Title:
Predictable sequential structure enhances auditory sensitivity in clinical audiograms

Authors:
Nadège Marin[1,2], Grégory Gérenton[1], Hadrien Jean[3], Nihaad Paraouty[3], Nicolas Wallaert[3], Diane S. Lazard[1,4], Luc H. Arnal[1*], Keith B. Doelling[1*]

1. Institut Pasteur, Université Paris Cité, Inserm UA06, Institut de l'Audition, F-75012 Paris, France
2. Aix-Marseille University, Inserm, Institut de Neurosciences des Systèmes (INS), Marseille, France
3. My Medical Assistant SAS, Reims, France
4. Institut Arthur Vernes, ENT Department, F-75006 Paris, France

***Corresponding Authors :**
LHA : luc.arnal@pasteur.fr
KBD : keith.doelling@pasteur.fr
These authors jointly supervised this work

Classifications:
Biological Sciences – Psychological and Cognitive Sciences

Keywords: Audiology, Pure Tone Audiogram, Sequential Predictions, Bayesian Active Learning

Abstract
Human hearing is highly sensitive and allows us to detect acoustic events at low levels and at great distances. However, sensitivity is not only a function of the integrity of cochlear mechanisms, but also constrained by central processes such as attention and expectation. While the effects of distraction and attentional orienting are generally acknowledged, the extent to which probabilistic expectations influence sensitivity is not clear. Classical audiological assessment, commonly used to assess hearing sensitivity, does not distinguish between bottom-up sensitivity and top-down gain/filtering. In this study, we aim to decipher the influence of various types of expectations on hearing sensitivity and how this information can be used to improve the assessment of sensitivity to sound. Our results raise important critiques regarding common practices in the assessment of sound sensitivity, both in fundamental research and in audiological clinical assessment.

Significance Statement
The principal clinical measure of ear function is the pure-tone audiogram in which a clinician will test for the softest tones one can detect across the frequency range. Clinicians have long known listener strategy can influence these measures whereby patients can guess when tones may occur based on regularities in the tone presentation structure. Our results demonstrate that this predictability and other forms influence not only listener strategy but also the sensitivity of the peripheral auditory system itself. This finding that prediction not only biases low-level perception but also enhances it has wide-ranging consequences for the fields of audiology and cognitive neuroscience, and furthermore suggests that prediction ability should be tracked as well as ear sensitivity as patients age.

Introduction

The sense of hearing allows us to detect and interpret even very weak sounds in our environment. Typically, sensitivity is primarily attributed to the intricate mechanics of the middle and inner ear, and the synaptic transduction of signals to the auditory nerve (Hudspeth, 2014; Marin et al., 2022; Pickles, 1998). The efficacy of these mechanics is largely measured using pure-tone detection (PTD) thresholds, task paradigms designed to gauge the softest intensities at which a person can detect pure tones, which are standard practice in clinical audiology (Brant & Fozard, 1990).

In recent decades, neuroscience has increasingly embraced the idea that sensory processing is more than just a result of bottom-up signal transduction. According to predictive processing theories (Friston, 2005), perception is an active process heavily relying on our capacity to anticipate upcoming events based on our sensory experience. These predictions are proposed to be organized in a hierarchical manner across various levels of abstraction, such as acoustic, phonemic, or semantic levels, for instance. Top-down signals proactively filter and modulate the gain of expected stimulus inputs (Hesselmann et al., 2010). As such, even without intending to, the human brain extracts statistical regularities to exploit predictable structures in a task (Saffran et al., 1999). Does the effect of prediction extend to even the most fundamental and low-level auditory function, pure-tone sound detection?

Previous work has shown that hearing sensitivity is contingent on more than the mere transduction of ascending signals at the cochlear level but is also affected by the allocation of cognitive constructs such as attentional resources towards upcoming events depending on their behavioural relevance. The probe-signal task (Greenberg & Larkin, 1968), for example,

demonstrated that detection of a tone is enhanced when it is preceded by a probe tone around the same frequency. A wealth of studies in psychoacoustical experimentation have demonstrated this improvement in performance based on this sort of selective attention *in noise* (Borra et al., 2013; Scharf et al., 1987; Tanner & Norman, 1954). These studies operate under the assumption that the effects of top-down modulation of signal detection would be the same in quiet at threshold and that the added noise merely externalizes and enhances the neural noise of the system.

On the other hand, the audiological field considers the pure-tone audiogram (signal detection in quiet) to diagnose only the bottom-up sensory pathway (Musiek et al., 2017), even though current audiometric protocols contain considerable predictive structure. Do the findings of psychoacoustics in noisy settings apply to signal detection in quiet? Or does expectation only play a role when separating auditory signal from noise?

While previous work has studied the role of attention in quiet, this has largely been in terms of attention to hearing vs visual input. Rather than manipulating the expected information of attended stimuli, these studies consider attention as a means of removing distraction from one or another sensory domain. Lukas (1980), for example, demonstrated that brainstem responses from auditory nerve to colliculus were filtered by sustained concentrated attention to the visual domain, suggesting top-down control to avoid distraction. Later evidence further suggested that attention has cochlear effects, demonstrated by changes in oto-acoustic emissions during attention to one ear vs the other (Giard et al., 1994) and by changes in rhythmic modulation of the auditory nerve fiber during attention to the auditory vs visual domain (Gehmacher et al., 2022). These findings provide a mechanism for top-down modulation of cochlear sensitivity via the olivocochlear bundle. Given that a decidedly cortical construct can have direct effect on peripheral function, it is reasonable to explore whether other more stimulus-specific cognitive functions may also influence cortical processing.

Of particular importance in the case of audition is the perception of not only the content of a sound but also its timing. For example, speech contains prosodical cues such as the speeding up or slowing down of syllabic rate which provides further nonverbal information to the listener. As such, predictions about an upcoming sensory event can not only regard its content but also its moment in time. This delineation between content-based ("what") and time-based ("when") predictions has been maintained in the field suggesting that the two rely on different top-down mechanisms (Arnal & Giraud, 2012). In the case of timing, a growing body of evidence supports the coupling of neural oscillations in different timescales to statistical regularities in the sensory input as a mechanism for predictive timing, particularly in the case of auditory stimuli (Arnal et al., 2015; Doelling et al., 2019; Nobre et al., 2007). Meanwhile, content predictions are sent as inhibitory top-down signals (Bastos et al., 2012; Friston, 2010; Heilbron & Chait, 2017), reducing the processing of predicted signals so that only what is unexpected is given greater neural resources.

These components of prediction, while developed over a large body of research across several decades in the fields of psychology and neuroscience, have not been heavily considered in the audiological domain. In the clinic, one baseline measurement for auditory health focuses on pure tone detection, identifying the softest presentation level at which a participant can detect a pure tone. However, such measurements overlook the predictive nature of perception, often presenting the same tone repeatedly and regularly, thereby allowing for the participant to predict when and what a tone is before it arrives. Recent protocols in audiological paradigms have begun to take this into account, instructing clinicians

to randomize the timing of tones in pure tone detection tasks to avoid guessing (British Society of Audiology, 2018). Still, content predictions (i.e., pitch) are ignored and the role such predictions play and how they interact with temporal predictions in threshold evaluation is unknown. While much work on prediction has focused on signal recovery in noisy environments, very little work has been done at auditory detection threshold. Our goal is to assess how much sequential structures contribute to thresholds evaluation under current clinical techniques and to what extent the predictions in either time or content can affect sensory outcomes.

In this study, we take advantage of new advances in pure tone detection paradigms (Schlittenlacher et al., 2018) by employing automated detection threshold techniques using Bayesian Machine Learning to infer audiograms based on a more flexible paradigm, allowing for a fully randomized structure in both the timing and frequency. We use this advance to compare audiograms within subject by asking participants to complete audiograms with varying levels of structure: 1) Randomized Paradigm, 2) Predictive Paradigms, where prediction in time and frequency are tightly manipulated, and 3) 3-AFC Staircase Paradigm, a traditional clinical paradigm carrying much predictive information.

We expect that predictive structure will facilitate sensory detection at thresholds. Our aim is to quantify this gain by measuring both the decibel shift in thresholds as a result of a change in structure, and also by comparing the number of tones detected against the probabilistic fits of the unpredictable stimuli, determining the decibel shift in detectability of each tone. This experiment will provide key information regarding the role of prediction at detection thresholds in the normal hearing listener and will therefore provide the foundation for future work to show how this relationship may be altered in hearing-impaired populations.

Methods

Participants

31 participants aged 18 to 45 with self-reported normal hearing were recruited using the RISC (Relai d'Information sur les Sciences de la Cognition) volunteers' database (Risc., https://www.risc.cnrs.fr/.). Normal hearing was confirmed (Pure Tone Average < 20 dB HL) in the first audiometric test (described below) and no participants were removed on this basis. Three participants were excluded from all the analyses, for either not following the instructions, or due to a data collection error. We present the results from the remaining 28 participants (12 men, 16 women) with an average age of 26.25 years-old (sd: 5.73). Informed consent was obtained prior to testing and participation in this study was compensated at a rate of 10€/h. One session lasted about 1h30 in total. We assessed self-reported musical ability by administering the Goldsmiths Musical Sophistication Index (Gold-MSI). This study was approved by the Comité de Protection des Personnes Tours OUEST 1 on 10/09/2020 (project identification number 2020T317 RIPH3 HPS).

Stimuli and Material

All experiments were created using the open-source python library Psychopy 2022.1.1 (Peirce et al., 2019). Stimuli consisted of 200 ms pure tones with frequencies spanning the 125-8000 Hz range. Stimuli were presented *binaurally* in a sound attenuated booth, through Etymotic Research ER-2 insert earphones fitted with 13mm Echodia plugs and connected to a Solid State Logic SSL 2 soundcard. The 200ms pure tones were apodized with 5ms hanning windows and hearing thresholds were measured for tones calibrated to ISO 389-2, with reference to an IEC-711 occluded-ear simulator. Psychopy's sound module interprets

loudness as a number comprised between 0 and 1, therefore the calibration was performed for an arbitrary Psychopy volume of 0.005 before applying the correction to dB HL. The presentation level was linearly interpolated for frequencies not included in the correction table.

Experimental setup

Participants completed 4 audiometrical experiments in total hereinafter referred to as the Randomized paradigm, the 3-alternative forced choice (3-AFC) adaptive staircase paradigm, the Sweeping Cluster paradigm, and the Sweeping Continuous paradigm. Sounds were presented binaurally which likely accounts for the significantly lower thresholds overall than typically found in audiometry which would usually examines one ear at a time. The 3-AFC adaptive staircase (Figure 1A) was designed to reproduce the current gold-standard in audiometry procedures and allow us to compare its threshold estimation with other assessments made using tones of varying predictability. The Randomized task (Figure 1B) measures thresholds for pure tones that are as unpredictable in timing and frequency as possible within the limitations of this study, while the two Sweeping designs recorded 4 audiograms each, in different conditions of predictability in time and frequency. Every participant performed the Randomized task first, to confirm inclusion in the study (< 20 dB HL hearing loss) and to let us choose the sound presentation levels in both Sweeping paradigms based on the Randomized audiogram estimation (see below). The order of the remaining 3 experiments was shuffled across participants.

To estimate thresholds, all tasks except for the 3-AFC adaptive staircase rely on a state-of-the-art, automated pure-tone audiometry Application Programming Interface (API) by My Medical Assistant SAS (iAudiogram, www.iaudiogram.com) validated against customary adaptive staircase procedures and based on the automated machine learning described previously (Schlittenlacher et al., 2018). The API's algorithm starts by collecting initialization data by testing tones spanning the frequency range (1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz, 8000 Hz, 750 Hz, 500 Hz, 250 Hz, 125 Hz) to provide an initial coarse estimation of the audiogram. After each tone presentation, the algorithm requires a positive or negative answer to select the next stimulus. During the initialization phase, the frequency of the tones is predictable, as one frequency is tested several times in a row at decreasing hearing levels, until a negative answer is recorded for that frequency. To avoid repeating this phase in the two Sweeping paradigms where we aim to control the predictability of oncoming tones, we reused the initialization data recorded during the Randomized task to start all the audiograms measured in the Sweeping paradigms.

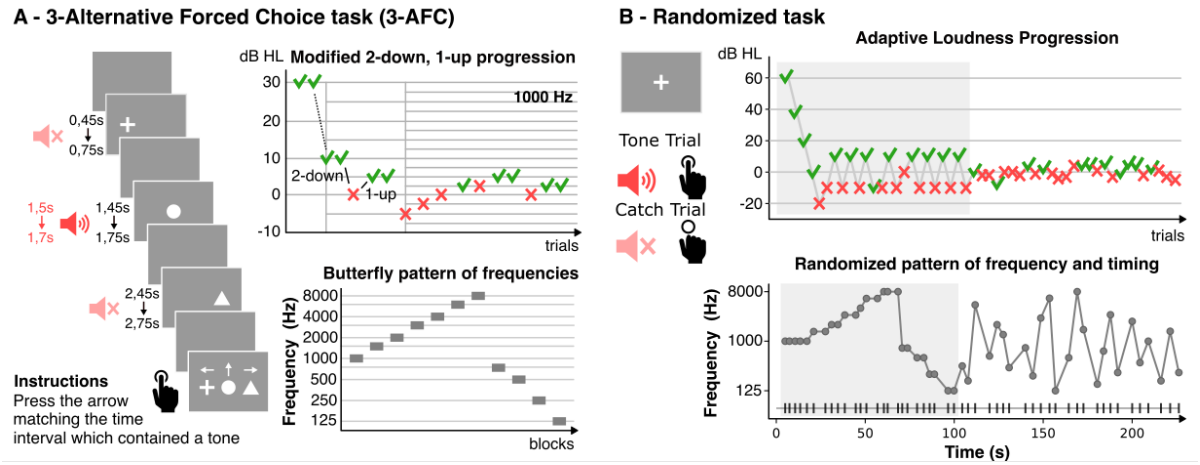

Figure 1 – Experimental paradigms

A – 3-AFC task. *Participants were asked to identify which of 3 possible time intervals contained a tone. A 2-down, 1-up adaptive staircase starting at 30 dB HL determined the level of the presentation of the tones. The initial 10 dB HL step was halved after the first step down and halved again after 3 reversals of the staircase. The next frequency started after 6 reversals. In total, 11 frequencies were tested successively.* ***B – Randomized task.*** *A fixation cross was displayed on screen while tones chosen by an automated pure-tone audiometry API were presented with randomized ISI (1-3 sec). Instructions were to press a key as soon as a tone was detected. For this task and both Sweeping paradigms, participants were advised that only key presses within a 1-second window following the tone would be considered positive detections.*

Three Alternative Forced Choice (3-AFC) paradigm

At each trial, participants were asked to identify one of three possible time intervals during which a tone had been played. A sequence of 3 shapes flashed on screen (a cross, a disk and a triangle respectively on the left, in the middle and on the right) to indicate the three intervals in time (Fig. 1A, left). Each shape was displayed during 300 ms and the next one appeared after a delay of 700 ms. Participants had to press the keyboard arrow corresponding to the correct interval, randomly chosen on a trial-by-trial basis. The experiment started with five practice trials to familiarize participants with the task. We tested 11 frequencies ordered in a butterfly pattern spanning the 125-8000 Hz range (1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz, 8000 Hz, 750 Hz, 500 Hz, 250 Hz, 125 Hz) (Fig. 1A, bottom right). The first tone of each frequency was presented at 30 dB HL. The level of presentation of the next tone was determined based on a 2 down – 1 up adaptive staircase procedure (Fig. 1B, top right). The initial step of 10 dB HL was reduced to 5 dB HL after the first decrease in level. We halved the step again after 3 reversals of the staircase, and stopped after 6 reversals. Participants had the opportunity to take a short break before starting the next frequency's staircase. Pure tone thresholds were estimated as the average level across the last 4 reversals.

Randomized paradigm

Participants were asked to detect the presence of a tone at low decibel levels, indicating their detection by pressing a key whenever they were able to hear the tone (Fig. 1B, top). They were informed that detections would only be considered correct if they occurred within a 1-second window following the tone. A fixation cross was displayed on the screen during the whole task. Participants were familiarized with the paradigm using three practice trials at 40 dB HL. After the initialization phase of the API, the following sound presentation levels and frequencies were chosen to maximally reduce the uncertainty of the threshold estimation,

based on Bayesian inference (Fig. 1B, bottom). The estimation is based on recent work designed to automate pure tone detection threshold (Schlittenlacher et al., 2018). It uses a Gaussian Process with prior knowledge about the relationship between points in the frequency-level space, to assign a probability that a tone of any level and frequency will be heard. The prior knowledge is encoded through kernels: one with a covariation matrix using a squared exponential in frequency space, denoting that neighboring frequencies are correlated, and another as a linear function in intensity space, denoting that increased loudness should increase the probability of detection. The next tone is chosen as the point leading to the greatest decrease in the uncertainty of these probabilities. The timing in between each tone, the Inter-Stimulus Interval (ISI), was randomly chosen for each tone to be between 1 and 3 seconds to reduce the participants' ability to predict the timing of upcoming tones. In addition, the call to the API takes progressively longer over the course of the experiment, meaning the ISI gets longer over the course of the experiment which could lead up to a 6 second delay between tones by the end of the paradigm. Each tone had a 20% chance of being replaced with silence to be analyzed as catch trials. The true average percentage of catch trials across participants amounted to 18.81% ± 1.57% of trials. The paradigm completes when the Bayesian estimator reaches a confidence interval of 6 dB. With this constraint, the complete paradigm comprised an average of 60.9 trials (min: 51, max: 101, sd: 17.9), including an initialization phase of 25.1 trials (min: 22, max: 28, sd: 1.4).

Predictive Sweeping paradigms

For both Sweeping paradigms, tones were organized in 'sweeps' whose structures defined four conditions of predictability: predictable timing (T), predictable frequency (F), predictable frequency and timing (FT), or random frequency and timing (R) (Fig. 2AB). The main difference between these two tasks is the timescales used to separate successive tones. In the Continuous sweeping task, we recorded whether participants detected each tone within a sweep, to emulate clinical procedures where an answer is given for each tone. This required using large ISIs to allow time for participants to answer. Each tone thus consisted in a trial and each sweep represented one block. In the Cluster task, meanwhile, each trial tested the detection of a target tone preceded by a rapid cluster of 4 cue tones, in a fashion more relevant to prediction studies that tend to use shorter ISIs. In this task, one sweep represented one trial.

The four conditions were obtained by building sweeps using the same structure in both paradigms: we made some of the sweeps predictable by regularly spacing their tones in time and/or frequency. For predictable timing, the ISI of a sweep was kept constant and was randomly selected from a paradigm-dependent range (Continuous: 1-3s, Cluster: 400-800ms). For frequency predictability, the distance between two successive tones was picked at random from a list of three musical intervals: (whole tone, major third or perfect fifth) and was the same for all tones in a sweep. To create unpredictable sweeps, the intervals between successive tones were randomized (ISIs for time and musical intervals for frequency). We subsequently shuffled the order of the tones for sweeps with unpredictable frequency (T, R).

We measured a separate audiogram in each of the four conditions of predictability and the order of presentation of the conditions was randomized for each sweep. Participants were made aware of the predictability of each sweep through a visual cue. The word 'PITCH' would flash on screen for 1 second before sweeps with predictable frequency, and the word 'RHYTHM' flashed the same way for sweeps with predictable timing. Both words flashed

simultaneously before FT sweeps, while fully random sweeps were preceded by a flashing cross.

Sweeping Continuous paradigm

Instructions were the same as those for the Randomized task: participants had to press the spacebar less than 1 second after the onset of tones they managed to detect. Stimuli were presented in sweeps of 8 to 15 tones organized in time and pitch, or not (Fig. 1C). After flashing the next sweep's condition for 1 second, a fixation cross was displayed until the end of the sweep. The API chose one tone per sweep and each sweep was built around its frequency so that the API-picked tone was located near the middle of the sweep in conditions of predictable frequency. Each sweep was initially designed using 15 tones with the API tone in 8th position. Sweeps with tones whose frequency fell outside of the 125-8000 Hz range were truncated, and subsequently redesigned by using a smaller interval if the number of tones left in the sweep after truncating was lower than 8. The presentation level of tones not chosen by the API were determined randomly between the threshold estimated in the Randomized paradigm (always completed first) for the tone's frequency and the level of the API-chosen tone. ISIs were randomly chosen from a range of 1 to 3 seconds. Each sweep contained one catch trial which replaced a randomly chosen tone in the sweep from the third position onward (9.20% ± 0.99% of tones across participants). For the F, T and FT conditions, all analyses were conducted after removing data from the first 2 tones of each sweep since they cannot be predicted using previous tones.

Sweeping Cluster paradigm

Participants were instructed to report whether they could hear a tone while a disk was displayed on the screen. Before the disk was displayed, a cluster of four pure tones presented at the same time as a fixation cross served as cues for the target tone. The delay following each of the four cues was drawn from a uniform distribution between 400 and 800 ms in trials with random timing (F, R), while it was drawn per trial (per sweep) from the same range in predictable-timing conditions (T, FT). The disk appeared as soon as the last cue ended and remained displayed during 1.2s. Only key presses collected one second or less after the start of the target tone were considered positive detections. Participants were familiarized with the experiment in an initial training phase consisting of five mock trials with feedback.

Because the call to the API could take up to several seconds, waiting for its response for every trial would have significantly slowed down the experiment. Instead, the API determined the frequency and level of every other target tone presented in each of the 4 conditions. The target frequency of the remaining 50% of trials was randomly picked from a log-uniform distribution ranging 125 to 8000 Hz, while the hearing level was interpolated from the audiogram estimation made in the Randomized paradigm. The level of the cues was also interpolated from the Randomized audiogram and raised by 6 dB to increase the probability of their detection. Each target tone had a 20% chance of being replaced with silence to be analyzed as a catch trial (19.53% ± 6.47% of trials across participants).

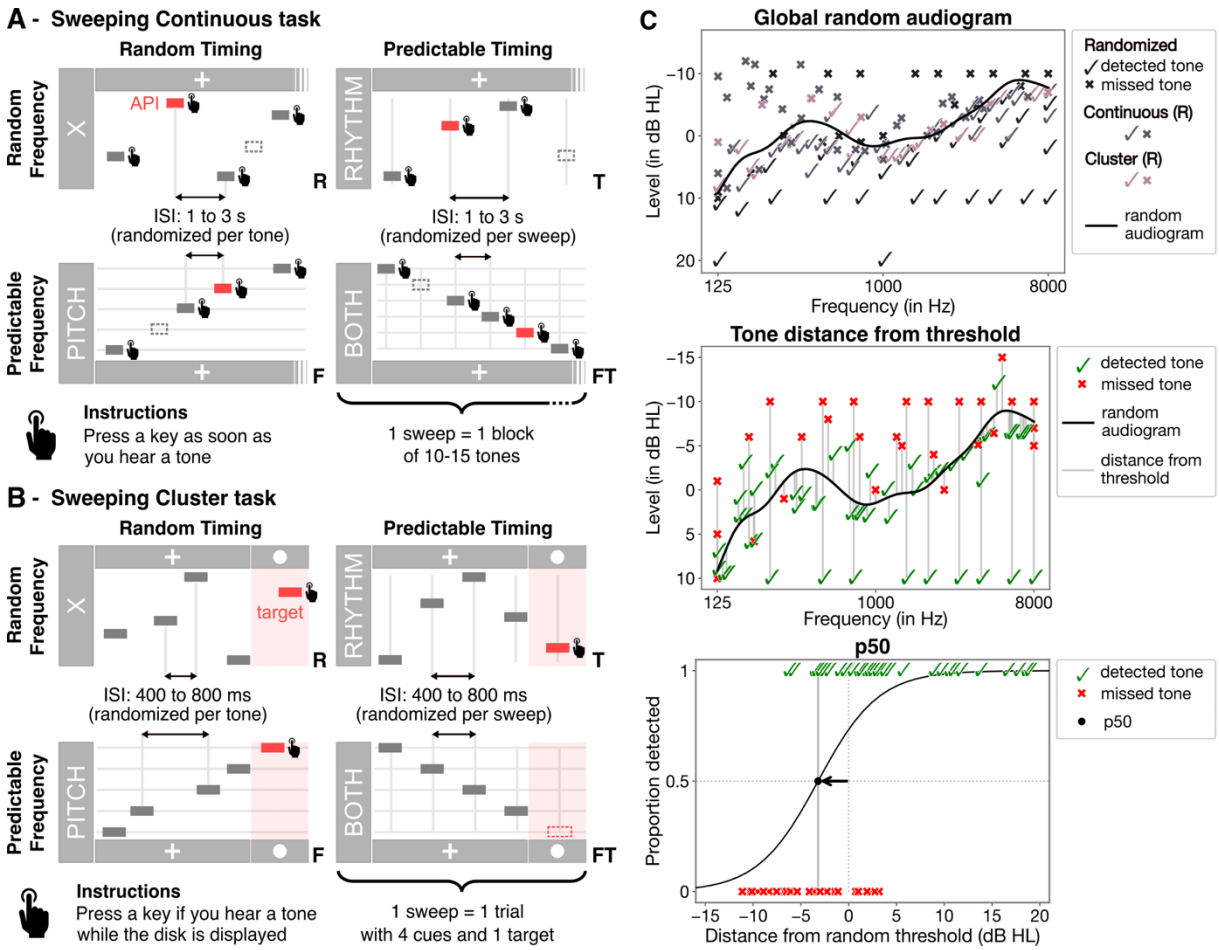

Figure 2 – Sweeping tasks design.

Four conditions of predictability were established for both paradigms: predictable timing (T), predictable frequency (F), predictable frequency and timing (FT), or random frequency and timing (R). Predictability in each modality was achieved by designing sweeps with constant intervals, i.e., by regularly spacing tones in time and/or frequency. The ISI was chosen from a paradigm-dependent range, while the frequency interval was selected from a list of three musical intervals (whole tone, major third, perfect fifth). In conditions of unpredictability, we randomized the intervals (ISIs and/or musical intervals) for each tone in a sweep. We further randomized frequency by shuffling the order of tones in a sweep. ***A - Sweeping Continuous task.*** *Sweeps consisting of 8 to 15 tones were presented while a fixation cross was displayed on screen. The instructions were to press a key as soon as a tone was detected. The API chose one tone near the middle of each sweep and each sweep contained one catch trial.* ***B – Sweeping Cluster task.*** *A cluster of 4 cue tones served to indicate the timing and frequency of a 5th target tone, or not. A fixation cross was displayed on screen until the end of the last cue. A disk then replaced the cross to visually indicate the interval during which the target tone could be presented. Participants were to press a key if they detected a tone while a disk was displayed on screen.* ***C – Method for estimating p50 values (example data).*** *Top panel: A 'global random' audiogram (solid black line) is computed for each participant from aggregating the answers of the Randomized task and condition R of the Continuous and Cluster tasks. Middle panel: for each tested tone or a sweeping condition (R, F, T, or FT; checkmarks: positive detections, crosses: missed detections), the distance from the 'global random' threshold at the tested tone's frequency within each condition is estimated in dB HL (vertical grey lines). Bottom panel: The tones tested in each condition are ordered by their distance from the 'global' threshold and their proportion of detection is estimated by fitting a sigmoid to the participant's answers. We defined p50 as the distance from the 'global random' threshold corresponding to the 50% mark of detection.*

Analysis

Comparison of 3-AFC and Randomized Paradigm

The 3-AFC and Randomized paradigms use different algorithms to calculate the final threshold: Adaptive Staircase and Bayesian Active Learning, respectively. These methods differ in the continuity of the outputted thresholds (the staircase is necessarily discrete, whereas the Bayesian algorithm is continuous), but otherwise are readily comparable. As such, to account for this difference we sample the threshold values of the Randomized paradigm at the frequencies tested in the 3-AFC. We then average the values across frequency for each participant to get a mean detection threshold comparable with the 3-AFC in two ways: first, averaging across all 11 frequencies of the 3-AFC, and second, averaging over the 4 critical frequencies commonly tested in the clinic, 500, 1000, 2000 and 4000 Hz. The difference in thresholds from each metric are compared using a T-test.

While the 3-AFC is designed to infer a threshold at 70.7% correct, this percentage corresponds to a 56.5% chance of tone detection (accounting for guessing at 33% when tones go undetected). We consider this percentage to be negligibly different from the 50% chance inferred by the Randomized. Furthermore, correcting for the 6.5% difference, if possible, would only enhance the size of the effect between the two conditions as we expect the more predictable task (3-AFC) to result in lower thresholds.

P50 Threshold Distance

After the above initial comparison, it is important to control for the difference in methods used to assess thresholds and compare all paradigms under the same footing. As such when considering all paradigms together (Randomized, Sweeping Tasks and 3-AFC), we first calculate a global threshold using Bayesian Active Learning as in the Randomized paradigm incorporating tones from all random conditions (Randomized paradigm and Random conditions from Sweeping tasks - Figure 2C, top panel). This threshold reflects an audiogram from unpredictable stimuli without dependence on the paradigm protocol. Then for each condition, we consider the distance of each tone presented in terms of its intensity from the global threshold (Figure 2C, middle panel). We use this distance as an independent variable in a logistic function to predict whether each tone will be detected or not (Figure 2C, bottom-panel). If the point of equivalence of the logistic function is equal to 0, then there is no difference in threshold for this condition relative to global threshold. However, if there is a significant difference in either direction, we can assume the threshold has shifted by this amount. This method allows us to treat each condition in the same manner comparing its outcome as a relative distance from all conditions with unpredictable stimuli.

Statistics

All statistical analyses were conducted using Python libraries, including SciPy (Virtanen et al., 2020), Statsmodels (Seabold & Perktold, 2010), Pingouin (Vallat, 2018) and Scikit-learn (Pedregosa et al., 2011). Audiograms were compared across frequencies and paradigms using two-factor repeated-measures ANOVA, with Greenhouse-Geisser correction applied where the sphericity assumption was violated. Paired t-tests were performed to compare mean thresholds across frequencies and paradigms, using the Benjamini Hochberg method to correct for False Discovery Rate (FDR) (Benjamini & Hochberg, 1995).

To ensure a consistent comparison across frequencies, paradigms and conditions, all subsequent analyses made use of the calculated p50 values, which serve as a normalized metric representing the relative deviation of the detection threshold from the reference threshold. Owing to variability in the number of calculable p50 values across tasks, separate

one-way ANOVAs were implemented for the Cluster and Continuous tasks to assess differences within paradigms based on predictability levels. This approach ensured that the statistical comparisons were adequately tailored to the available data for each task. Post-hoc pairwise comparisons with Benjamini Hochberg correction were performed to determine significant differences between different conditions.

To investigate potential decisional bias, we incorporated catch trials into all paradigms except the 3-AFC task and computed false alarm rates from these trials. We conducted a repeated-measures ANOVA, treating each testing condition – be it the single condition in the Randomized task or any of the four distinct conditions in the Cluster and Continuous tasks – as a separate group. To identify significant variations in false alarm rates across these conditions, we performed post-hoc comparisons using False Discovery Rate correction (Benjamini & Hochberg, 1995). We performed Pearson's correlation analysis to investigate relationships between false alarm rates and p50 threshold differences and between participants' characteristic features (age and musicianship) and their ability to benefit from predictive information. In all tests, the significance level was set at $p < 0.05$. We also used unsupervised hierarchical agglomerative clustering to evaluate the patterns of participant performance across paradigms and predictability conditions. Clustering algorithm used performed using SciPy's hierarchical clustering module with linkage using an average method and cosine distance to avoid clustering based on overall performance. All data are presented as means ± standard error of the mean.

Results

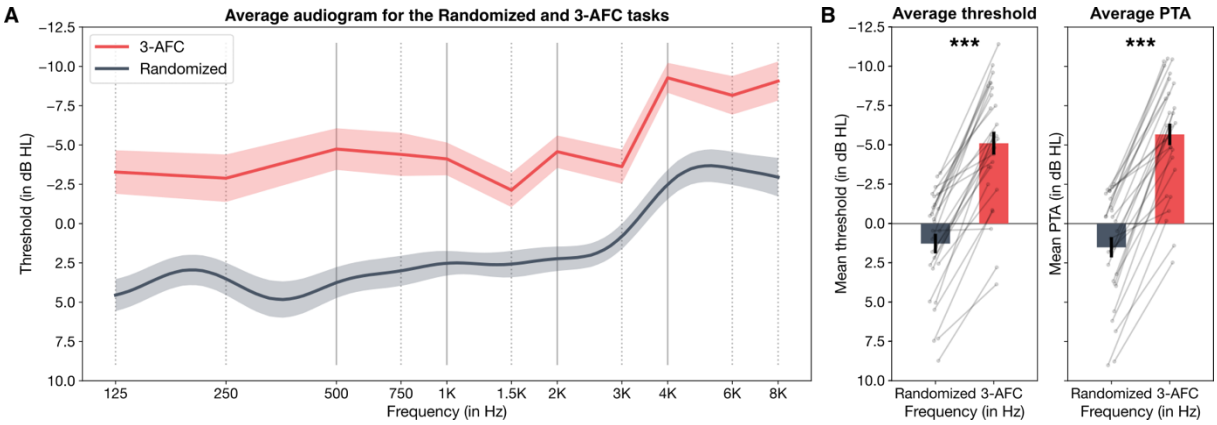

Figure 3: Thresholds are lower in the highly structured 3-AFC task than in the ***R***andom***ized*** task. *A – Average audiograms measured in the Randomized task (in dark grey) and in the 3-AFC task (in red). Shaded areas represent the standard error to the mean. B – left panel: Average threshold for both tasks, calculated as the mean for the 11 frequencies tested in the 3-AFC paradigm (all vertical lines in panel A). Individual threshold estimates are plotted as dark circles. right panel: Pure Tone Average (PTA) for both tasks, calculated as the mean threshold at 500, 1000, 2000 and 4000 Hz (solid vertical lines in panel A). Individual PTAs are plotted as dark circles. *** indicate significance at p < 0.001*

Audiometric protocols adjust thresholds by 7 dB

First, we compared audiograms from the Randomized version (estimated using Bayesian Machine Learning) and the 3-AFC tasks (two-up, one-down adaptive staircase), which represent our most extreme cases of random and predictable while also being the most standard in terms of previous testing. Figure 3A shows the average audiogram of all participants for the two paradigms across the tested frequency range from 125 – 8000 Hz. The two paradigms were compared at the 11 frequencies tested in 3-AFC. The result shows a significant difference in intensity across frequencies in the two paradigms (two-factor

repeated-measures ANOVA, paradigm effect: $F(1, 27) = 271.50$, $p < 0.001$; frequency effect: $F(10, 270) = 11.14$, $p < 0.001$; paradigm x frequency interaction: $F(10, 270) = 2.44$, $p = 0.009$). To estimate the size of this effect, we then extracted, from the Randomized threshold, the values only at frequencies that were tested in the 3AFC and compared the mean values. Figure 3B shows that this comparison yielded a difference of 6.4 dB HL ($t(27) = -13.24$, $p < 0.001$). We repeated this analysis using only the thresholds measured for the frequencies 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz to compute the Pure Tone Average (PTA), a commonly used metric in audiology. For these selected frequencies, the difference in thresholds between the Randomized method and the 3-AFC was 7.19 dB HL ($t(27) = -14.41$, $p = <0.001$).

Predictive structures induce increased sensitivity

We then tested for what role predictive sequential structure plays in this difference using the Continuous and Cluster sweep tasks described in the Methods section. In Figure 4A, we first compared the most random version of each paradigm to assess how changes in protocol for each compare across the tasks. A one-way ANOVA shows a significant difference across paradigms ($F(3, 81) = 191.91$, $p < 0.001$). Pairwise comparisons reveal significant differences between all paradigms (3AFC vs Randomized: $p < 0.001$; 3-AFC vs Cluster: $p < 0.001$; 3-AFC vs Continuous: $p < 0.001$; Randomized vs Cluster: $p < 0.001$; Randomized vs Continuous: $p = 0.001$; Cluster vs Continuous: $p = 0.006$). The effect of sweeping paradigms compared to Randomized reveal unsurprisingly that the change in protocols, even when cues are reduced as much as possible, provides added information to the presence of the tone compared to the fully randomized comparisons.

Next, we compared the predictable conditions within the Continuous and Cluster paradigms (Figure 4A, right) to see what role prediction plays in auditory sensitivity outside of the protocol differences analyzed above. To that end, we conducted a separate one-way ANOVA for each paradigm, with predictability (Random (R), Timing (T), Frequency (F), or Frequency and Timing (FT)) as the independent variables. For both paradigms, the ANOVA revealed a significant main effect of predictability (Continuous: $F(3, 81) = 6.92$, $p < 0.001$; Cluster: $F(3, 81) = 14.62$, $p < 0.001$), confirming that even in a tightly controlled paradigm, predictive structure enhances pure-tone detection sensitivity. Post-hoc paired t-tests further showed that all predictable conditions improved performance compared to the most random condition, in both tasks (Continuous - R vs. FT: $p = 0.008$; R vs. F: $p < 0.001$; R vs. T: $p = 0.028$; Cluster - R vs. FT: $p < 0.001$; R vs. F: $p < 0.001$; R vs. T: $p < 0.001$), showing a difference of about 2 dB with the most random condition. Thresholds also improved significantly in the fully predictable (FT) condition compared to the condition with predictable timing only (Continuous T vs. FT: $p = 0.036$; Cluster T vs. FT: $p = 0.006$).

Distinct predictive sources across timescales

We next investigated how performance correlated between task pairs. Our aim was to test whether performance was grouped by the category of predictability (frequency vs timing) or by the paradigmatic setup (continuous vs cluster). The correlation matrix, shown in Figure 4B, reveals that the structure of the task most drives similar performance rather than predictability in a particular dimension like time or frequency. This suggests that the two different time scales lead to different mechanisms through which predictive structure can support the analysis. To test this further, we included participant performance into an

unsupervised clustering algorithm (4B, right) to assess how the patterns of participant performance could help us to infer similar mechanisms used across paradigms. The clustering algorithm clearly grouped performance by paradigm rather than by the type of predictability in the condition.

Predictive gain mostly driven by sensitivity, not decisional bias

Another important question is what role decisional bias plays in this increase in performance. As the sweeping paradigms involved a detection task, it could be that participants claimed detection more even when they didn't hear anything. To test for this, we included catch trials in the experiment during which no tone occurred. Figure 4C reveals the results of the catch trial analysis. We conducted a repeated measures ANOVA with the false alarm rate as the dependent variable, treating each condition as a separate group to allow for comparisons among all conditions. The analysis revealed a significant effect of group ($F(8, 216) = 5.78$, $p = <0.001$). Post-hoc paired t-tests using FDR correction indicated that false alarm rates in the Randomized paradigm were significantly lower than in all conditions of the Continuous and Cluster tasks (Randomized vs Cluster/F: $p = 0.004$; Randomized vs Cluster/FT: $p = 0.002$; Randomized vs Cluster/R: $p < 0.001$; Randomized vs Cluster/T: $p < 0.001$; Randomized vs Continuous/F: $p = 0.047$; Randomized vs Continuous/FT: $T(27) = -5.13$, $p < 0.001$; Randomized vs Continuous/R: $p = 0.018$; Randomized vs Continuous/T: $p < 0.001$). Within the Continuous paradigm, false alarm rates were higher in both conditions of the task with predictable timing compared to the random condition (FT: $p = 0.022$; T: $p = 0.017$), and they were lower when only the frequency was predictable than when only the time was predictable (F vs. T: $p = 0.019$). While temporal predictability in the Continuous paradigm does appear to be modulated by decisional bias, it is unlikely to explain the entirety of the result. Predictability in frequency does not appear to affect false alarms (F vs R: $p = 0.965$) while still improving detection thresholds. At the same time the cluster paradigm shows a nonsignificant reduction in false alarms as predictability increases ($F(3, 69) = 0.668$, $p = 0.575$), excluding decisional bias as a possible explanation of the result. To further test this hypothesis, we compared false alarm rates of individual participants against their threshold differences. While the Continuous paradigm shows some contribution of decisional bias to the overall performance with the FT condition indicating a significant correlation ($R^2 = 0.456$, $p < 0.001$), there is no such significant correlation in the Cluster paradigm (R: $R^2 = 0.116$, $p = 0.076$; T: $R^2 = 0.051$, $p = 0.247$; F: $R^2 = 0.026$, $p = 0.408$; FT: $R^2 = 0.052$, $p = 0.244$).

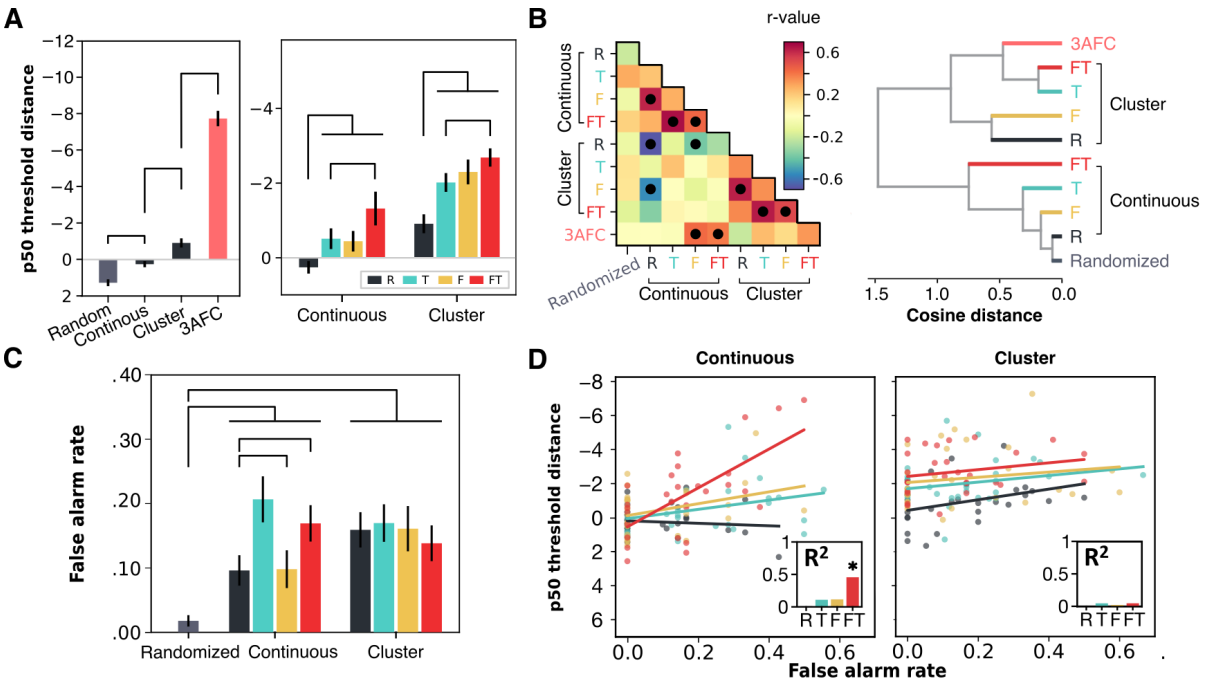

Figure 4: Effect of prediction on detection thresholds*. A – threshold distance effects due to protocol (left) and to predictability (right). Protocol distance is assessed by comparing the random conditions within each protocol. Significant post-hoc comparisons indicated with black lines. B – protocol clustering based on performance correlations across subjects. Left, Correlation matrix of subject performance by protocol conditions. Filled circles indicate significant correlations. Right, Hierarchical clustering of protocols by cosine distance of subject performance in each cluster. C – False alarm rates in each condition. Significant post-hoc pairs indicated with black lines. D – correlation between threshold distance and false alarm rates for continuous (left) and cluster (right) paradigms. Scatter plots indicate single subject plots, lines indicate linear fit between the two parameters, insets reveal R^2 of the correlation between false alarms and threshold distance. * indicate significant relationship.*

Individual differences not explained by musical experience

We then sought to explain the variability in participants' ability to extract predictive information in terms of other characteristic features that we had tracked through surveys: musical expertise and age (Fig. 5). Neither feature had a particularly strong effect on participant performance. In the continuous paradigm, Age had no effect on threshold distance (R: $R^2 = 0.001$, $p = 0.878$; T : $R^2 = 0.001$, $p = 0.872$; F : $R^2 = 0.008$, $p = 0.658$; FT: $R^2 = 0.000$, $p = 0.914$). In the cluster paradigm, Age had a worsening effect on Random and Time conditions but no effect on Frequency and Both conditions (R: $R^2 = 0.154$, $p = 0.039$; T : $R^2 = 0.226$, $p = 0.011$; F : $R^2 = 0.007$, $p = 0.68$; FT: $R^2 = 0.011$, $p = 0.601$). Surprisingly, Musicianship (as indexed by the General Sophistication score of the GMSI) had no effect on threshold distance in any condition (Cluster: R - $R^2 = 0.017$, $p = 0.508$; T - $R^2 = 0.001$, $p = 0.861$; F - $R^2 = 0.061$, $p = 0.204$; FT - $R^2 = 0.008$, $p = 0.653$; Continuous: R - $R^2 = 0.026$, $p = 0.416$; T - $R^2 = 0.002$, $p = 0.827$; F - $R^2 = 0.023$, $p = 0.438$; FT - $R^2 = 0.003$, $p = 0.776$).

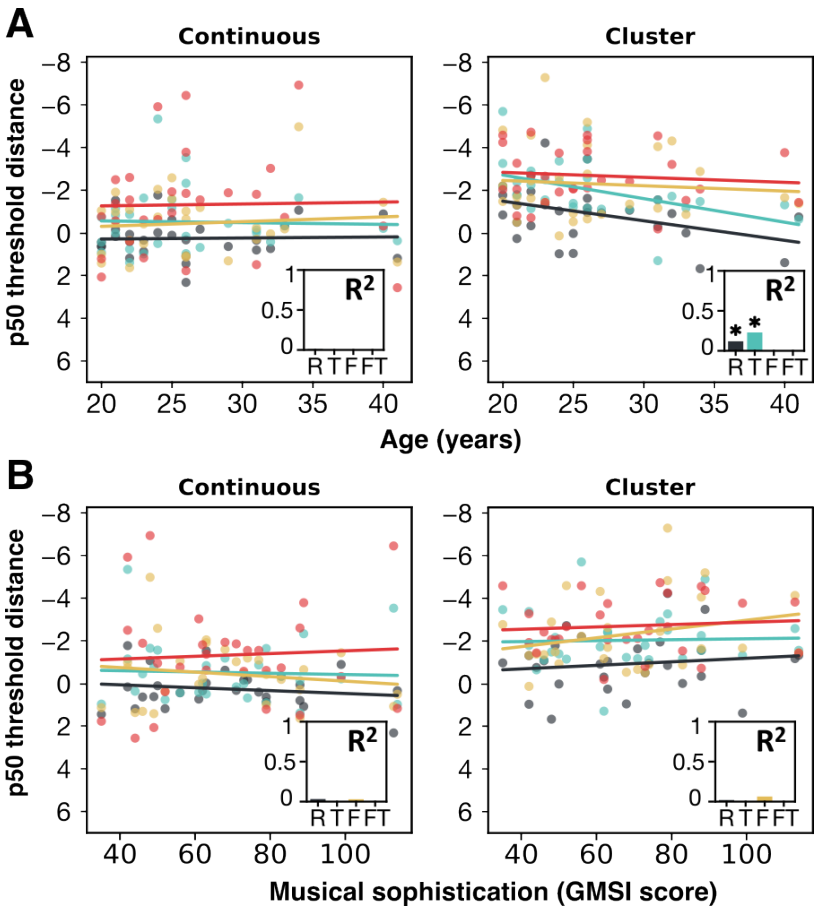

Figure 5: Predictive effects variability largely unexplained by age and musicianship*. A – correlation between threshold distance and participant Age, in continuous paradigm. Scatter plots indicate individual participants, lines indicate linear regression. R^2 of variance explained is reported in upper right insets. Asterisks indicate significance. B – Correlation between threshold distance and participant musicianship. Plot organization is the same as in A.*

Discussion

Our findings suggest that the human brain uses predictive information to enhance auditory sensitivity even at pure tone detection threshold, previously thought to be a largely low-level peripheral process. Between our least and most predictable tasks (the Randomized to the clinical 3AFC paradigm), we found a 7-decibel average difference within the same subjects. This effect is sizeable when compared against the 5 dB step size used for clinical measurement (Favier et al., 2018). We then use more high-resolution methods to identify how predictive structure in time and frequency contribute to this difference. We find evidence that predictability in both frequency and time contribute substantially to this predictive effect at two timescales, a slow pace consistent with audiological protocols 1 - 3 seconds per tone) and a faster pace consistent with time scales of sequences of acoustic events like speech and music (400 - 800 ms). Our findings reveal the influence of top-down (cortical) modulation of peripheral *sensitivity*, something unexplored in audiological fields.

Our study investigated the role of prediction under two scenarios: a slow-paced condition we refer to as *continuous* in which participant's continuously report detection on every tone they hear, spaced out about 1-3 seconds, and a faster-paced condition we refer to as *cluster*, whereby participants first hear a cluster of cue tones (400-800 ms between them) which cue predictively (or not) the timing or frequency (or both) of a target detection tone. Our

strongest results occur in the cluster case. This finding is in keeping with the previous literature on prediction. In this faster paced case, tones are detected as part of a sequence occurring at a pace typical of music. Previous work has shown that predictions are strong within these sorts of sequences and peak particularly at around this time scale (Arnal et al., 2015; Morillon et al., 2019). While the continuous task also consists of a sequence of tones, the amount of time spaced between tones is quite large making it more likely to be processed as isolated inputs. Previous work has shown that particularly in non-musicians, sequences are processed differently when item rate is less than 1 Hz (Doelling & Poeppel, 2015).

Further proving that the two protocols reveal distinct mechanisms of prediction is our clustering analysis. We initially expected that, regardless of pacing, participants would be grouped by the type of prediction (frequency or time). Instead, we found the opposite effect, participant performance was best clustered by the experimental protocol. Furthermore, each predictable condition within each protocol clustered together, away from the random condition. These findings suggest that similarity across participants is driven more so by prediction ability at distinct timescales rather than over distinct features. This clustering is not due to the overall mean shift in performance across the protocols. Our clustering method operates over the cosine similarity across participants, which ignores this overall shift.

Our results exclude the possibility of two alternative hypotheses to explain the improvement in thresholds: 1) a change in decisional biases and 2) a change in cochlear state. A change in decisional bias would plausibly lead to improved thresholds if for example participants indiscriminately indicated detecting the sound regardless of the stimulus. We address this concern by including catch trials in which no sound was presented. By analyzing these trials, we find that decisional bias cannot explain our results. While in the slower paced experiment, false alarms increased with temporal prediction, predictability in frequency shows no increase in false alarms. In the faster-paced experiment, false alarms go down with predictability (though not significantly). From this we can negate the first alternative hypothesis: a significant portion of our results must be due to an increase in *sensitivity* not to a change in decisional bias.

The second alternative hypothesis suggests that the increase in sensitivity is due not to central processing but instead to a local change in the cochlear processor. A myriad of studies have shown that local peripheral responses can change as a result of preceding and concurrent stimuli including masking effects (Harris & Dallos, 1979), and noise protection through the modulatory effect of the Medial Olivocochlear Reflex (Morand-Villeneuve et al., 2002). Our experimental design also negates this possibility as the gain in performance is demonstrated in comparison against a random control which maintains the stimulus context in every way except predictability. For example, comparing the effect of predictability in timing vs random, in this case the frequency distributions of tones, the temporal intervals between tones and the visual cue for the target window are exactly the same. The only change between these two conditions is that the temporal intervals are held constant per trial in the predictable case and shuffled for the random. To take advantage of this, the cochlea would need to execute computations that are expected in cerebellum, basal ganglia and motor cortex (for a recent review, see Cannon & Patel, 2021). From our perspective, a temporal prediction redundancy in the cochlea is a far less plausible hypothesis.

Instead, we propose that our findings result from an interaction of top-down predictive information with cochlear responses, enhancing sensitivity for specific frequencies and time points based on expectation. Whether this interaction involves a change in cochlear processing directly or instead of the central interpretation of cochlear responses will require

future work. However, either hypothesis has major implications for the fields of audiology and neuroscience. For example, if the hypothesis of altering cochlear responses via central input is correct, it would provide a mechanism for phase locked loops whereby the central perceptual system can alter its bottom-up input, similar to eye movements or whisking (Ahissar, 2003; Ahissar & Arieli, 2012), a pathway that has often been suspected but remains missing in the auditory system. Alternatively, if the results are due to a change in central interpretation, the findings upend the assumption in audiology that pure tone audiogram measure purely cochlear health (Musiek et al., 2017).

While audiologists are aware that predictability can alter their results, – it is standard practice in several countries to present tones with random timing, for example – this effect is thought to be due to our first alternative hypothesis: decisional bias, patients may report detecting a sound regardless of what they have heard, purely because they know where the tone should have been. Here we show for the first time that prediction can not only alter decisional bias but also enhance sensitivity and to a significant degree. Therefore, depending on the chosen clinical set up of the paradigm, the pure tone audiogram may be diagnosing more fully the auditory neural apparatus, mixing effects of both peripheral cochlear health and central, predictive ability. From the perspective of standard practice, this finding may not represent a significant issue for clinicians: the typical sounds that patients experience in their lives have some degree of predictability in them and as such current clinical setups match patients' daily experience. Still, our findings reveal two separable components present in sensory detection which may have separate trajectories of development and degradation over the course of lifespan. Future work will investigate these trajectories more fully.

To that end, while the current task is not designed to study the individual differences in perception, we used the natural variance in our participant cohort to assess how age could affect the predictive trajectory. The only significant effect we found was in the random and timing predictions of our fast-paced task. Interestingly predictability in frequency (F and FT) removed this effect, suggesting that even in younger populations, age reduced sensitivity but that this effect could be masked by improved predictability. Future work will assess in more detail the trajectory of this effect both with age and with clinical sensory hearing loss to assess how prediction can be used either as a coping mechanism or an indicator of future decline.

Together, our findings reveal that pure tone audiometric measurements contain the influence of both bottom-up sensory responsivity and top-down predictive modulation. We show that prediction in both frequency and time and at multiple timescales can influence the sensitivity of the auditory system. In particular, signal detection is most improved in the context of sequences with timing similar to those of typical auditory stimuli like speech and music. We also show key differences in behavior between these timescales including that temporal predictions at slower timescales may influence more decisional bias whereas frequency predictions improve sensitivity. Lastly, we look for individual differences to understand the variability in predictive performance. Even in our younger population, we find an effect of age on sensory detection in fast paced sequences which is can be recovered by frequency predictability. We expect our findings to inspire audiologists and auditory neuroscientists to track raw predictive ability more carefully across age and lifespan as a potential key factor in understanding hearing abilities.

Acknowledgements

The authors would like to thank Benjamin Morillon for his input and discussions on this project. In addition, we thank Céline Quinsac, Paul Avan, and Lavinia Slabu for their insights and discussions from the audiological perspective.

Funding:
Fondation Fyssen Postdoctoral Fellowship (KBD)
Fondation Pour l'Audition grant RD-2020-10 (LHA)

References

Ahissar, E. (2003). Closed-loop Neuronal Computations: Focus on Vibrissa Somatosensation in Rat. *Cerebral Cortex*, *13*(1), 53–62. https://doi.org/10.1093/cercor/13.1.53

Ahissar, E., & Arieli, A. (2012). Seeing via Miniature Eye Movements: A Dynamic Hypothesis for Vision. *Frontiers in Computational Neuroscience*, *6*. https://doi.org/10.3389/fncom.2012.00089

Arnal, L. H., Doelling, K. B., & Poeppel, D. (2015). Delta-Beta Coupled Oscillations Underlie Temporal Prediction Accuracy. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(9), 3077–3085. https://doi.org/10.1093/cercor/bhu103

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), Article 4. https://doi.org/S0896-6273(12)00959-2 [pii] 10.1016/j.neuron.2012.10.038

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. In *Journal of the royal statistical society. Series B (Methodological)* (pp. 289–300).

Borra, T., Versnel, H., Kemner, C., Van Opstal, A. J., & Van Ee, R. (2013). Octave effect in auditory attention. *Proceedings of the National Academy of Sciences*, *110*(38), 15225–15230. https://doi.org/10.1073/pnas.1213756110

Brant, L. J., & Fozard, J. L. (1990). Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging. *The Journal of the Acoustical Society of America*, *88*(2), 813–820. https://doi.org/10.1121/1.399731

British Society of Audiology. (2018). *Pure tone air and bone conduction threshold audiometry with and without masking*. https://www.thebsa.org.uk/resources/pure-tone-air-bone-conduction-threshold-audiometry-without-masking/

Cannon, J. J., & Patel, A. D. (2021). How Beat Perception Co-opts Motor Neurophysiology. *Trends in Cognitive Sciences*, *25*(2), 137–150. https://doi.org/10.1016/j.tics.2020.11.002

Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences*, *116*(20), 10113–10121. https://doi.org/10.1073/pnas.1816414116

Doelling, K. B., & Poeppel, D. (2015). Cortical entrainment to music and its modulation by expertise. *Proceedings of the National Academy of Sciences*, *112*(45), E6233–E6242. https://doi.org/10.1073/pnas.1508431112

Favier, V., Vincent, C., Bizaguet, É., Bouccara, D., Dauman, R., Frachet, B., Le Her, F., Meyer-Bisch, C., Tronche, S., Sterkers-Artières, F., & Venail, F. (2018). French Society of ENT (SFORL) guidelines (short version): Audiometry in adults and children. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, *135*(5), 341–347. https://doi.org/10.1016/j.anorl.2018.05.009

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), Article 2. https://doi.org/10.1038/nrn2787

Gehmacher, Q., Reisinger, P., Hartmann, T., Keintzel, T., Rösch, S., Schwarz, K., & Weisz, N. (2022). Direct Cochlear Recordings in Humans Show a Theta Rhythmic Modulation of Auditory Nerve Activity by Selective Attention. *Journal of Neuroscience*, *42*(7), 1343–1351. https://doi.org/10.1523/JNEUROSCI.0665-21.2021

Giard, M.-H., Collet, L., Bouchet, P., & Pernier, J. (1994). Auditory selective attention in the human cochlea. *Brain Research*, *633*(1), 353–356. https://doi.org/10.1016/0006-8993(94)91561-X

Greenberg, G. Z., & Larkin, W. D. (1968). Frequency-Response Characteristic of Auditory Observers Detecting Signals of a Single Frequency in Noise: The Probe-Signal Method. *The Journal of the Acoustical Society of America*, *44*(6), 1513–1523. https://doi.org/10.1121/1.1911290

Harris, D. M., & Dallos, P. (1979). Forward masking of auditory nerve fiber responses. *Journal of Neurophysiology*, *42*(4), 1083–1107.

Heilbron, M., & Chait, M. (2017). Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? *Neuroscience*. https://doi.org/10.1016/j.neuroscience.2017.07.061

Hesselmann, G., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2010). Predictive Coding or Evidence Accumulation? False Inference and Neuronal Fluctuations. *PLOS ONE*, *5*(3), e9926. https://doi.org/10.1371/journal.pone.0009926

Hudspeth, A. J. (2014). Integrating the active process of hair cells with cochlear function. *Nat Rev Neurosci*, *15*(9), Article 9. https://doi.org/10.1038/nrn3786

Lukas, J. H. (1980). Human Auditory Attention: The Olivocochlear Bundle May Function as a Peripheral Filter. *Psychophysiology*, *17*(5), 444–452. https://doi.org/10.1111/j.1469-8986.1980.tb00181.x

Marin, N., Lobo Cerna, F., & Barral, J. (2022). Signatures of cochlear processing in neuronal coding of auditory information. *Molecular and Cellular Neuroscience*, *120*, 103732. https://doi.org/10.1016/j.mcn.2022.103732

Morand-Villeneuve, N., Garnier, S., Grimault, N., Veuillet, E., Collet, L., & Micheyl, C. (2002). Medial olivocochlear bundle activation and perceived auditory intensity in humans. *Physiol Behav*, *77*(2–3), 311-20.

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neuroscience and Biobehavioral Reviews*, *107*, 136–142. https://doi.org/10.1016/j.neubiorev.2019.09.012

Musiek, F. E., Shinn, J., Chermak, G. D., & Bamiou, D.-E. (2017). Perspectives on the Pure-Tone Audiogram. *Journal of the American Academy of Audiology*, *28*(07), 655–671. https://doi.org/10.3766/jaaa.16061

Nobre, A. C., Correa, A., & Coull, J. (2007). The hazards of time. *Current Opinion in Neurobiology*, *17*(4), 465–470. https://doi.org/10.1016/j.conb.2007.07.006

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. In *Journal of machine learning research* (Vol. 12, Issue Oct, pp. 2825–2830).

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pickles, J. (1998). An Introduction to the Physiology of Hearing: Fourth Edition. In *An Introduction to the Physiology of Hearing*. Brill. https://brill.com/display/title/23098

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception & Psychophysics*, *42*(3), 215–223. https://doi.org/10.3758/BF03203073

Schlittenlacher, J., Turner, R. E., & Moore, B. C. J. (2018). Audiogram estimation using Bayesian active learning. *The Journal of the Acoustical Society of America*, *144*(1), 421–430. https://doi.org/10.1121/1.5047436

Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and Statistical Modeling with Python*. 92–96. https://doi.org/10.25080/Majora-92bf1922-011

Tanner, W., & Norman, R. (1954). The human use of information–II: Signal detection for the case of an unknown signal parameter. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 222–227. https://doi.org/10.1109/TIT.1954.1057462

Vallat, R. (2018). Pingouin: Statistics in Python. *The Journal of Open Source Software*, *3*(31), 1026.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), Article 3. https://doi.org/10.1038/s41592-019-0686-2