Distinguishing fine structure and summary representation of sound textures from neural activity

Author list: Martina Berto*[1], Emiliano Ricciardi[1], Pietro Pietrini[1], Nathan Weisz[2,3], Davide Bottari[1]

Affiliations:

1 *Molecular Mind Lab, IMT School for Advanced Studies Lucca, Lucca, Italy*
2 *Centre for Cognitive Neuroscience and Department of Psychology, University of Salzburg, Austria*
3 *Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria*

Corresponding author

Martina Berto

martina.berto@imtlucca.it

ABSTRACT

The auditory system relies on detailed and summary representations; when local acoustic details exceed system constraints, they are compacted into a set of average statistics, and a summary structure emerges. Such compression is pivotal for abstraction and sound-object recognition. Here, we assessed whether computations subtending local and statistical representations of sounds could be distinguished at the neural level. A computational auditory model was employed to extract auditory statistics from natural sound textures (i.e., fire, wind, rain) and to generate synthetic exemplars in which local and statistical properties were controlled. Participants were passively exposed to auditory streams while the EEG was recorded. In distinct streams, we manipulated sound duration (short, medium, long) to vary the amount of acoustic information. Short and long sounds were expected to engage local or summary statistics representations, respectively. Data revealed a clear dissociation. As predicted, in discriminations based on local information – compared to summary-based ones – auditory responses of greater magnitude were measured selectively for short sounds, while the opposite pattern emerged for longer sounds. Neural oscillations revealed that local features and summary statistics representations rely on neural activity occurring at different temporal scales, faster (beta) or slower (theta-alpha), respectively. These dissociations in neural response emerged without explicit engagement in a discrimination task, strongly suggesting that such processing may be pre-attentive in nature. Overall, this study demonstrates that the auditory system developed a neural architecture relying on distinct coding to automatically discriminate changes in the auditory environment based on acoustic details and their summary representations.

SIGNIFICANCE STATEMENT

Prior to this study, it was unknown whether we could directly measure auditory discriminations based on local features or statistical properties of sounds. Results show that the two auditory modes (local and summary statistics) are pre-attentively attuned to the temporal resolution (high or low) at which a change has occurred. In line with the temporal resolutions of auditory statistics, faster or slower neural oscillations (temporal scales) code sound changes based on local or summary representations. These findings expand our knowledge of some fundamental mechanisms underlying the function of the auditory system.

INTRODUCTION

The human auditory system is capable of discriminating sounds at both high and low temporal resolutions (McAdams, 1993; Griffiths, 2001). The processing of fine details relies on extracting and retaining local acoustic features (on the order of a few milliseconds) to detect transient changes over time (Plomp, 1964; McDermott, Schemitsch, and Simoncelli, 2013; Dau, Kollmeier, and Kohlrausch, 1997). These temporal variations characterize different sound objects and help the system discern among acoustic sources. However, environmental inputs typically comprise long-lasting sounds, in which the number of local features to be retained exceeds sensory storage capacity. This prohibits discrimination based on temporally detailed analysis from giving way to compressed representations (McDermott, Schemitsch, and Simoncelli, 2013). As the duration of the entering sounds increases, summary representations are built upon fine-detailed acoustic features to condense information into a more compact, and retainable structure (Yabe et al., 1998). The emergence of summary representations allows abstraction from detailed acoustic features and prompt sound categorization (McDermott and Simoncelli, 2011; McDermott, Schemitsch, and Simoncelli, 2013).

For stationary sounds (such as sound textures, e.g., rain, fire, waterfall, typewriting; Saint-Arnaud and Popat, 1995), characterized by a constant repetition of similar events over time, this form of compression consists of a set of auditory statistics

comprising averages over time of acoustic amplitude modulations (Giraud et al., 2000; Lorenzi et al., 1999; McDermott and Simoncelli, 2011; Figure S1A). Computational approaches in auditory neuroscience allow the mathematical formalization of this set of auditory statistics. The basic assumption is derived from information theories (Barlow, 1961) and suggests that if the brain represents sensory input with a set of measurements (statistics), any signal containing values matching those measurements will be perceived as the same (Figure 1A).

Psychophysical experiments reveal that stimuli including the same summary statistics but with different local features are easy to discriminate when they are short, but that as duration increases and summary representation takes over, they are progressively more difficult to tell apart. On the other hand, when sounds comprise different statistics, their perceived dissimilarity will increase with duration as their summary representations diverge (Berto et al., 2021; McDermott, Schemitsch, and Simoncelli, 2013; Figure 1B, right panel). While some evidence exists in the animal model (Zhai et al. 2020), in humans the neural activity underpinning auditory analyses based on local features and summary statistics is unknown (see Zhai et al., 2020 for results in rabbits). Moreover, previous behavioral studies required participants to actively attend to the stimuli to perform tasks. From this evidence alone, it thus remains unanswered whether discrimination based on local features and their summary statistics can also occur automatically and, possibly, pre-attentively (e.g., Triesman et al., 1992).

To fill these gaps, we used a validated computational auditory model (McDermott and Simoncelli, 2011) to extract a set of auditory summary statistics from natural sounds and to generate synthetic sounds that feature this same set of measurements (see Material and Methods). With this approach, it is possible to impose the same set of statistics to sounds (white noise samples) that originally had different local structures (Figure 1B, S1B). Employing this synthesis approach, we could thus create sounds that differ at high temporal resolutions (e.g., local features) but are perceptually indistinguishable at lower ones (summary statistics), and vice-versa.

We acquired EEG measurements in participants being passively exposed to a stream composed of triplets of sounds, presented at a fast stimulation rate (2Hz). To ensure generalizability, sounds were randomly drawn from a large set of synthetic excerpts (see Material and Methods). Within each triplet, the first two sounds were

repeated, while the third was novel. Two experiments were designed based on the sound property to be discriminated. (1) In Local Features, the novel and repeated sounds differed only in their local structures, with unaltered auditory statistics; (2) in Summary Statistics, the novel sound differed from the repeated sounds in auditory statistics. As statistical variability is expected to change with sound duration (McDermott, Schemitsch, and Simoncelli, 2013), we presented separate sound streams comprising stimuli of different lengths (either 40, 209, or 478ms; Figure 1C). We first investigated simple auditory evoked responses to uncover magnitude changes in the neural activity associated with the two modes of representation. We predicted that short and long sounds would prompt larger auditory-discriminative responses for local features and summary statistics, respectively. In line with this prediction, we expected local information to be encoded at a faster timescale than encoding for summary statistics. To this end, we investigated neural oscillations and assessed whether information conveyed at different timescales (e.g., Giraud and Poeppel 2012; Panzeri et al., 2010) could reveal specific fingerprints of discriminations based on local details and summary statistics (see Figure 1C, bottom panel).

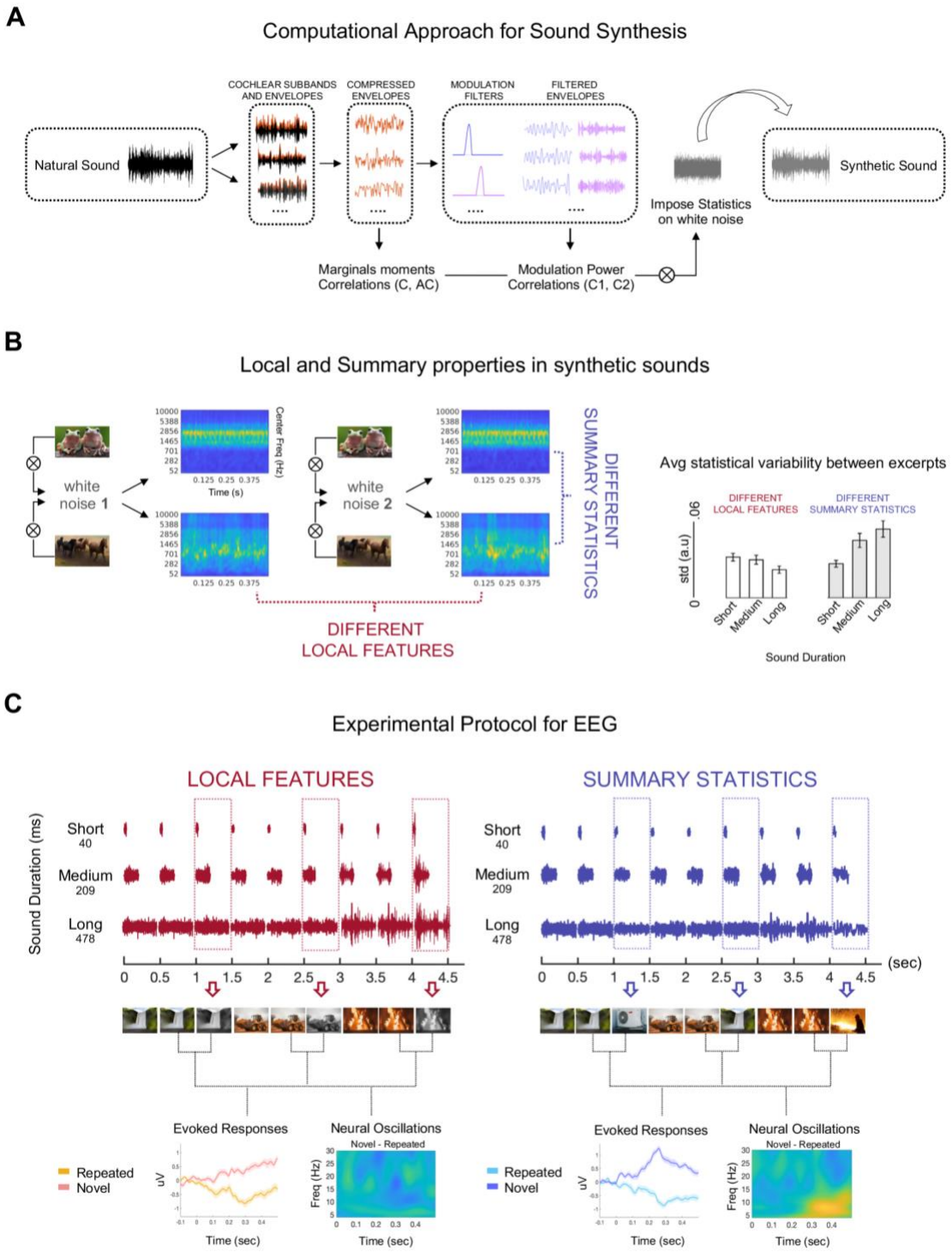

Figure 1. Using a computational approach to unveil cortical representations of basic auditory processing.

(A) Computational Approach for Sound Synthesis. An original recording of a natural sound texture is passed through an auditory model.

The model provides a mathematical formulation of the computations (auditory statistics) performed by the auditory system to recognize that sound object. The signal is filtered with 32 audio filters to extract analytic and envelope modulation for each cochlear subband. Envelopes are downsampled and multiplied by a compression factor. From the compressed envelopes, a first set of statistics is computed: marginal moments (including envelope mean, variance, skewness), auto-correlation between temporal intervals, and cross-band

correlations. Compressed envelopes are then filtered with 20 modulation filters. The remaining set of statistics is extracted from the filtered envelopes: modulation power, cross-band correlations between envelopes filtered with the same modulation filter (C1) and between the same envelope filtered through different filters (C2). Auditory statistics values are then imposed on a white noise sample. The result is a synthetic sound, which is constrained only by the imposed set of measurements. This sound will be perceptually similar to the original and will represent a synthetic exemplar of that sound texture.
(B) Local and Summary properties between synthetic sounds. Two different sets of statistics are extracted from two sound textures: “frogs” and “horse trotting”. Each set of values is imposed on two different random-white-noise samples. When the same statistics are imposed on different white-noise samples, the outcomes are two synthetic exemplars of the same sound texture. These exemplars will have the same summary statistical representation but will diverge in their local features, as they will be influenced by the original input sound. When different statistics are imposed on the same white-noise sample, the results are two synthetic exemplars that will diverge in their overall summary statistics and that will be perceptually associated with different sound objects. The cochleograms of 0.5s synthetic exemplars are displayed in the left panel. In the right panel, bar plots represent the standard deviation of all statistics (averaged) across sound excerpts of different durations (Short= 40ms; Medium= 209ms; Long= 478ms). If the sound excerpts feature the same summary statistics but originate from different white noise samples, their statistical variability decreases with duration. When they are derived from the same input sound, but different statistics are imposed, variability increases. Error-bars represent the standard error of the mean (SE).
(C) Experimental protocol for EEG. Triplets of sounds were presented at a fast rate (one sound every 500ms). Two sounds were identical (Repeated), while the third was different (Novel) and could vary in its local features (left) or summary statistics (right), depending on the experiment (Local Features or Summary Statistics). Three sound durations, equally spaced logarithmically, (Short, Medium, and Long: 40, 209, and 478ms) were employed (in different sound streams) to tap into each auditory mode separately (local features vs. summary statistics processing). By comparing the response elicited by the repeated sound with the one associated to the novel one, it was possible to measure neural correlates of sound discrimination based on local features or summary statistics (in this study: magnitude and neural oscillations).

MATERIALS AND METHODS

Participants

Twenty-four normal-hearing right-handed young adults (F= 12; mean age= 27.13 years, std= 2.83) took part in the experiment. This sample size was estimated via simulations. We used the procedure described by Wang and Zhang (2021; for details on sample size estimation see Supplementary Information and Figure S1C). All participants were healthy; they were fully informed of the scope of the experiment; they signed written, informed consent prior to testing, and they received monetary compensation. The study was approved by the regional Ethical Committee

(CEAVNO protocol n 24579), and the protocol adhered to the guidelines of the Declaration of Helsinki (2013).

Stimuli

Synthetic sounds were generated using a previously validated computational auditory model of the periphery. The auditory model and synthesis toolbox are available here: http://mcdermottlab.mit.edu/downloads.html.

This auditory model emulates basic computations occurring in the cochlea and the mid-brain (McDermott and Simoncelli, 2011).

The signal (7s original recording of a sound texture, N=54; see Table S1) was decomposed into 32 cochlear subbands, using a set of gammatone filter-banks with different central frequencies spaced on an ERBs scale. Absolute values of the Hilbert transform for each subband was computed to extract the envelope modulation of each cochlear channel over time. Envelopes were then compressed to account for the non-linear transformations performed by the cochlea, and the first set of statistics was measured from the transformed envelopes: mean, skewness, variance, auto-correlation (within each cochlear channel), and cross-correlation (between channels). Additional filtering was applied to the envelopes to account for the modulatory response of the spectro-temporal receptive fields of neurons in the midbrain (Bacon and Wesley Grantham, 1989; Dau et al., 1997). Three additional statistics resulting from these operations could be derived: modulation power, C1, and C2 (respectively, the correlation between different envelopes filtered through the same modulation filter and the correlation between the same envelopes filtered by different modulation filters; Figure 1A).

The resulting set of statistics – extracted from the original recording of sound textures – was imposed on four 5s white noise samples (Figure 1A, S1A). This allowed the generation of different sound exemplars, which varied selectively in their local features but included the same long-term summary representation (Figure 1B, C). All synthetic exemplars featuring the same auditory statistics were perceptually very similar to the original sound texture from which they were derived, even when their input sounds (white noise) varied (Figure 1B; see also Figure S1B). Synthetic sounds with the same imposed auditory statistics represent different exemplars of the same sound texture with the same summary statistics but a different fine structure. This is because, in the synthesis procedure, the imposed statistics are

combined with the fine structure of the original white noise sample (Figure S1A). Due to inconstant local acoustic features, the statistical variability between different exemplars of the same texture will be high for short excerpts and will progressively decrease with increased sound duration: as the imposed summary statistics emerge, sound statistics converge to the same set of original values (Figure 1B, right panel). On the other hand, when excerpts are derived from different Sound Textures, their variability will increase with sound duration, because the emerging summary representations will converge on different original values. In other words, with increasing sound duration, it is easier to discriminate sounds based on their summary representation, as they diverge (Figure 1B).

Importantly, to create experimental stimuli, all four 5-second synthetic exemplars were cut from the beginning to the end into excerpts of different lengths, either short (40ms), medium (209ms), or long (478ms). These lengths were chosen based on results in previous behavioral investigations (Berto et al., 2021; McDermott, Schemitsch, and Simoncelli, 2013). Excerpts were equalized to the same root mean square (RMS= 0.1) and had a sampling rate of 20kHz. Which experimental stimuli to present for each run were randomly drawn from all available excerpts, according to the experiment requests (see below).

Procedure

Participants were tested in a sound-isolation booth. After reading instructions on a monitor, they listened to the sounds in the absence of retinal input (participants were blindfolded to prevent visual input).

For each run of the experimental session, a sound sequence lasting 108sec was presented. The sequence contained triplets of sounds (n = 216), presented one after the other to form an almost continuous sound stream, in which sound onsets occurred every 500ms (Figure 1D). Within each sequence, all sounds had the same duration (either 40, 209, or 478ms).

Two experiments were implemented: (1) In Local Features, two different 5s synthetic exemplars of the same sound texture were selected (out of the four we had created); these exemplars were cut into brief excerpts of either 40, 209 or 478ms. According to the selected duration (which was different for each sequence), two excerpts – one for each exemplar – were selected from among those available. The two excerpts had the same starting point (in seconds) from the onset of the 5s exemplar. The first

sound excerpt was repeated twice, and afterwards, the other was presented as the third element in the triplet.

Thus, two sounds within a triplet were identical (repeated), while the third one (novel) comprised different temporal local features but the same summary statistics; importantly, repeated and novel sounds had the same identity (both could be, e.g., waterfall) but different acoustic details (Figure 1C, left panel; Table S1, column 1).

(2) In Summary Statistics, sound textures were coupled according to their perceived similarity (McDermott, Schemitsch, and Simoncelli, 2013; see Table S1, column 1 and 2). For each sound texture, one of the four 5s synthetic exemplars was selected, and an excerpt of the required duration (40, 209, or 478ms) was picked randomly from among those available. The same was done for the coupled texture, matching the exemplar number (so that the original input noise they were derived from was the same and the sounds only varied in their imposed statistics) and the starting point in seconds.

Again, the first excerpt was repeated twice, while the other was used as the last sound in the triplet. The novel sound thus deviated from the other two (repeated) in its auditory statistics, extracted as it was from an exemplar of a different sound texture. This means that the novel sound was a different sound object (e.g., the repeated sounds might both be waterfall excerpts, and the novel one an air conditioner; see Figure 1C, right panel).

To ensure generalizability, the sound textures were different across triplets, so the statistical variability between repeated and novel sounds was kept constant within an experiment while presenting different types of sound objects.

Discriminative responses emerging from the contrast between the novel and repeated sounds did not depend on specific properties (e.g., a change in frequency between a particular type of sound category) but only on their local or statistical changes.

In both experiments, the order of the triplets was shuffled for each participant and run. Moreover, excerpts were selected randomly from among those that shared the required characteristics, so not only the presentation order but also stimuli *per se* were never the same across participants.

Crucially, in the two experiments, the statistical variability between repeated and novel sounds changed as a function of sound duration in opposite directions:

decreasing variability with longer duration in Local Features, while increasing variability in Summary Statistics.

A total of six conditions were employed: two experiments (Local Features and Summary Statistics) for three sound durations (40, 209, 478). Two sequences/runs per condition (Experiment * Duration) were presented, for a total of twelve runs. The order of runs was randomized across participants, and short breaks were taken between runs. Disregarding duration, in the sound stream, excerpts were always presented in triplets, with those repeated presented twice. This was to prevent potential differences (e.g., standard formation; Sussman and Gumeyuk, 2005) and expectancy effects from influencing results.

Since in each run (for both experiments), the interstimulus gap depended on sound duration (sound onset was kept constant at every 500ms), comparisons were assessed only between experiments but within duration. That is, we tested whether neural response was different when local changes occurred, as compared to statistical, expecting local processing to be favored for short sounds and statistical processing for long.

Participants had to listen to the sound stream but were asked to perform an orthogonal task, consisting of pressing a button when a beep sound was heard. The beep was a pure tone higher in pitch and intensity than the sound-texture stream. The pure tone was 50ms in length, had a frequency of 2200Hz, an amplitude of 50dB, a sampling rate of 20kHz, and RMS of 5. The beep randomly occurred during the stimulation. The number of beeps varied randomly across runs, from 0 to 3. Detection was considered valid when the participant pressed the key within an arbitrary window of 3 seconds from beep occurrence (behavioral results are reported in Supplementary Information and Figure S1D).

EEG recording

Electroencephalography (EEG) was recorded from an EGI HydroCel Geodesic Sensor Net with 65 EEG channels and a Net Amps 400 amplifier (Electrical Geodesics, Inc., EGI, USA). The acquisition was obtained via EGI's Net Station 5 software (Electrical Geodesics, Inc., EGI, USA). Central electrode E65 (Cz) was used as reference. Four electrodes were located above the eyes and on the cheeks to capture eye movements. Electrode impedances were kept below 30 kΩ.

Continuous EEG signal was recorded throughout the session with a sampling rate of 500Hz.

Experiment sounds were played from a stereo speaker (Bose Corporation, USA) positioned in front of the participant and at 1 meter distance from the eyes; sound loudness was kept constant across participants and runs. The experiment ran on MATLAB (*R2018b*; Natick, Massachusetts: The MathWorks Inc.); written instructions were displayed only at the beginning of the experimental session, via Psychtoolbox version 3 (Brainard and Vision, 1997; PTB-3; http://psychtoolbox.org/).

EEG Data Analysis

Preprocessing

Data were preprocessed with a semi-automatic pipeline implemented in MATLAB (see Stropahl et al., 2018; Bottari et al., 2020). Preprocessing was performed using EEGLAB (Delorme and Makeig 2004; https://sccn.ucsd.edu/eeglab/index.php). Data were loaded, excluding electrode E65 (Cz), which was the reference channel of our EEG setup (thus consisting only of zero values).

To remove slow drifts and DC offset, a high-pass filter (windowed sinc FIR filter, cut-off frequency 0.1 Hz, filter order 10000) was applied to the continuous signal.

A first segmentation in time was performed by epoching the signal according to event onset. To avoid boundary artifacts, for each run, the signal was cut 2 seconds before its onset event and until 2 seconds after the end of the presentation (thus, from -2 to +114 sec). For each participant, epochs were then merged in a single file containing only the parts of the signal referring to significant stimulation (thus excluding breaks in between trials).

Independent Component Analysis (ICA; Bell and Sejnowski, 1995; Jung et al., 2000a,b) was used to identify stereotypical artifacts. To improve ICA decomposition and reduce computational time, data were low-pass filtered (windowed sinc FIR filter, cut-off frequency 40Hz, filter order 50), downsampled to 250Hz, high-pass filtered (windowed sinc FIR filter, cut-off frequency 1Hz, filter order 500), and segmented into consecutive dummy epochs of 1sec to spot non-stereotypical artifacts. Epochs with joint probability larger than 3 standard deviations were rejected (Bottari et al., 2020). PCA rank reduction was not applied prior to ICA to avoid compromising its quality and effectiveness (Artoni, Delorme, and Makeig, 2018).

For each subject, ICA weights were computed using the EEGLAB runica algorithm and then assigned to the corresponding original raw (unfiltered) dataset. Topographies for each component were plotted for visual inspection. Artefacts associated with eye movements and blinks were expected, and so a CORRMAP algorithm (Viola et al., 2009) was used to semi-automatically remove components associated with such artefacts. Automatic classification of components was performed using the EEGLAB plugin ICLabel (Pion-Tonachini, Kreutz-Delgrado, & Makeig, 2019). Components representing eye movements and blinks were identified from their topographical map within the components that ICLabel had marked as 'Eye' with a percentage above 95%. Among these components, those with the highest rankings were selected from a single dataset and used as templates (one for eye movements and one for blinks). CORRMAP algorithm clusters ICA components with similar topography across all datasets to highlight the similarity between the IC template and all the other ICs. A correlation of the ICA inverse weights was computed, and similarity was allocated with a threshold criterion of correlation coefficient being equal to or greater than 0.8 (default value of CORRMAP; Viola et al., 2009). For all participants, on average 1.92 components were removed (std= 0.88; range= 0-4).

Bad channels were interpolated after visually inspecting the scroll of the entire signal and the power spectral density for each electrode. On average, 3.75 (range= 1-8; std= 2.21) channels were interpolated. Interpolation of noisy channels was performed via spherical interpolation implemented in EEGLAB.

Finally, the reference channel (Cz) was reintroduced in the EEG data of each participant and the datasets were re-referenced to the average across all channels.

Time-domain analysis

This analysis was performed to extract auditory evoked potentials and uncover phase-locked magnitude changes associated with the two modes of sound representation (Local Features or Summary Statistics).

Pre-processed data were low-pass filtered (windowed sinc FIR filter, cut-off frequency= 40Hz, filter order= 50). Additionally, de-trend was applied by filtering the data above 0.5Hz (windowed sinc FIR filter, cut-off frequency= 0.5Hz, filter order= 2000). Consecutive epochs (from -0.1 to 0.5sec) were generated, including segments of either the novel sounds or the repeated one (the second) of the triplets

for each participant and condition. Data were baseline corrected using the −0.1 to 0 sec pre-stimulus period. Joint probability was used to prune non-stereotypical artefacts (i.e., sudden increment of muscular activation); rejection threshold was 4 standard deviations (Stropahl et al., 2018). For novel sounds, on average, 16.58 epochs per participant were removed (std=5.42; range 5-30) out of the 144 concatenated epochs that each Experiment * Duration comprised; for repeated sounds, on average, 16.15 epochs were removed (std= 5.11; range 5-29), again out of 144 trials per condition.

Datasets were converted from EEGLAB to FieldTrip (Oostenveld, Fries, Maris, and Schoffelen, 2011; http://fieldtriptoolbox.org). Grand averages across participants were computed for each experiment, duration, and stimulus type (either repeated or novel). Data across trials were averaged generating Auditory Evoked Potentials (Figure S2).

We subtracted from the response to the novel sound the response to the preceding, repeated one in each triplet. Since all stimuli in the triplets (repeated and novel) were never the same across runs and participants, the subtraction was performed to ensure that neural responses were not driven by idiosyncratic differences in the stimuli that were presented in that specific run, but by the statistical difference between novel and repeated ones.

A non-parametric permutation test was performed between experiments (Local Features vs. Summary Statistics) for each duration (short, medium, and long), employing the differential auditory responses between the novel and repeated sounds. The permutation test was carried out under the null hypothesis that probability distributions across condition-specific averages were identical across experiments.

The cluster-based permutation approach is a nonparametric test that has the advantage of solving the multiple comparison problem of multi-dimensional data (in which you must control for several variables, such as time, space, frequencies, and experimental conditions. Maris and Oostenveld, 2007).

Notably, statistical analyses between experiments were performed only within each duration, to avoid possible confounds associated with refractoriness effects due to different interstimulus intervals (ISI) at long and short durations.

Thus, the contrasts of interest were: (1) Local Features short vs. Summary Statistics short; (2) Local Features medium vs. Summary Statistics medium; (3) Local Features long vs. Summary Statistics long.

A series of cluster-based permutation tests (Maris and Oostenveld, 2007; cluster alpha threshold of 0.05 (two-tailed, accounting for positive and negative clusters); 10000 permutations; minimum neighboring channels = 2) was performed. Cluster-based analyses were performed within a pool of central channels (according to EGI system, channels: E3, E4, E6, E7, E9, E16, E21, E41, E51, E54, E65), typically capturing auditory response and including all samples from 0 to 0.5. We expected novel sounds to elicit larger responses compared to repeated sounds.

Time-Frequency analysis

Following the differences in magnitude changes that we observed between experiments for long and short durations, we performed data decomposition in the time-frequency domain to test whether sound changes at a high temporal resolution (local features in short sounds) were encoded at faster timescales compared to the ones occurring at a low temporal resolution (summary statistics in long sounds). We investigated frequencies below 40Hz which have been associated with auditory processing in studies including both humans and animals (for review see Gourevic et al, 2020). Specifically, several studies have marked the relevance of lower (theta, alpha) and higher (beta) frequency bands, with respect to auditory feature integration (e.g., VanRullen, 2016; Teng et al., 2018) and detection of deviant sounds (e.g., Fujioka et al., 2012; Snyder and Large, 2005).

Preprocessed data were low-pass filtered to 100Hz (windowed sinc FIR filter, cut-off frequency= 100Hz, filter order= 20) to attenuate high frequencies and above Nyquist one and high pass filtered at 0.5Hz (as with time-domain data). Data were epoched into segments from -0.5 to 1sec from stimulus onset: either the second repeated or the novel. Joint probability was used to remove bad segments with a threshold of 4 standard deviations. On average, 11.96 epochs were removed for repeated sounds (range= 4-25; std= 4.28) and 11.58 for novel (range 4-26; std= 4.23). The resulting epoched datasets were converted to Fieldtrip for time-frequency analysis. We used complex Morlet wavelets to extract the power spectrum at each frequency of interest and time point. The frequencies spanned from 4 to 40Hz in steps of 2; the time window for decomposition comprised latencies from -0.5 to 1, around stimulus onset

(either novel or repeated) in steps of 20ms. Finally, the length of the wavelets (in cycles) increased linearly from to 3 to 6.32 cycles with increasing frequency (depending on the number of frequencies to estimate; N=19). The signal was zero-padded at the beginning and at the end to ensure convolution with the central part of the window. The resulting power spectrum for each participant was averaged across trials. Then, to account for the power scaling (1/f), we performed baseline correction. We applied a condition-averaged baseline (e.g., Cohen and Donner, 2013; Cohen and Cavanagh, 2011) corresponding to the 100ms prior to the occurrence of the repeated sound preceding the novel one. That is, within each duration, at the single participant level, we selected the period from -100 to 0ms before the onset of the repeated sound preceding the novel one separately for each experiment (Local Features and Summary Statistics) and averaged them. As a baseline normalization method, we selected relative change:

$$\frac{(pow(t) - bsl)}{bsl}$$

Where *pow* is the total power at each sample (t), within the latencies of interest for repeated and novel grand-averaged trials, and *bsl* is the averaged baseline (across Experiment and time). Grand-average of baseline-corrected power spectrums of all participants were computed.

We investigated the neural activity underlying the discrimination of novel and repeated sounds across experiments for short and long durations. Thus, we first subtracted the power at repeated trials from the one at novel trials and then used cluster-based permutation (Maris and Oostenveld, 2007) to investigate differences between neural responses to sound changes across experiments (Local Features vs. Summary Statistics) at each of the selected durations (short or long), at any latency (0 500ms) and across all (65) channels (minimum neighboring channels = 1). We used the period of the oscillatory activity as an index of the temporal scale of discriminative auditory processing. Following the inspection of power change between novel trials and repeated trials, oscillatory activity above 30 Hz was not considered. To avoid frequency band biases, we divided the power change into equally spaced frequency band ranges (8Hz each, in steps of 2Hz), creating a slow, medium, and fast oscillation range between 4 and 30Hz. These frequencies of

interest included canonical theta, alpha and beta oscillations (theta and alpha: 4-12Hz; low beta: 12-20Hz; high beta: 20-28Hz).

Depending on sound duration, we expected to detect different power modulations in response to Local Features changes as compared to Summary Statistics at different timescales (frequency bands). Cluster permutation was performed separately for each frequency range (10000 permutations). The directionality of the test was based on results in the Auditory Evoked Responses (see Time-domain results) and on the specific frequency ranges: specifically, for short duration, we expected power changes in higher frequencies in Local Features as compared to Summary Statistics. Conversely, at long duration, we expected greater power changes in the lower-frequency range in response to sound discriminations based on Summary Statistics compared to those baded on Local Features. For the short duration, we thus expected: Local Features > Summary Statistics in the 4-12Hz range, and Local Features < Summary Statistics in 12-20Hz and/or 20-28Hz. The opposite outcome was anticipated for the long duration: Summary Statistics > Local Features in the alpha-theta range; Summary Statistics < Local Features for beta bands (given the predefined directions of the effects, cluster alpha threshold was 0.05, one-tailed).

RESULTS

Time-domain results

By comparing Local Features vs. Summary Statistics separately for each sound duration, cluster permutation revealed a significant positive cluster, selectively for the short sound duration 40 ($p < 0.02$), lasting from 188ms to 220ms after stimulus onset. Following the prediction, results revealed a greater auditory potential of Local Features compared to Summary Statistics for short duration. No significant positive cluster was found for the medium (209) and long (478) sound durations (all $p > 0.39$). Conversely, a significant negative cluster was found, selectively for the long duration 478 ($p < 0.001$), lasting from 220ms to 308ms after stimulus onset. These results indicate a greater response for Summary Statistics compared to Local features at long durations only. No differences emerged for short and medium sound durations (all ps >0.33).

462 Results clearly reveal, at the neural level, a double dissociation based on stimulus
463 length and mode of representation (Figure 2). Findings support behavioral outcomes
464 for which discriminations based on local features processing is favored for brief
465 sound excerpts, while summary statistics are built at a slower temporal rate as
466 information is accumulated (i.e., Berto et al., 2021; McDermott, Schemitsch, and
467 Simoncelli, 2013). Going beyond past behavioral effects, our results clearly show
468 that local and summary representations can emerge automatically (and, putatively,
469 pre-attentively) from exposure to systematic sound changes. The perception of such
470 changes is based on the variability between sound excerpts in their acoustic details
471 and summary representation and can be manipulated as a function of the amount of
472 incoming information (i.e., sound duration), eliciting magnitude differences that can
473 be detected from brain responses and that match behavioral expectations.

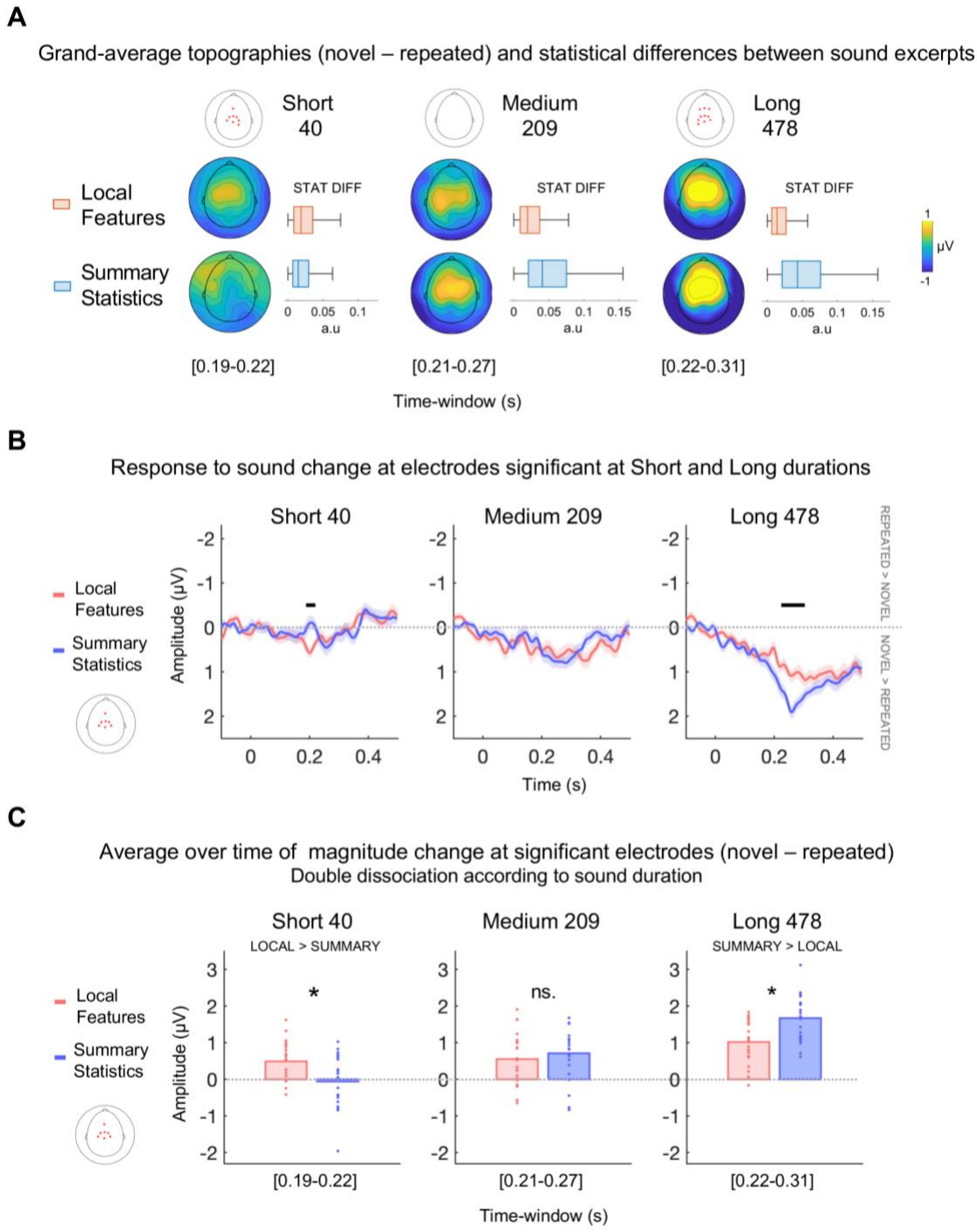

Figure 2. Results of time-domain analysis

(A) Grand-average topographies of the differential response associated with the sound change (novel sound minus repeated sound) at significant latencies, for each experiment and duration. For each latency, electrodes associated with significant clusters are displayed above as red stars on the scalp. ** $p < 0.025$.*

On the right side of topographical maps, the boxplots represent objective differences between the novel and repeated sounds of all auditory statistics (averaged). The difference was computed between the statistics of sounds presented for each run, experiment and duration and averaged across all participants. Within each duration, medians differed at the 5% significance level between experiments. Local Features > Summary Statistics in short (40) duration and Summary Statistics > Local Features for medium (209) and long (478) durations.

Evoked response in the EEG is in line with the objective statistical difference measured from sound excerpts.
(B) Grand-average electrical activity (negative values are plotted up) of the differential response (novel minus repeated) at significant electrodes (in red) for both short and long durations. Shaded regions show interpolated repeated error of the mean (SE) at each time point. Positive values indicate that novel elicited a greater response than repeated.
Results of cluster permutation are displayed as black bars extending through significant latencies. – $p < 0.025$.
(C) Grand-average magnitude change (novel minus repeated) across participants at the same significant electrodes, averaged within significant latencies for each experiment and duration. For visualization only, since no significant latencies existed for the duration 209, an intermediate time window was used, by computing an average between the time windows of significant latencies for durations 40 and 478. A double dissociation according to sound duration emerges, with local features eliciting higher-magnitude changes at short duration and summary statistics at long. Bar plots represent average values across participants. Individuals' values are separately displayed by each point.

Time-Frequency Results

Since summary statistics emerge over time, we expected statistical variations to be encoded by slower oscillations than local feature changes. For such encoding, we expected power modulations at faster oscillations in response to local feature change in short sounds, and at slower oscillations in response to the emergence of a different set of summary statistics in long acoustic excerpts. To test this, we separated the power between 4 and 30Hz into three ranges, equally spaced: slow, 4-12Hz; medium, 16-20Hz; and fast, 20-28Hz. Then we used a nonparametric permutation approach to address whether differences between Local Features and Summary Statistics emerged according to sound duration (short or long) within the three frequency ranges.

Results followed the predicted pattern. For short sound duration, the analysis revealed a significant cluster between 100 and 220ms, in which sound change in Local Features elicited a greater decrease of power in the fastest oscillation range (20-28Hz; $p < 0.05$) compared to Summary Statistics (Figure 3A, left panel). This significant effect was located over left fronto-central and right posterior sensors (see Grand-average topography in Figure 3A, left). Conversely, for long sound duration, we found a greater increase of power in the slow oscillation range for Summary Statistics compared to Local Features (4-12Hz; $p < 0.03$); the significant cluster consisted mostly of left fronto-central channels and bilateral posterior channels and spanned from 260 to 500ms (Figure 3A, right panel). No differences of power were found between Local Features and Summary Statistics for any sound duration at

medium frequency range (12-20Hz ranges, at any latency; all ps > 0.24). Overall, results revealed that when sound duration is short, neural oscillations at higher frequency bands (canonically corresponding to high-beta band) desynchronize more when the acoustic discrimination is driven solely by local details; vice-versa when sound duration is long, i.e., higher low-frequency oscillations (alpha and theta bands) are associated with stimulus changes based on different summary statistics (Figure 3B).

Overall, these findings show that different temporal scales at the neural level underpin the discrimination of variant elements in the auditory environment based on the amount of information available and the type of sound change that has occurred. Notably, beta desynchronization for Local Features (short duration) peaks 100-150ms after stimulus onset, while the same effect in the time domain has a peak that builds up around 200ms. The opposite was found for Summary Statistics (long duration), in which theta-alpha synchronization starts about 40ms later than the effect observed in the time-domain and is more sustained over time (i.e., it lasts the entire time window). These differences suggest that the two measures are capturing at least partially different aspects of sound discrimination.

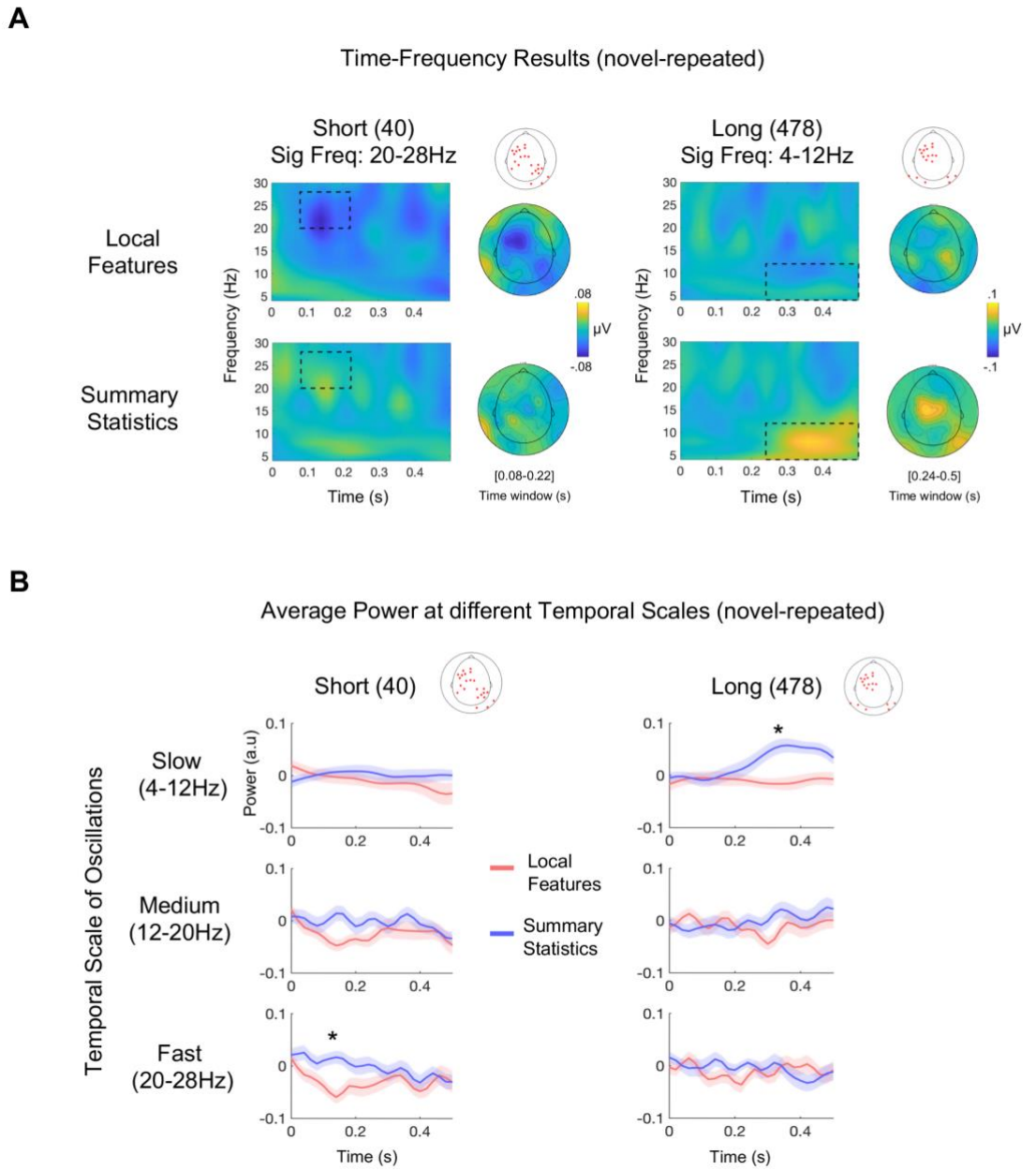

Figure 3. Results of Time-Frequency analysis

(A) Grand-average difference (novel minus repeated) of total power for short and long sound durations in both experiments (Local Features and Summary Statistics) at significant channels. Rectangular regions comprise the latencies and frequency ranges in which power changes were significant between experiments after cluster-based permutation. Significant channels are marked as red stars over the sketch of a scalp (* $p < 0.05$).

In the left panel, results for the short duration are displayed and show higher-power desynchronization in the 20-28Hz frequency range (high beta) for Local Features as compared to Summary Statistics. In the right panel, results for the long duration show higher 4-12Hz (alpha-theta) power synchronization for Summary Statistics as compared to Local Features. Grand-average topographical maps at significant latencies and frequency ranges are displayed next to the corresponding power-spectrum plots.

(B) Average power difference between novel and repeated sounds for each range of frequency bands (Slow, Medium, and Fast), averaged across all significant channels, plotted at all latencies (from 0 to 0.5s). Significant channels are marked as red stars over the sketch of a scalp. Shaded regions show interpolated standard error of the mean (SE) at each time point. * $p < 0.05$.

DISCUSSION

The auditory system extracts information at high (local) and low (summary) temporal resolution. Here, we aimed to assess whether discriminative responses to local or summary representations could be measured at the neural level and whether they are encoded at different temporal scales (Panzeri et al., 2010). We employed a computational model (McDermott and Simoncelli, 2011) to synthetically create stimuli with the same summary statistics but different local features. We used these synthetic stimuli to present streams of triplets containing repeated and novel sounds that could vary in their local features or summary statistics. Results in the time domain showed that, when sound duration was short, the magnitude of auditory potential increased selectively for local features changes. By contrast, when sound duration was long, changes in auditory statistics elicited a higher response compared to changes in local features (Figure 2A, B, C). Thus, according to sound duration, we observed an opposite trend in the magnitude change of the evoked response (Figure 2C). This trend perfectly matched expectations based on previous psychophysics evaluations (i.e., Berto et al., 2021). Importantly, analysis in the time-frequency domain revealed that neural activity at different temporal scales characterized discriminative responses to local features or summary statistics. Faster oscillations (beta range) were associated with discriminations based on local features, and slower oscillations (theta-alpha) with changes based on summary statistics.

Automaticity of Local Features and Summary Statistics Processing

Auditory responses to novel local features or summary statistics were associated with differences in magnitude that could be automatically detected. This finding suggests that the auditory system pre-attentively attunes its response to specific sound changes. This evidence expands seminal studies measuring the MisMatch Negativity (MMN) response (Näätänen et al., 1978; Tiitinen et al., 1994). MMN is the neural marker of a process by which the system “scans” for regularities in entering sounds and uses them as references to detect variations in the auditory scene (for reviews see Näätänen et al., 2001, 2010). In our study, expectations that a change would occur in the third element of the triplet had a probability of 1 in each experiment (Local Features and Summary Statistics; Figure 1D). Thus, our effects cannot be simply explained by spurious expectancy or attentional effects.

Coherently, MMN response to a deviant sound is not affected by prior expectations that the novel element will occur (Rinne et al., 2001), rather the auditory system automatically orients attention towards it. Here, we highlighted another ability of the system. Beyond automatic orientation toward a relevant deviant sound, our results show that it is possible to categorize the acoustic change according to the representation (detailed or summary) and temporal resolution (high or low) at which it has occurred. Importantly, discriminative neural responses could be detected even if the task *per se* did not involve any discrimination or in-depth processing of either local features or summary statistics. The sound changes were processed even when irrelevant to the task participants were attending to (rare beep detection), strongly suggesting that such processing occurs not only automatically but also pre-attentively. Furthermore, the double dissociation we observed based on sound duration (with Local Features eliciting greater magnitude change than Summary Statistics for short sounds and vice-versa for long sounds) rules out the possibility of results being explained by a mere saliency effect (i.e., the fact that, in Summary Statistics, a different sound object was presented).

Importantly, results emerged despite the fact that sound objects between the triplets were continuously changing (the only fixed parameter was the variability in local features or summary statistics between novel and repeated sounds) and thus can be generalized to a variety of sound textures (Figure 1C; see also Supplementary Information, Table S1). The exact moment in which the summary percepts emerge likely depends on the specific comparisons across sound-objects (repeated and novel). In line with this, the use of many different sounds for the creation of sound streams led to grand-averaged signals associated with discriminations based on summary statistics with a rather spread-out shape (see Figure 2B, right).

Local features changes are encoded by fast oscillations

By comparing the difference in total power between novel and repeated sounds in the two experiments we found that, for short sounds, the power between 20 and 28Hz decreased when a change in local features was detected, as compared to when summary statistics were changed. This desynchronization occurred between 80 and 200ms after stimulus onset (Figure 3A, B, left).

The 20-28Hz band includes frequencies that are canonically attributed to high-beta oscillations. Previous studies correlated power synchronization at such frequency

rates with performance in tasks involving the detection of temporal or intensity deviations (Arnal et al., 2015; Herrmann et al., 2016). This evidence suggests that, among other operations, brain activity in the high-beta range could be engaged in the processing of low-level properties of a stimulus.

In the auditory domain, beta-band activity has been investigated in several instances, especially in the context of rhythmic perception. A disruption in beta power can be observed in non-rhythmic sequences or when an attended tone is omitted from a regular series (e.g., Fujioka et al., 2012). Interestingly, beta synchronization not only captures irregularities in a pattern but also reflects the type of change that has occurred. For instance, it has been shown that beta desynchronization was higher prior to the occurrence of a deviant sound whose pitch varied in a predictable way, as compared to an unpredictable variation. Accordingly, beta desynchronization has been proposed as a marker of predictive coding (Engel and Fries, 2010; Chang, Bosnyak, and Trainor, 2018).

In our model, stimuli could be derived from the same white-noise sample or a different one (Figure 1A, B). In Local Features, the novel sound is derived from another white-noise sample, as compared to the repeated sound, on which we imposed the same summary statistics. Thus, with this synthesis approach, in terms of acoustic fine details, when sounds were short, novel sounds were more different than the repeated one in the Local Features experiment as compared to Summary Statistics (Figure 2A, left side; see also Figure S1A for method details). Overall, these results suggest that, in the absence of enough information to build summary representation, faster oscillations are in charge of small, acoustic change-detection to be used to discriminate sound excerpts.

Slower oscillations are engaged in Summary Statistics processing

By comparing Local Features with Summary Statistics at long durations, we observed that the emergence of different auditory statistics in the novel sound, as compared to the previous, repeated one, elicited higher power at slower frequencies, compatible with canonical alpha-theta oscillations. This power synchronization emerged at late latencies from stimulus onset (between 240 and 500ms; Figure 3A, B, right). When solely local features were driving sound change (as in Local Features), this power synchronization was not present.

The involvement of relatively slow oscillations for processing auditory statistics, especially those derived from envelope transformations (Figure S1A, S1B), makes sense considering previous evidence on amplitude modulation processing. For example, envelope detection reaches its greatest sensitivity at 4Hz (Viemeister, 1979). Interestingly, several studies have shown that the auditory system groups information within an integration window about 150-300ms long, roughly corresponding to a full cycle in the theta band (Ghitza & Greenberg, 2009; Ghitza, 2012). A recent study showed that acoustic changes occurring within such a temporal window could explain the modulations of phase synchronization in theta band (Teng et al., 2018). The general idea is that brain activity processes sounds through an active chunking mechanism, which condenses entering acoustic information within a temporal window (~ 200ms), in accordance with ongoing oscillatory cycles (VanRullen, 2016; Riecke, Sacks, & Schroeder., 2015; Teng et al., 2018). In the same spirit, the higher power in the theta-alpha range observed in our study (approximately 240ms after stimulus onset) could reflect the integration of local features into summary envelope statistics. The different statistical representation leads to a higher-power synchronization which is not present in Local Features, because in the latter case, after the chunking period, the novel representation matches the previous one.

Overall, these results support findings revealing that summary representations are built after parsing a continuous sound into chunks of approximately 200ms length (Poeppel, 2003; Ghitza & Greenberg, 2009; Panzeri et al., 2010; Ghitza, 2012; Giraud & Poeppel, 2012; VanRullen, 2016; Teng et al., 2018). This mechanism may be a prerequisite for the recognition of sound identity, leading to an increased synchronization when a novel set of summary statistics, pointing towards a different sound source, is presented.

CONCLUSION

By combining a computational-synthesis approach with electrophysiology, we revealed distinct cortical representations associated with detailed and summary representations. We showed that different neural codes, at faster and slower temporal scales, are entrained to automatically – and possibly pre-attentively – detect changes in entering sounds, based on these two auditory modes of

representation. These results promote the usage of computational methods to appoint neural markers of temporal discrimination and for studying basic auditory computation in both fundamental and applied research. Furthermore, the automaticity of the protocol and the fast implementation allow the testing of different populations (including newborns, infants, children, and clinical patients) that do not have the resources to attend to complex tasks.

ACKNOWLEDGMENTS

We thank all the students who helped with recruiting participants and/or data collection: Nicolò Castellani, Irene Sanchez, Chiara Battaglini, and Dila Suay. Funding: Davide Bottari (PRIN 2017 research grant. Prot. 20177894ZH).

CONFLICT OF INTEREST

The authors declare no competing interest.

SUPPLEMENTARY INFORMATION

Sample size

The number of participants was estimated via simulations. We used the procedure described in Wang and Zhang (2021) and simulated a dataset with two conditions (Local Features and Summary Statistics) of Auditory Evoked Potentials data. First, we selected three electrodes of interest at central locations (E7, E65, E54) that typically capture auditory responses. For the simulation, we selected a time window between 0.1 and 0.3s, based on previous MMN studies (see Näätänen et al., 2007 for review). The amplitude values at the electrodes of interest for the two conditions were sampled from a bivariate normal distribution (within-subject design), in which mean and repeated deviation were chosen based on results of four pilot datasets (mean Local Features= 0.16; mean Summary Statistics= 0.56; sd Local Features= 0.52; sd Summary Statistics= 0.54).

We then ran a cluster-based permutation on simulated datasets to test whether any statistical cluster (t-values) exhibited significant difference between the two

conditions with an alpha level of 0.05. The procedure started with a sample size of 10 and increased in steps of 1 until it reached a power of 0.80. We ran 1000 simulations for each sample size and calculated the power as the proportion of the number of times significant clusters were found in these 1000 simulations. The simulation results showed that, in order to obtain a power above 0.8, a sample size of N=24 was required (see Figure S1C).

The algorithm to perform such analyses can be downloaded from this link: https://osf.io/rmqhc/

Behavioral Results

For each condition, percentage of correct beep detections was above 90% (Local Features 40: mean= 0.99, std= 0.03; Local Features 209: mean= 0.99, std= 0.05; Local Features 478: mean= 1, std= 0; Summary Statistics 40: mean=0.99, std= 0.05; Summary Statistics 209: mean= 0.97, std=0.08; Summary Statistics 478: mean= 0.97, std= 0.11; Figure S1D). We ran a two-way ANOVA for repeated measures with factors Experiment (2 levels, Local Features vs. Summary Statistics) and Duration (3 levels, 40, 209, and 478) to address whether experiment type and stimulus length had any impact on beep detection and participant attention to the task. No significant main effects were observed (Experiment, $F_{(1,23)}= 3.62$, $p =0.07$, $n^2= 0.14$; Duration, $F_{(2,46)}= 0.58$, $p= 0.56$, $n^2= 0.3$), nor their interaction (Experiment*Duration, $F_{(2,46)}= 0.45$, $p= 0.64$, $n^2= 0.2$).

These behavioral results provide evidence that participants were attentive and responsive during sound presentation throughout the experiment and that attention to this orthogonal task was not influenced by duration of either sound or experiment.

REFERENCES

McAdams, S. (1993). Recognition of sound sources and events. Thinking in sound: The cognitive psychology of human audition, 146-198.

Griffiths, T. D. (2001). The neural processing of complex sounds. Annals of the New York Academy of Sciences, 930(1), 133-14

Plomp, R. (1964). Rate of decay of auditory sensation. The Journal of the Acoustical Society of America, 36(2), 277-282.

718 McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory
719 perception. Nature neuroscience, 16(4), 493-498.
720
721 McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the
722 auditory periphery: evidence from sound synthesis. Neuron, 71(5), 926-940.
723 Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude
724 modulation. I. Detection and masking with narrow-band carriers. The Journal of the
725 Acoustical Society of America, 102(5), 2892-2905.
726
727 Yabe, H., Tervaniemi, M., Sinkkonen, J., Huotilainen, M., Ilmoniemi, R. J., & Näätänen, R.
728 (1998). Temporal window of integration of auditory information in the human brain.
729 Psychophysiology, 35(5), 615-619.
730
731 Saint-Arnaud, N., Popat, K. (2021). Analysis and synthesis of sound textures
732 D.F. Rosenthal, H.G. Okuno (Eds.), Computational Auditory Scene Analysis, CRC Press
733 (2021), pp. 293-308
734
735 Giraud, A. L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., &
736 Kleinschmidt, A. (2000). Representation of the temporal envelope of sounds in the human
737 brain. *Journal of neurophysiology*, *84*(3), 1588-1598.
738
739 Lorenzi, C., Berthommier, F., Apoux, F., & Bacri, N. (1999). Effects of envelope expansion
740 on speech recognition. *Hearing research*, *136*(1-2), 131-138.
741
742 Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages
743 Sensory Communication ed WA Rosenblith. Cambridge, MA: MIT Press), 7, 1-34.
744
745
746 Berto, M., Ricciardi, E., Pietrini, P., & Bottari, D. (2021). Interactions between auditory
747 statistics processing and visual experience emerge only in late development. Iscience,
748 24(11), 103383.
749
750 Zhai, X., Khatami, F., Sadeghi, M., He, F., Read, H. L., Stevenson, I. H., & Escabí, M. A.
751 (2020). Distinct neural ensemble response statistics are associated with recognition and
752 discrimination of natural sound textures. Proceedings of the National Academy of Sciences,
753 117(49), 31482-31493.
754
755 Treisman, A., Vieira, A., & Hayes, A. (1992). Automaticity and preattentive processing. *The*
756 *American journal of psychology, 341-362*
757
758 Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging
759 computational principles and operations. *Nature neuroscience*, *15*(4), 511-517.
760
761 Panzeri, S., Brunel, N., Logothetis, N. K., & Kayser, C. (2010). Sensory neural codes using
762 multiplexed temporal scales. *Trends in neurosciences*, *33*(3), 111-120.
763
764 Wang, C., & Zhang, Q. (2021). Word frequency effect in written production: Evidence from
765 ERPs and neural oscillations. Psychophysiology, 58(5), e13775.
766
767 World Medical Association. (2013). World Medical Association Declaration of Helsinki:
768 ethical principles for medical research involving human subjects. Jama, 310(20), 2191-2194.
769
770 Bacon, S. P., & Grantham, D. W. (1989). Modulation masking: Effects of modulation
771 frequency, depth, and phase. *The Journal of the Acoustical Society of America*, *85*(6), 2575-
772 2580.

Sussman, E. S., & Gumenyuk, V. (2005). Organization of sequential sounds in auditory memory. Neuroreport, 16(13), 1519-1523.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. Spatial vision, 10(4), 433-436.

Stropahl, M., Bauer, A. K. R., Debener, S., & Bleichner, M. G. (2018). Source-Modeling auditory processes of EEG data using EEGLAB and brainstorm. Frontiers in neuroscience, 12, 309.

Bottari, D., Bednaya, E., Dormal, G., Villwock, A., Dzhelyova, M., Grin, K., … & Röder, B. (2020). EEG frequency-tagging demonstrates increased left hemispheric involvement and crossmodal plasticity for face processing in congenitally deaf signers. NeuroImage, 223, 117315.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 134(1), 9-21.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. Neural computation, 7(6), 1129-1159.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., Mckeown, M. J., Iragui, V., & Sejnowski, T. J. (2000a). Removing electroencephalographic artifacts by blind source separation. Psychophysiology, 37(2), 163-178.

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2000b). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. Clinical Neurophysiology, 111(10), 1745-1758.
Tonachini et al., 2019

Artoni, F., Delorme, A., & Makeig, S. (2018). Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition. NeuroImage, 175, 176-187.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. NeuroImage, 198, 181-197.

Viola, F. C., Thorne, J., Edmonds, B., Schneider, T., Eichele, T., & Debener, S. (2009). Semi-automatic identification of independent components representing EEG artifact. Clinical Neurophysiology, 120(5), 868-877.

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.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. Journal of neuroscience methods, 164(1), 177-190.

Gourévitch, B., Martin, C., Postal, O., & Eggermont, J. J. (2020). Oscillations in the auditory system and their possible role. *Neuroscience & Biobehavioral Reviews*, *113*, 507-528.

VanRullen, R. (2016). Perceptual cycles. *Trends in cognitive sciences*, *20*(10), 723-735.

Teng, X., Tian, X., Doelling, K., & Poeppel, D. (2018). Theta band oscillations reflect more than entrainment: behavioral and neural evidence demonstrates an active chunking process. *European Journal of Neuroscience*, *48*(8), 2770-2782.

Fujioka, T., Trainor, L. J., Large, E. W., & Ross, B. (2012). Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *Journal of Neuroscience*, *32*(5), 1791-1802.

Snyder, J. S., & Large, E. W. (2005). Gamma-band activity reflects the metric structure of rhythmic tone sequences. *Cognitive brain research*, *24*(1), 117-126.

Cohen, M. X., & Donner, T. H. (2013). Midfrontal conflict-related theta-band power reflects neural oscillations that predict behavior. Journal of neurophysiology, 110(12), 2752-2763.

Cohen, M. X., & Cavanagh, J. F. (2011). Single-trial regression elucidates the role of prefrontal theta oscillations in response conflict. *Frontiers in psychology*, *2*, 30.

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. Acta psychologica, 42(4), 313-329.;

Tiitinen, H., May, P., Reinikainen, K., & Näätänen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. Nature, 372(6501), 90-92.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., & Winkler, I. (2001). 'Primitive intelligence'in the auditory cortex. Trends in neurosciences, 24(5), 283-288.

Näätänen, R., Astikainen, P., Ruusuvirta, T., & Huotilainen, M. (2010). Automatic auditory intelligence: An expression of the sensory–cognitive core of cognitive processes. Brain research reviews, 64(1), 123-136.

Rinne, T., Antila, S., & Winkler, I. (2001). Mismatch negativity is unaffected by top-down predictive information. NeuroReport, 12(10), 2209-2213.

Arnal, L. H., Doelling, K. B., & Poeppel, D. (2015). Delta–beta coupled oscillations underlie temporal prediction accuracy. *Cerebral Cortex*, *25*(9), 3077-3085.

Herrmann, C. S., Strüber, D., Helfrich, R. F., & Engel, A. K. (2016). EEG oscillations: from correlation to causality. *International Journal of Psychophysiology*, *103*, 12-21.

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—signalling the status quo?. *Current opinion in neurobiology, 20(2), 156-165.*

Chang, A., Bosnyak, D. J., & Trainor, L. J. (2018). Beta oscillatory power modulation reflects the predictability of pitch change. *Cortex, 106, 248-260.*

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America*, *66*(5), 1364-1380.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113-126.

881 Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech:
882 intelligibility of speech with a manipulated modulation spectrum. *Frontiers in psychology*, *3*,
883 238.
884
885 Riecke, L., Sack, A. T., & Schroeder, C. E. (2015). Endogenous delta/theta sound-brain
886 phase entrainment accelerates the buildup of auditory streaming. *Current Biology*, *25*(24),
887 3196-3201.
888
889 Poeppel, D. (2003). The analysis of speech in different temporal integration windows:
890 cerebral lateralization as 'asymmetric sampling in time'. *Speech communication*, *41*(1), 245-
891 255.
892