-post change in target tracking will be referred to as ***Δtarget tracking***.

Regarding the distractor sounds, we calculated their neural tracking (***distractor tracking***) separately for each sound occurring within a trial based on the TRF trained on all sounds that occurred in all the other trials.

Salience computation

We used an existing, behaviorally validated model for acoustic salience (Huang & Elhilali, 2017). In brief, this model was developed to predict human decisions about which of two concurrent streams, each presented to one ear, is capturing the listeners attention moment-to-moment. Psychoacoustic features were extracted from the streams and underwent a factor analysis. Unsurprisingly, loudness was found to be the strongest driver for salience, but other factors such as harmonicity also play a role. Here we rms-equalized the distractor sounds (see Stimulus material), such that their variability in loudness was minimized and other features dominated the variability in overall computed salience.

We submitted the distractor sounds embedded in the stationary noise to the salience model as provided by Huang & Elhilali (2017). We used the default parameters. The salience model returned a salience time course, which we averaged across the 1-second window in which the distractor was inserted. Hence, we ended up with one value of ***computed salience*** for each distractor sound. These computed salience values were transformed into z-scores.

585 ## Statistical analyses

586 We tested if predictors (such as computed salience) affected our
587 response variables (such as neurophysiological responses or
588 behavior) in the manner we hypothesized. Additionally, we
589 controlled for several covariates, as described below.

590 We used a hierarchical series of (generalized) linear mixed
591 models to explain variance of dependent variables (i.e.,
592 responses) by independent variables (i.e., predictors).
593 Residuals of earlier models were used as predictors for
594 subsequent models. For example, we first tested if computed
595 salience predicted distractor tracking. After we found a
596 significantly positive estimate for modelled salience on distractor
597 tracking, we regressed it out from distractor tracking and used
598 the residuals of distractor tracking as a predictor for distractor-
599 evoked MPD. This way, a common modulation by computed
600 salience of both distractor tracking and distractor-evoked MPD
601 would not spuriously lead to a positive estimate of distractor
602 tracking on distractor-evoked MPD. Only if distractor tracking
603 explained additional variance beyond the variance that
604 computed salience explained directly, would a significant
605 relationship be revealed.

606 An overview of all statistical models, including their order, can
607 be found in tables 3-1 & 3-1 (supplemental material). Besides
608 the variables of interest (Fig. 3), we used several covariates,
609 such as trial number and baseline of the neurophysiological
610 responses. Each response was modelled in a stepwise
611 (generalized) linear mixed modelling approach with random and

fixed effects. We started with a basic model and added predictors one at a time (i.e., increased model). At each step, a likelihood-ratio test between the increased model and the current model was performed to determine if a predictor significantly improved the model fit. If this was the case, the predictor was added to the current model. When a predictor was added, all previously included predictors were removed one at a time (i.e., reduced model). If their removal did not significantly degrade the model fit (according to a likelihood-ratio test between the current and the reduced model), the predictor was removed from the model. That way we ensured that we accounted for the collinearity of some predictors such as age and PTA.

We ran this stepwise procedure on two levels: 1) the within-subject level: here we modelled the neurophysiological and behavioral response to single distractor sounds (Table 3-1, supplemental material), 2) the between-subject level, where predictors and response variables only varied between subjects, such as age, means of the neurophysiological and behavioral measures, as well as subject-wise random slopes from the within-subject level (Table 3-2, supplemental material). This way, we ensured that we could distinguish between interpretations on the within-subject level, such as “A relatively strong response to a unique distractor indicates that a subject was more distracted by this particular sound compared to another sound where a weaker response was found” and the between-subject level such as “A subject with a strong average

response to distractor sounds was more distractable than another subject showing a weaker response". By taking subject-wise random slopes of computed salience to the between-subject level (i.e., ***salience sensitivity***), we could test if a strong dependency of the neurophysiological responses on salience indicates higher distractibility, which we expect to impact behavioral performance (i.e., target comprehension and distractor recall).

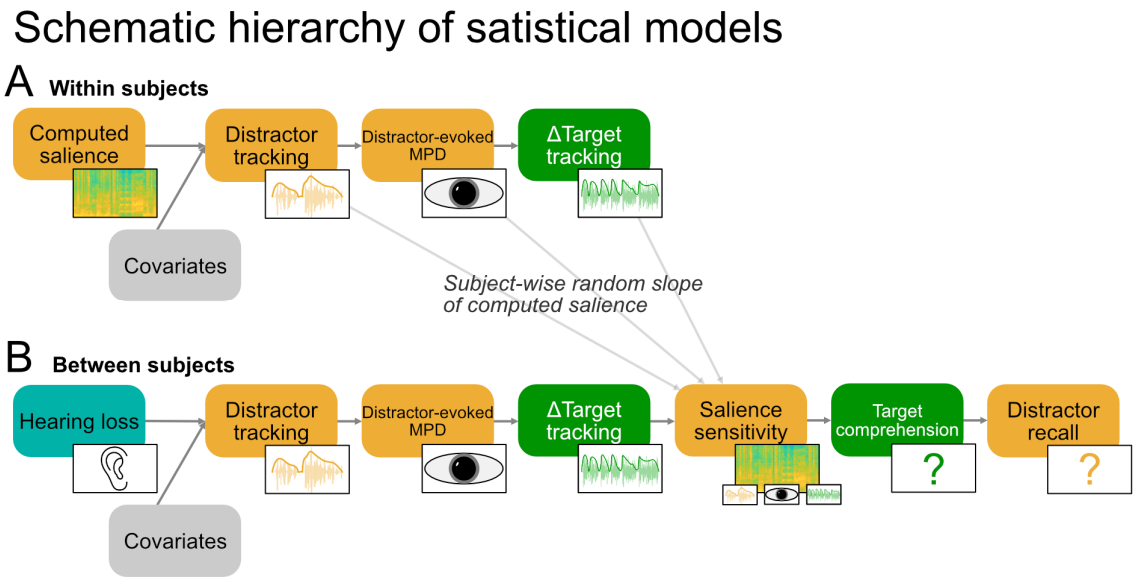


Figure 3: Hierarchical Statistical Modelling: A) Within-subject statistical modelling estimated the effect of stimulus features of single distractor sounds (computed salience) on single neurophysiological responses (distractor tracking, distractor-evoked mean pupil dilation (MPD), Δtarget tracking) as well as their interdependence, while controlling for baseline measures (covariates). **B)** Between-subject statistical modelling estimated the effect of hearing loss on subjects' average neurophysiological responses, their salience sensitivity (random slopes from within-subject models) and their common effect on behavioral performance, while controlling for covariates (age, average baseline measures).

Results

We asked human subjects to listen to continuous speech while distractor sounds of varying computed salience were presented at unpredictable times and locations. We statistically modelled

both the neurophysiological and behavioral responses to the distractors and target speech to test if they can be interpreted as a proxy of distraction. Furthermore, we tested if distractibility increases with hearing loss. We will present the results following the order in Figure 3. Results are summarized in Tables 3-1 and 3-2.

Distractor tracking

We found consistent temporal response functions to distractors (Fig. 4A). Above-zero distractor tracking was observed in all participants (Fig. 4B), which was confirmed by a significantly positive intercept (ß = .15, SE = $7.0*10^{-3}$, t(46) = 21.04, p = $3.03*10^{-25}$). On the within-subject level, the models for distractor tracking included trial number ($\chi^2(1) = 22.41$, p = $2.2*10^{-6}$) and computed salience ($\chi^2(1) = 10.18$, p = $1.4*10^{-3}$). Interestingly, distractor tracking decreased over the course of the experiment: trial number predicted distractor tracking with a negative estimate (ß = $-.75*10^{-3}$, SE = $1.6*10^{-4}$, t(12531) = -4.75, p = $2.1*10^{-6}$, partial η^2 = .0018). As hypothesized, distractor tracking was stronger for sounds with higher computed salience (Fig. 4C & 4D, ß = $5.8*10^{-3}$, SE = $1.8*10^{-3}$, t(12531) = 3.19, p = $1.4*10^{-3}$, partial η^2 = .0008). The final model explained only a small amount of the variance ($R^2_{adj.}$ =.048).

On the between-subject level, none of the predictors were significant, such that none of the inter-subject variance could be explained by the model ($R^2_{adj.}$ = 0). Even though we found considerably higher mean distractor tracking in HI compared to

NH, PTA predicted neither distractor tracking directly (Fig. 4B), nor its salience sensitivity (Fig. 4D), contrary to our predictions.

In sum, we found above-zero distractor tracking that depends positively on computed salience, while neither distractor tracking nor its salience sensitivity is modulated by PTA (Fig. 4).

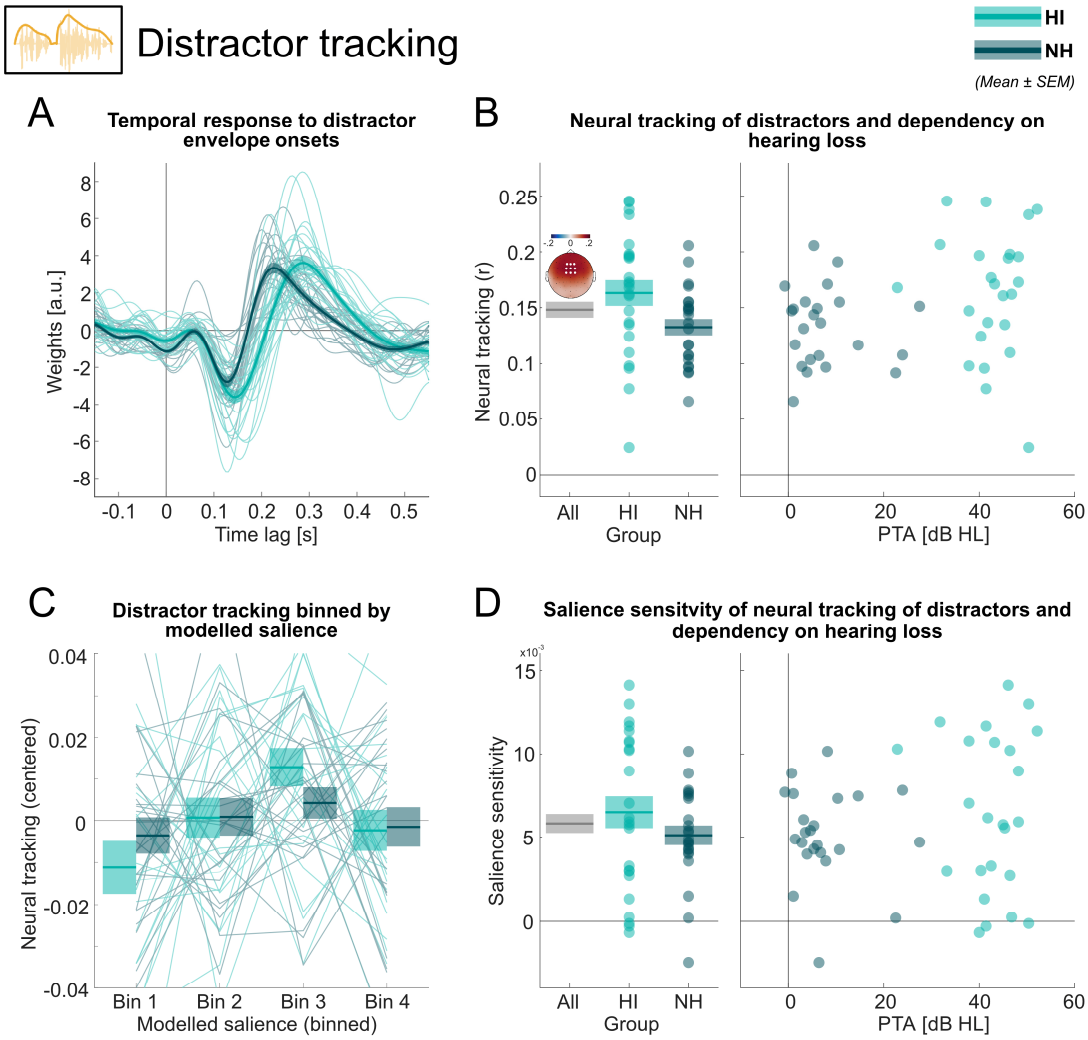


Figure 4: Neural tracking of distractor sounds (distractor tracking). Hearing aid users (HI) depicted in lighter green. Non-hearing-aid users (NH) depicted in darker green. Grey depicts average across all subjects. Thin lines and dots depict individual subject's data points. Thick lines and shaded areas depict the mean ±1 standard error of the mean (SEM) across subjects. **A)** Temporal response to distractor sounds as estimated from an onset representation of the sounds and EEG. **B)** Left: Average distractor tracking at fronto-central channels as estimated in a within-subject leave-one-out procedure. Right: Scatter plot of distractor tracking vs. hearing loss. **C)** Average distractor tracking (centered) binned by computed salience.

D) Left: Salience sensitivity of distractor tracking as estimated from fixed and random slopes in the linear models. Right: Scatter plot of salience sensitivity of distractor tracking vs. hearing loss.

Distractor-evoked mean pupil dilation

We found consistent pupil dilation in response to distractors (Fig. 5A). Above-zero distractor-evoked mean pupil dilation was observed in all participants (Fig. 5B), which was confirmed by a significantly positive intercept in the linear mixed model (ß = .043, SE = $2.6*10^{-3}$, t(45) = 16.36, p = $1.5*10^{-19}$).

On the within-subject level, the prediction model included pre-distractor baseline pupil size ($\chi^2(1) = 1823$, p = $2.2*10^{-308}$), computed salience ($\chi^2(1) = 56.37$, p = $6.0*10^{-14}$) and subject-wise random slopes of computed salience ($\chi^2(1) = 14.31$, p = $1.5*10^{-4}$). Relatively large pupils just before the distractor sound lead to smaller distractor-evoked MPD (and vice versa): pre-distractor baseline pupil size predicted distractor-evoked MPD with a negative estimate (ß = -.22, SE = $.50*10^{-3}$, t(11335) = -45.24, p = $2.2*10^{-308}$, partial η^2 = .17). As hypothesized, distractor-evoked MPD was larger for sounds with higher computed salience: computed salience predicted distractor-evoked MPD (Fig. 5C & 5D, ß = $8.7*10^{-3}$, SE = $1.6*10^{-3}$, t(11335) = 5.31, p = $1.1*10^{-7}$, partial η^2 = .0055). Additionally, there were considerable between-subject differences in how sensitive distractor-evoked MPD was to the modulation of computed salience (i.e., salience sensitivity): subject-wise random slopes of computed salience were significant ($ß_{std} = 7.81*10^{-3}$). Random slopes were taken to the between-subject level for

further investigations. The final model explained a considerable amount of within-subject variance ($R^2_{adj.}$ = .18).

On the between-subject level, the model only included trial MPD ($\chi^2(1) = 31.06$, $p = .002*10^{-5}$). Subjects with a stronger decrease in pupil size (during listening compared to trial baseline pupil size) showed a larger distractor-evoked MPD: trial MPD predicted distractor-evoked MPD with a negative estimate (ß = -.10, SE = .015, $t(45) = -6.63$, $p = 3.55*10^{-8}$, partial $\eta^2 = .48$). The final model explained a considerable amount of the between-subject variance ($R^2_{adj.}$ = .47). Against our hypothesis, PTA predicted neither distractor-evoked MPD directly (Fig. 5B), nor its salience sensitivity (Fig. 5D).

In sum, we found above-zero distractor-evoked MPD and that it positively depends on computed salience, while neither distractor-evoked MPD nor its salience sensitivity is affected by PTA (Fig. 5).

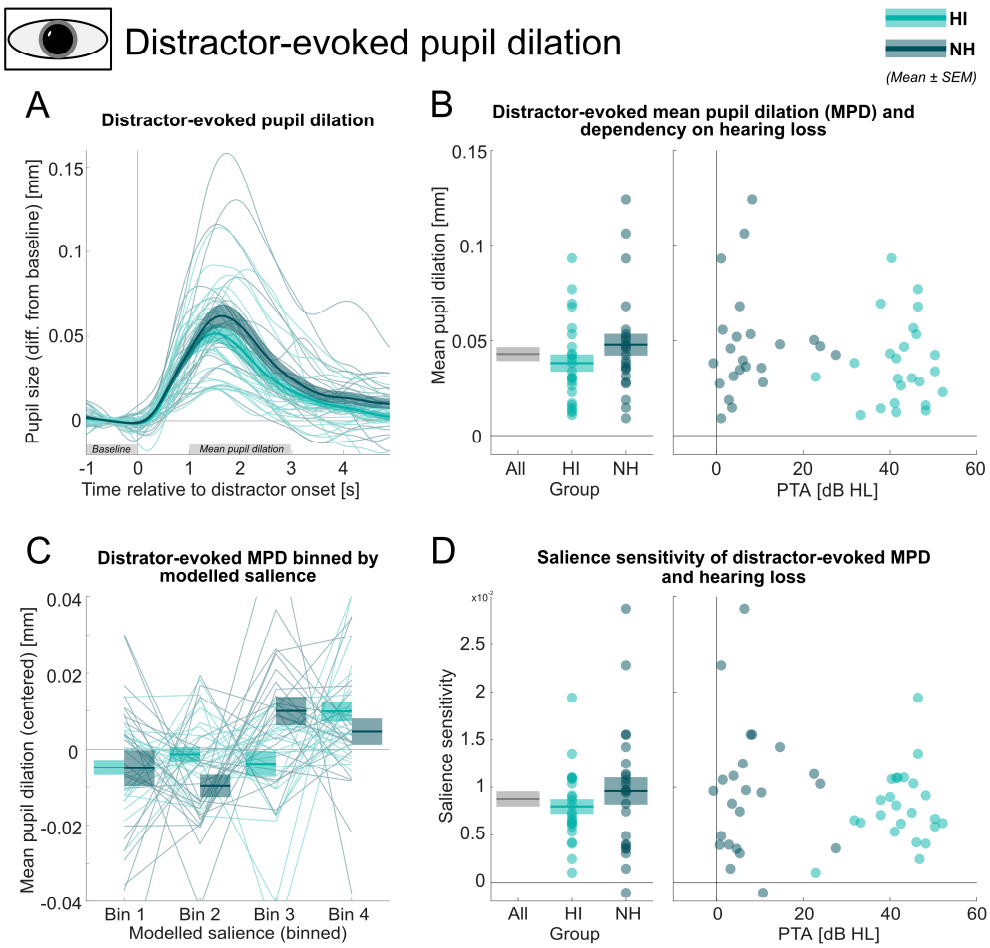


Figure 5: Distractor-evoked pupil dilation. Hearing aid users (HI) depicted in lighter green. Non-hearing-aid users (NH) depicted in darker green. Grey depicts average across all subjects. Thin lines and dots depict individual data points. Thick lines and shaded areas depict the mean ±1 standard error of the mean (SEM) across subjects. **A)** Distractor-evoked pupil dilation. **B)** Left: Average distractor-evoked mean pupil dilation (MPD). Right: Scatter plot of distractor-evoked MPD vs. hearing loss. **C)** Average distractor-evoked MPD (centered) binned by computed salience. **D)** Left: Salience sensitivity of d distractor-evoked MPD as estimated from fixed and random slopes in the linear models. Right: Scatter plot of salience sensitivity of distractor-evoked MPD vs. hearing loss.

Distractor-related change in neural target tracking

We found consistent temporal response functions to target speech (Fig. 6A inset). More importantly, we found consistent drops in target tracking post distractor onset (Fig. 6A). A drop in target tracking (negative Δtarget tracking) could be observed in almost all subjects (Fig. 6B), which was confirmed by a

significantly negative intercept in the linear mixed model (ß = -.022, SE = .0021, t(44) = -10.17, p = $4.00*10^{-13}$).

At the within-subject level, the model included baseline target tracking ($\chi^2(1) = 8312$, p = $.23*10^{-307}$), distractor tracking ($\chi^2(1) = 65.45$, p = $5.5*10^{-16}$) and a subject-wise random slope of computed salience. Relatively strong target tracking just before the distractor sound led to a larger drop in target tracking: baseline target tracking predicted Δtarget tracking with a negative estimate (ß = -.98, SE = .0089, t(12531) = -110.00, p = $2.23*10^{-308}$, partial η^2 = .50). Stronger distractor tracking was accompanied with less of a drop in target tracking: distractor tracking predicted Δtarget tracking (ß = .056, SE = .0069, t(12531)= 8.10, p = 5.92^{-16}, partial η^2 = .0055). Even though there was a trend towards a negative relationship between computed salience and Δtarget tracking (as we would have hypothesized, Fig. 6D), computed salience did not predict Δtarget tracking. However, the significant subject-wise random slopes of computed salience indicate that the strength of relation between computed salience and Δtarget tracking (i.e., salience sensitivity) varied considerably across participants ($ß_{std}$ = $5.19*10^{-3}$). The final model explained a considerable amount of variability in Δtarget tracking ($R^2_{adj.}$ = .49), which could be mainly attributed to pre-distractor target tracking.

On the between-subject level, the model included baseline target tracking ($\chi^2(1) = 13.66$, p = $2.0*10^{-4}$) and distractor-evoked MPD ($\chi^2(1) = 6.06$, p = .014). Subjects with higher baseline target tracking showed a larger drop in target tracking:

baseline target tracking predicted Δtarget tracking with a negative estimate (ß = -.39, SE = .086, t(44) = -4.54, p = $4.73*10^{-5}$, partial η^2 = .31). Subjects with larger distractor-evoked MPD showed a larger drop in neural target tracking: distractor-evoked MPD predicted Δtarget tracking with a negative estimate (ß = -.28, SE = .11, t(44) = -2.54, p = .015, partial η^2 = .12). The final model explained a moderate amount of the between-subject variance in Δtarget tracking ($R^2_{adj.}$ = .31).

In sum, we found that target tracking dropped when a distractor was presented, but the size of the drop does not depend on computed salience. Further neither the drop in target tracking after distractor presentation, nor its salience sensitivity is modulated by PTA (Fig. 6).

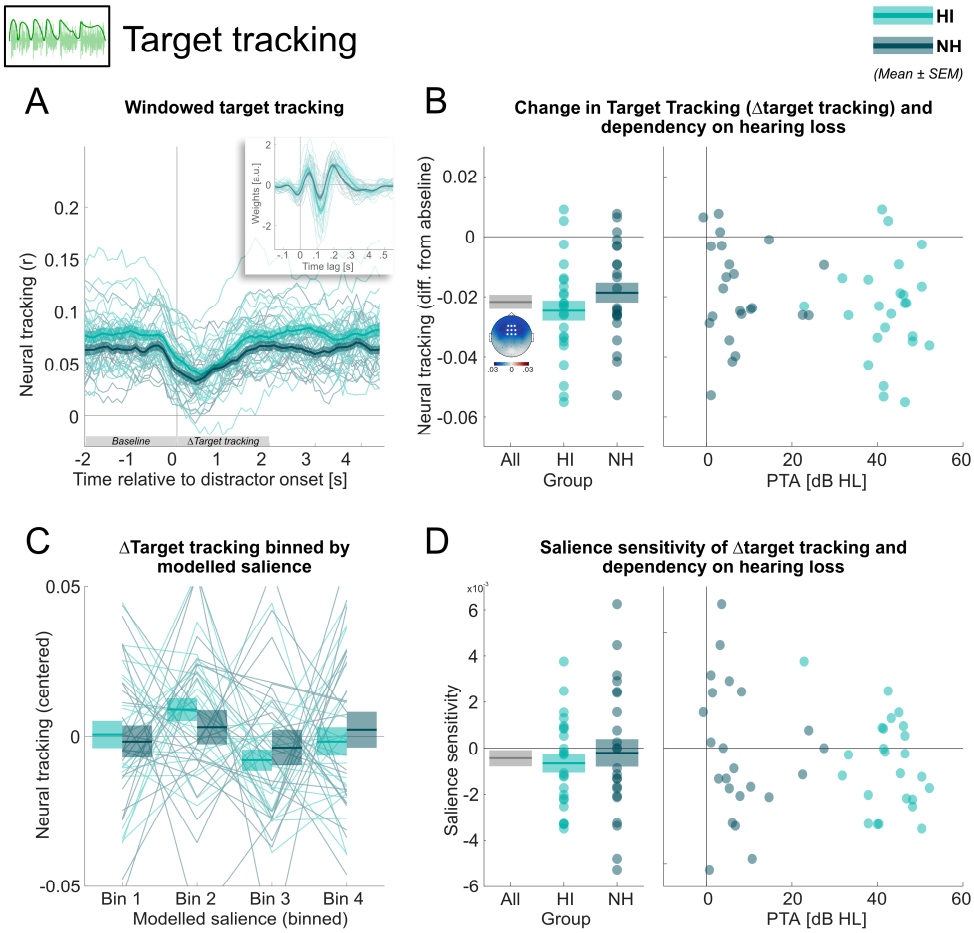


Figure 6: Distractor-related change in neural tracking of target speech (Δtarget tracking). Hearing aid users (HI) depicted in lighter green. Non-hearing-aid users (NH) depicted in darker green. Grey depicts average across all subjects. Thin lines and dots depict individual subject's data points. Thick lines and shaded areas depict the mean ±1 standard error of the mean (SEM) across subjects. **A)** Windowed target tracking as estimated in a within-subject leave-one-out procedure and a sliding prediction window. Inset: Temporal response to target speech as estimated from an onset representation of speech signal and EEG. **B)** Left: Average Δtarget tracking at fronto-central channels. Right: Scatter plot of Δtarget tracking vs. hearing loss. **C)** Average Δtarget tracking (centered) binned by computed salience. **D)** Left: Salience sensitivity of Δtarget tracking as estimated from fixed and random slopes in the linear models. Right: Scatter plot of salience sensitivity of Δtarget tracking vs. hearing loss.

Target comprehension

The percentage of correctly answered comprehension questions was significantly above chance (> 59%) for each participant, which indicates everyone was able to follow the task instructions (Fig. 7A). This was also confirmed by a significantly

positive intercept (ß = 1.84, SE = .0049, t(41) = 37.28, p = $2.97*10^{-33}$), which corresponds to a mean comprehension of 86.3 %.

The model included PTA ($\chi^2(1) = 13.66$, p = $.2*10^{-3}$), baseline target tracking ($\chi^2(1) = 30.5$, p = $.358*10^{-8}$), distractor-evoked MPD ($\chi^2(1) = 5.76$, p = .016), the random slopes by subject of trial number on baseline pupil size ($\chi^2(1) = 12.06$, p = .0005) and the random slopes by subject of computed salience on Δtarget tracking ($\chi^2(1) = 7.11$, p = .0076). Participants with higher PTA showed worse target comprehension (Fig. 7B, ß = -.316, SE = .052, t(41) = -6.09, p = $3.26*10^{-7}$). At the point of the intercept, this corresponds to a change of roughly 3.7% per SD in PTA (SD = 18.9 dB). Interestingly, participants with relatively high baseline target tracking (given their age) showed better target comprehension: baseline target tracking (residualized for age) predicted target comprehension (Fig. 7C, ß =.278, SE = .051, t(41) = 5.48, p = $2.34*10^{-6}$). At the point of the intercept, this corresponds to a change of roughly 3.3% per SD in baseline target tracking (SD = 0.023). Crucially, participants with a larger distractor-evoked MPD showed worse target comprehension: distractor-evoked MPD predicted target comprehension with a negative estimate (Fig. 7D, ß = -.127, SE = .050, t(41) = -2.56, p = 0.0144). At the point of the intercept, this corresponds to a change of roughly 1.5% per SD in distractor-evoked MPD (SD = 0.018 mm). Participants with a steeper decrease in baseline pupil size over the course of the experiment showed better target comprehension: the random slope of trial number on

baseline pupil size predicted comprehension with a negative estimate (ß = -.207, SE = .052, t(41) = -3.98, p = $2.75*10^{-4}$). At the point of the intercept, this corresponds to a change of roughly 2.5% per SD in random slope of trial number on baseline pupil size (SD = 0.0073 mm/trial). Importantly, participants: with higher salience sensitivity in Δtarget tracking showed worse target comprehension: the random slope of computed salience on Δtarget tracking predicted target comprehension (Fig. 7E, ß = .132, SE = .0535, t(41) = 2.47, p = 0.0177). At the point of the intercept, this corresponds to a change of roughly 1.5% per SD in random slope of computed salience on Δtarget tracking (SD = 0.0023). The final model explained a considerable amount of between-subject variance in target comprehension ($R^2_{adj.}$ = .43).

In sum, we found that higher PTA, higher distractor-evoked MPD, lower baseline target tracking (given age), and higher salience sensitivity in Δtarget tracking lead to worse target comprehension.

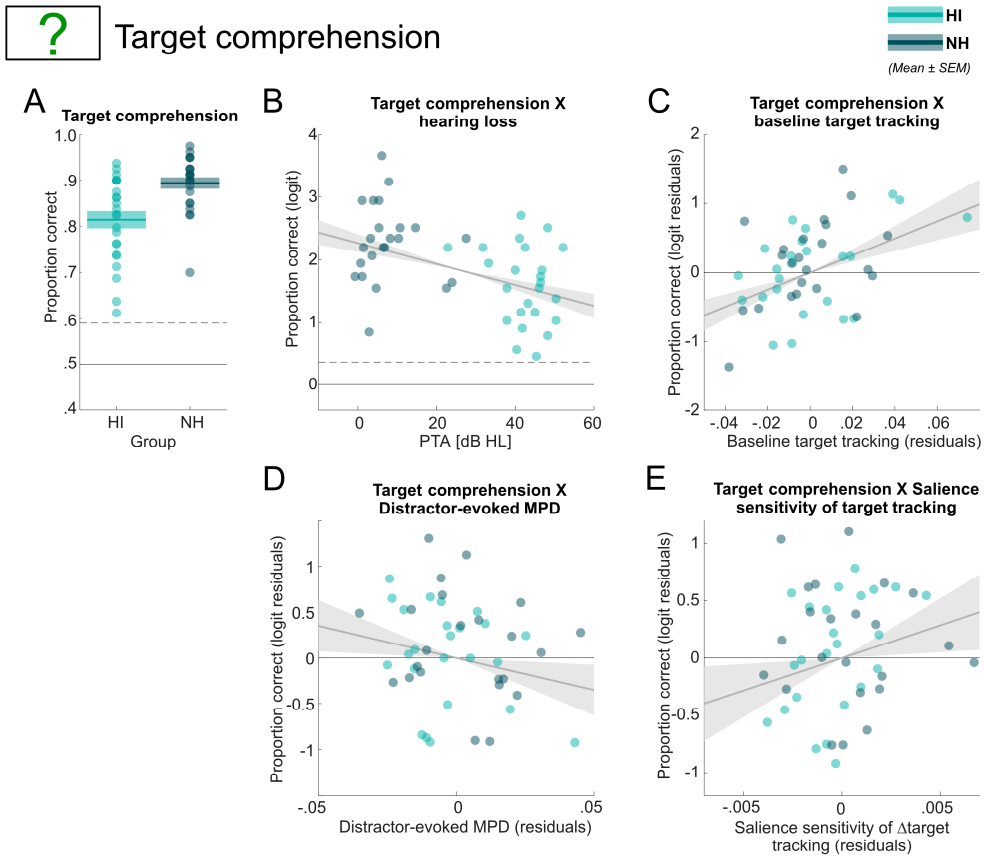


Figure 7: Correctly answered comprehension questions (Target comprehension). Hearing aid users (HI) depicted in lighter green. Non-hearing-aid users (NH) depicted in darker green. Thin lines and dots depict individual subject's data points. Thick lines and shaded areas depict the mean ±1 standard error of the mean (SEM) across subjects. Grey lines depict slope estimates from generalized linear mixed model together with their confidence bands (shaded areas). **A)** Proportion of correctly answered comprehension questions. Dotted line depicts the empirical chance level ($p < 0.05$) on an individual level. **B)** Scatter plot of logit-transformed target comprehension vs. PTA. **C)** Scatter plot of logit-transformed target comprehension (residuals) vs. baseline target tracking (residuals). **D)** Scatter plot of logit-transformed target comprehension (residuals) vs. distractor-evoked MPD (residuals). **E)** Scatter plot of logit-transformed target comprehension (residuals) vs. salience sensitivity of Δtarget tracking (residuals).

Distractor recall

The percentage of correctly recalled distractors rarely exceeded binomial chance at an individual level (<56%, Fig. 7A). Participants were not able to perform the task very well.

On the between-subject level, for the prediction of recall the model only included age ($\chi^2(1) = 9.92$, $p = .160*10^{-2}$). The apparent group difference between HI and NH could be explained by age: age predicted recall with a negative estimate (Fig. 8B, ß = $.425*10^{-2}$, SE = $.134*10^{-2}$, $t_{(44)} = 3.16$, $p = .285*10^{-2}$), The model explained a moderate amount of the inter-subject variance in recall ($R^2_{adj.} = .139$).

In sum, neither distractor-tracking, nor distractor-evoked mean pupil dilation during the experiment influenced later distractor recall, probably because distractor recall was not far above chance, even in the normal hearing group.

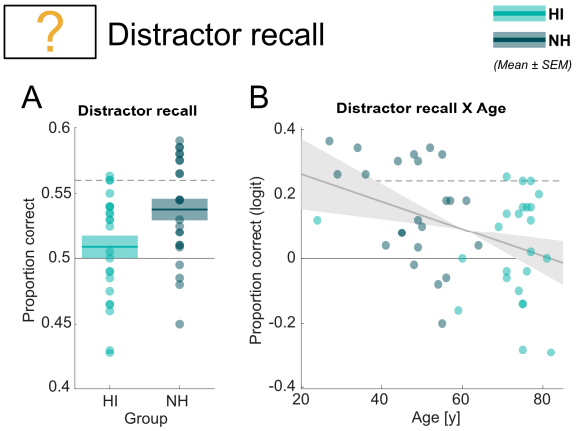


Figure 8: Correctly recalled distractor sounds (distractor recall). Hearing aid users (HI) depicted in lighter green. Non-hearing-aid users (NH) depicted in darker green. Thin lines and dots depict individual subject's data points. Thick lines and shaded areas depict the mean ±1 standard error of the mean (SEM) across subjects. Grey lines depict slope estimates from generalized linear mixed model together with their confidence bands (shaded areas). **A)** Proportion of correctly recalled distractor sounds. Dotted line depicts the empirical chance level ($p < 0.05$) on an individual level. **B)** Scatter plot of logit-transformed distractor recall vs. age.

Discussion

We tested whether neurophysiological responses to irrelevant sounds (distractors) indicate the degree to which attentional focus on continuous target speech is captured by such distractors, in both hearing-aid users and non-hearing-aid users.

We found that both neural tracking of the distractors (distractor tracking) and distractor-evoked mean pupil dilation (MPD) could be predicted from salience of the distractors. We found a distractor-related disruption in neural tracking of target speech (Δtarget tracking), which did not directly depend on computed salience, but did predict behavioral performance. Participants who showed a stronger distractor-related drop in target tracking for more salient sounds (i.e., higher salience sensitivity) performed worse in target speech comprehension. Similarly, participants with stronger distractor-evoked MPD showed greater distractor-related disruption of target tracking, and performed worse in target speech comprehension. While (compensated) hearing loss did not have a direct effect on the distraction-related responses, a general decrease in target speech comprehension due to hearing loss was observed.

Neural tracking and pupil dilation are salience-dependent indicators of distraction

We used recent, behaviorally validated models to estimate the salience of individual distractors (Huang & Elhilali, 2017, 2020). This computed salience is, by definition, a proxy of the degree to which sound captures attention in a stimulus-driven fashion (Kaya & Elhilali, 2014). As we hypothesized, more salient sounds led to stronger neural tracking of distractors and larger distractor-evoked pupil dilation, and the larger these two distractor-related responses, the worse was target comprehension. Together, these findings indicate that these two neurophysiological measures (distractor tracking and distractor-

evoked MPD) each reflect how distracting a task-irrelevant stimulus is.

These measures allow us to index momentary distraction while subjects are focusing on a continuous stimulus. This may help to overcome challenges in the behavioral assessment of distraction: sustained attention does not need to be interrupted by a distractor-related task, which would interrupt the target task and would make subjects perceive distractors as task relevant. The neurophysiological responses we measure here enable monitoring of distraction in the background without any interference with the target task.

Recent studies yield similar findings. In an EEG study by Holtze et al. (2021) subjects were listening to concurrent continuous speech while their own name was inserted into the to-be-ignored stream. In contrast to our findings, they observed that neural tracking of both the to-be-attended and the to-be-ignored streams increased after the name occurrence. While one's own name is undoubtedly a very strong attention- capturing stimulus (Cherry, 1953; Moray, 1959), it is a unique example. Here we enhanced generalizability of results to other stimuli by using non-repetitive, real-world sounds of high spectro-temporal diversity. A similar approach was taken by Straetmans et al. (2021). Similar to our findings, they observed a drop in neural target tracking, but both for the to-be-attended and to-be-ignored speech after the occurrence of a distractor sound. They also observed that the P3-component in the ERP to the distractor sounds (300 ms after target onset and thought to

reflect attention orienting among others) was positively correlated with computed salience. Here we extend these results with pupillometry and instead of studying the event-related potential relative to sound onset, we examine the temporal response function in response to the sound envelope, which not only captures the initial onset of a sound but also its temporal modulation.

A large portion of the variance in the neurophysiological responses was left unexplained. On the one hand, this may be due to EEG and pupil size being inherently noisy signals, on the other hand, salience modelling may still have room for improvement (Duangudom & Anderson, 2007; Kalinli & Narayanan, 2007; Kaya & Elhilali, 2012; Kayser et al., 2005; Kim et al., 2014). Temporal and spatial context of a sound could be incorporated into salience modelling: for example, the sound of running water may not be very salient if you are out hiking, but would be very salient if you heard it in the walls of your home. Here we only accounted for several seconds and only within the same spatial location and context.

A further reason for substantial unexplained variance may be low variability in the predictors. By presenting all distractors at the same sound level, we minimized their variability in loudness, which is in the strongest determinant of salience in the model we used (Huang & Elhilali, 2017). We kept distractors equal in level, but allowed their spectrotemporal features to vary. This allowed us to study to isolate attentional aspects in the neurophysiological responses and not tap into sensory aspects

alone. This is also confirmed by the fact that sounds scoring lowest in computed salience evoked consistent pupil dilation. The equalization of the level of the distractor sounds may have resulted in a rather subtle manipulation of salience, such that the effect of acoustic salience on neurophysiological responses was underestimated. The fact that we still found an effect shows that current models of salience tap into aspects of sound beyond its level, that relate to stimulus-driven attention.

Subjective factors may also play a strong role here: for example, a cat may be more salient to someone who is afraid of cats than someone who has a neutral attitude towards cats. These beyond-acoustic aspects of salience may be captured in the statistical dependencies in the neurophysiological measures and their ability to predict behavior within-subjects. For example, a sound that that is subjectively highly salient should both lead to strong neural tracking and pupil dilation, such that additional variance (beyond computed salience) in pupil dilation may be explained by distractor tracking. Although we found a connection between neural distractor tracking and neural target tracking within subjects, on the between-subject level a larger pupil dilation and a stronger salience sensitivity of (the change in) neural target tracking indicated worse comprehension. Together, these findings provide evidence for a correlation between the neurophysiological measures that goes beyond what computed salience can explain.

Under the assumption that a more distracting sound is better encoded into memory than is a less distracting sound, we

assumed that a stronger neurophysiological response to a distractor should make its recall in a surprise recall task more likely. However, we found that distractor recall performance could not be predicted by the neurophysiological responses, but only depended on age. Between-subject variability in distractor recall performance was low and subjects barely performed above chance. Our conclusion is that the surprise recall task may have been too difficult and hence was not sensitive to the salience manipulations.

In sum, we provide evidence that stimulus-driven attention modulates neurophysiological responses during sustained attention to a target. Stimulus-driven attention is synonymous with distraction here, since the salient sounds were task-irrelevant but their presence had a consequence on behavioral performance (Makov et al., 2022).

Neurophysiological measures of distraction are orthogonal to hearing thresholds

We hypothesized that subjects with higher hearing thresholds would show enhanced distraction due to both less sharp attentional filters and fewer cognitive resources for attentional control. This should be reflected in stronger neurophysiological responses to distractors and worse comprehension even after compensation for hearing thresholds (i.e., with hearing aids and tailored amplification of sound). We did not find a direct link between PTA and neurophysiological measures, but we found that higher PTA still led to worse comprehension. Within our design, it is not clear whether worse performance was due to

distraction or if we would have received the same results with target speech in stationary noise only (i.e., without distractors), as might be expected, given general hearing loss. Still, we found neurophysiological responses to distractors that predict speech comprehension, such that we conclude that we captured neurophysiological markers which are orthogonal to hearing thresholds and explain variance beyond conventional measures of hearing loss such as PTA.

Finding measures that improve the explanation of inter-individual differences in speech-in-noise performance under high ecological validity are desirable for diagnostics and treatment of hearing challenges. Conventional speech-in-noise tests are performed in stationary noise, such that the factor of distractibility has not been considered to a greater extent, but it may be an important aspect in the design and fitting of hearing prosthesis. Furthermore, recent developments towards hybrid workplaces lead to the development of noise-adaptive hearing solutions (EPOS, 2023), which could benefit from objective measures of distraction.

Conclusion

By relating neurophysiological responses evoked by task-irrelevant stimuli to both a current model for salience a well as behavioral performance, we provide evidence that both neural tracking of distractor sounds, and distractor-evoked pupil dilation, can be used as a proxy of momentary distraction. Hence, these measures may be suited to monitor the neural

processing of distractors, which cannot be assessed behaviorally in such a direct fashion.

Acknowledgements

EPOS Group A/S supported conference presentations. Torben Christiansen (EPOS Group A/S) contributed to the study design with discussion of potential application in noise-cancelling headsets. A first draft was proof-read by Bethany Plain.

Conflicts of interest

Eriksholm Research Centre is part of Oticon A/S, a manufacturer of hearing aids with commercial interests. Dissemination of study results (conference travel costs) was financially supported by EPOS Group A/S, a manufacturer of headsets with commercial interests.

References

Alickovic, E., Lunner, T., Wendt, D., Fiedler, L., Hietkamp, R., Ng, E. H. N., & Graversen, C. (2020). Neural Representation Enhanced for Speech and Reduced for Background Noise With a Hearing Aid Noise Reduction Scheme During a Selective Attention Task. *Frontiers in Neuroscience*, *14*, 846. https://doi.org/10.3389/fnins.2020.00846

Alickovic, E., Ng, E. H. N., Fiedler, L., Santurette, S., Innes-Brown, H., & Graversen, C. (2021). Effects of Hearing Aid Noise Reduction on Early and Late Cortical Representations of Competing Talkers in Noise.

Frontiers in Neuroscience, *15*, 636060. https://doi.org/10.3389/fnins.2021.636060

Ananthabhotla, I., Ramsay, D. B., & Paradiso, J. A. (2019). HCU400: An Annotated Dataset for Exploring Aural Phenomenology Through Causal Uncertainty. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 920–924. https://doi.org/10.1109/ICASSP.2019.8683147

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403–450.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00604

David, M., Tausend, A. N., Strelcyk, O., & Oxenham, A. J. (2018). Effect of age and hearing loss on auditory stream

segregation of speech sounds. *Hearing Research*, *364*, 118–128. https://doi.org/10.1016/j.heares.2018.03.017

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

Duangudom, V., & Anderson, D. (2007). *Using Auditory Saliency to Understand Complex Auditory Scenes*. 6.

EPOS. (2023). *Workplace of the Future*. https://www.eposaudio.com/contentassets/5e2bf080914f4b359a7ff4ed744fa324/epos-the-workplace-of-the-future.pdf

Fiedler, L., Seifi Ala, T., Graversen, C., Alickovic, E., Lunner, T., & Wendt, D. (2021). Hearing Aid Noise Reduction Lowers the Sustained Listening Effort During Continuous Speech in Noise—A Combined Pupillometry and EEG Study. *Ear & Hearing*, *42*(6), 1590–1601. https://doi.org/10.1097/AUD.0000000000001050

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., & Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of Neural Engineering*, *14*(3), 036020. https://doi.org/10.1088/1741-2552/aa66dd

Fiedler, L., Wöstmann, M., Herbst, S. K., & Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions.

NeuroImage, *186*, 33–42. https://doi.org/10.1016/j.neuroimage.2018.10.057

Foote, S. L., Freedman, R., & Oliver, A. P. (1975). Effects of putative neurotransmitters on neuronal activity in monkey auditory cortex. *Brain Research*, *86*(2), 229–242. https://doi.org/10.1016/0006-8993(75)90699-x

Gygi, B., & Shafiro, V. (2019). *Database of Environmental Sounds for Research Activities (DESRA)—Sound files and metadata.* [Computer software]. https://doi.org/10.5281/zenodo.2622626

Hess, E. H., & Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Holtze, B., Jaeger, M., Debener, S., Adiloğlu, K., & Mirkovic, B. (2021). Are They Calling My Name? Attention Capture Is Reflected in the Neural Tracking of Attended and Ignored Speech. *Frontiers in Neuroscience*, *15*, 643705. https://doi.org/10.3389/fnins.2021.643705

Huang, N., & Elhilali, M. (2017). Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, *141*(3), 2163–2176. https://doi.org/10.1121/1.4979055

Huang, N., & Elhilali, M. (2020). Push-pull competition between bottom-up and top-down auditory attention to natural soundscapes. *ELife*, *9*, e52984. https://doi.org/10.7554/eLife.52984

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, *89*(1), 221–234.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Citeseer.

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kalinli, O., & Narayanan, S. (2007). A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech. *Interspeech*, 4.

Kaya, E. M., & Elhilali, M. (2012). A temporal saliency map for modeling auditory attention. *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, 1–6. https://doi.org/10.1109/CISS.2012.6310945

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00327

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160101. https://doi.org/10.1098/rstb.2016.0101

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map. *Current Biology*, *15*(21), 1943–1947. https://doi.org/10.1016/j.cub.2005.09.040

Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Adaptation of the human auditory cortex to changing background noise. *Nature Communications*, *10*(1), 2509. https://doi.org/10.1038/s41467-019-10611-4

Kim, K., Lin, K.-H., Walther, D. B., Hasegawa-Johnson, M. A., & Huang, T. S. (2014). Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognition Letters*, *38*, 78–85. https://doi.org/10.1016/j.patrec.2013.11.010

Liao, H.-I., Kidani, S., Yoneya, M., Kashino, M., & Furukawa, S. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychonomic Bulletin & Review*, *23*(2), 412–425. https://doi.org/10.3758/s13423-015-0898-0

Lybarger, S. (1944). *US Patent Application SN 532* (Patent SN 532, 278).

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Sciences*, *8*(5), 204–210. https://doi.org/10.1016/j.tics.2004.03.008

Makov, S., Pinto, D., Har-shai Yahav, P., Miller, L. M., & Zion Golumbic, E. (2022). “Unattended, distracting or irrelevant”: Theoretical implications of terminological choices in auditory selective attention research. *Cognition*, 105313. https://doi.org/10.1016/j.cognition.2022.105313

Manunta, Y., & Edeline, J.-M. (2004). Noradrenergic Induction of Selective Plasticity in the Frequency Tuning of Auditory Cortex Neurons. *Journal of Neurophysiology*, *92*(3), 1445–1463. https://doi.org/10.1152/jn.00079.2004

May, T., Kowalewski, B., & Dau, T. (2018). Signal-to-Noise-Ratio-Aware Dynamic Range Compression in Hearing Aids. *Trends in Hearing*, *22*, 233121651879090. https://doi.org/10.1177/2331216518790903

Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60. https://doi.org/10.1080/17470215908416289

Nielsen, C., & Henriksen, T. B. (2022). 40 years of sound processing as described by hearing aid users. *Hearing Review*, *29*(2), 10–11.

Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., & Lunner, T. (2018). Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hearing Research*, *365*, 90–99. https://doi.org/10.1016/j.heares.2018.05.003

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., & Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79. https://doi.org/10.1016/j.heares.2017.05.012

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Oxenham, A. J. (2008). Pitch Perception and Auditory Stream Segregation: Implications for Hearing Loss and Cochlear Implants. *Trends in Amplification*, *12*(4), 316–331. https://doi.org/10.1177/1084713808325881

Peelle, J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear & Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Research Bulletin*, *35*(5), 607–616. https://doi.org/10.1016/0361-9230(94)90175-9

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and

cholinergic activity in cortex. *Nature Communications*, *7*(1), 1–7.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, *10*(1), e1003412. https://doi.org/10.1371/journal.pcbi.1003412

Serences, J. T., & Kastner, S. (2014). *A Multi-level Account of Selective Attention* (A. C. (Kia) Nobre & S. Kastner, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199675111.013.022

Straetmans, L., Holtze, B., Debener, S., Jaeger, M., & Mirkovic, B. (2021). Neural tracking to go: Auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering*, *18*(6), 066054. https://doi.org/10.1088/1741-2552/ac42b5

Tune, S., Alavash, M., Fiedler, L., & Obleser, J. (2021). Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. *Nature Communications*, *12*(1), 4533. https://doi.org/10.1038/s41467-021-24771-9

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in Hearing*, *22*, 233121651880086. https://doi.org/10.1177/2331216518800869

Zhao, S., Yum, N. W., Benjamin, L., Benhamou, E., Yoneya, M., Furukawa, S., Dick, F., Slaney, M., & Chait, M. (2019). Rapid Ocular Responses Are Modulated by Bottom-up-Driven Auditory Salience. *The Journal of Neuroscience*, *39*(39), 7703–7714. https://doi.org/10.1523/JNEUROSCI.0776-19.2019

Supplemental material

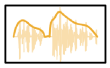
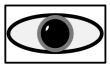
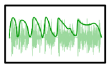
Predictor						
Intercept (+Random intercept)	$p = 2.0*10^{-6}$	$p = 2.3*10^{-109}$	$p = 3.0*10^{-96}$	$p = 9.0*10^{-35}$	$p = 1.6*10^{-12}$	
Trial number	$P = 3.5*10^{-16}$ η^2 (part.) = **.17**; RS		$p = 2.1*10^{-6}$ η^2 (part.) = **.0018**	RS		
Baseline pupil size				$p = 2.0*10^{-308}$ η^2 (part.) = **.17**		
Baseline target tracking					$p = 2.0*10^{-308}$ η^2 (part.) = **.49**	
Modelled Salience			$p = 1.4*10^{-3}$ η^2 (part.) = **.0008**	$p = 1.0*10^{-7}$ η^2 (part.) = **0.006**; RS	RS	
Distractor Tracking					$p = 5.9*10^{-16}$ η^2 (part.) = **0.006**	
Distractor-evoked pupil dilation						
Δtarget tracking						
R^2 (adjusted)	**0.24**	**0.015**	**0.05**	**0.18**	**0.49**	**0.024**
	Baseline pupil size	Baseline target tracking	**Distractor Tracking** 	**Distractor-evoked pupil dilation** 	**Δtarget tracking** 	**Distractor Recall**
	Response					

Table 3-1: Within-subject models: We used a hierarchical approach to model dependent variables of interest and covariates (Response) based on independent variables (predictors). Each column depicts one model. Grey cells depict tested predictors that were not included in the final model. Orange cells depict predictors that were included in the final model with a positive intercept. Blue cells depict predictors that were included in the final model with a

negative intercept. Text in cells of included predictors describe the p-values and partial eta-squared of the predictor in the final model. Residual of earlier modelled responses were used as predictors for models of later responses. RS marks significant random slopes.

Predictor	Trial baseline pupil size	Trial MPD	Baseline target tracking	Distractor Tracking	Distractor-evoked pupil dilation	Δtarget tracking	RS: TrialNr on baseline pupil size	RS: TrialNr on Distractor-evoked pupil dilation	RS: Mod. Salience on Dist.-evoked pupil dilation	RS: Mod. Salience on Δtarget tracking	Answers Correct	Distractor Recall
Intercept	$p = 3.4*10^{-36}$	$p = 1.0*10^{-7}$	$p = 2.0*10^{-27}$	$p = 3.0*10^{-25}$	$p = 1.5*10^{-20}$	$p = 4.0*10^{-13}$					$p = 8.0*10^{-34}$ ~86%	$p = 7.4*10^{-5}$ ~52.3%
Age	$p = 6.7*10^{-6}$ η^2 (part.) = **.36**		$p = 5.0*10^{-3}$ η^2 (part.) = **.16**									$p = 3.0*10^{-3}$ ~1.7%/16y
PTA		$p = 1.1*10^{-3}$ η^2 (part.) = **.20**									$p = 3.26*10^{-7}$ ~3.7%/19dB	
Trial baseline pupil size										$p = 2.6*10^{-2}$ η^2 (part.) = **.10**		
Trial MPD					$p = 3.6*10^{-8}$ η^2 (part.) = **.48**				$p = 8.3*10^{-8}$ η^2 (part.) = **.35**			
Baseline target tracking						$p = 4.7*10^{-5}$ η^2 (part.) = **.31**					$p = 2.34*10^{-6}$ ~3.3%/SD	
Distractor Tracking												
Distractor-evoked pupil dilation						$p = 2.5*10^{-2}$ η^2 (part.) = **.11**			$p = 6.9*10^{-4}$ η^2 (part.) = **.22**		$p = .014$ ~1.5%/.02mm	
Δtarget tracking												
RS: TrialNr on baseline pupil size								$p = 1.5^{-13}$ η^2 (part.) = **.70**			$p = 2.76*10^{-4}$ ~2.5%/SD	
RS: TrialNr on MPD												
RS: Mod. Salience on Dist.-evoked pupil dilation												
RS: Modelled Salience on Δtarget tracking											$p = .018$ ~1.5%/SD	
Answers Correct												
R^2(adjusted)	**0.34**	**0.19**	**0.14**	**0**	**0.48**	**0.31**	**0**	**0.69**	**0.43**	**0.082**	**0.43**	**0.14**

Response

Table 3-2: Between-subject models: We used a hierarchical approach to model dependent variables of interest and covariates (Response) based on independent variables (predictors). Each column depicts one model. Grey cells depict tested predictors that were not included in the final model. Orange cells depict predictors that were included in the final model with a positive intercept. Blue cells depict predictors that were included in the final model with a negative intercept. Text in cells of included predictors describe the p-values and partial eta-squared of the predictor in the final model. Residual of earlier modelled responses were used as predictors for models of later responses. RS refers to random slopes.