NEUROSCIENCE
RESEARCH ARTICLE

E. M. Kaya et al. / Neuroscience 440 (2020) 1–14

Pitch, Timbre and Intensity Interdependently Modulate Neural Responses to Salient Sounds

Emine Merve Kaya, Nicolas Huang and Mounya Elhilali *

Laboratory for Computational Audio Perception, Department of Electrical and Computer Engineering Johns Hopkins University, Baltimore, MD, USA

Abstract—As we listen to everyday sounds, auditory perception is heavily shaped by interactions between acoustic attributes such as pitch, timbre and intensity; though it is not clear how such interactions affect judgments of acoustic salience in dynamic soundscapes. Salience perception is believed to rely on an internal brain model that tracks the evolution of acoustic characteristics of a scene and flags events that do not fit this model as salient. The current study explores how the interdependency between attributes of dynamic scenes affects the neural representation of this internal model and shapes encoding of salient events. Specifically, the study examines how deviations along combinations of acoustic attributes interact to modulate brain responses, and subsequently guide perception of certain sound events as salient given their context. Human volunteers have their attention focused on a visual task and ignore acoustic melodies playing in the background while their brain activity using electroencephalography is recorded. Ambient sounds consist of musical melodies with probabilistically-varying acoustic attributes. Salient notes embedded in these scenes deviate from the melody's statistical distribution along pitch, timbre and/or intensity. Recordings of brain responses to salient notes reveal that neural power in response to the melodic rhythm as well as cross-trial phase alignment in the theta band are modulated by degree of salience of the notes, estimated across all acoustic attributes given their probabilistic context. These neural nonlinear effects across attributes strongly parallel behavioral nonlinear interactions observed in perceptual judgments of auditory salience using similar dynamic melodies; suggesting a neural underpinning of nonlinear interactions that underlie salience perception.

Keywords: Auditory salience, Electroencephalography, Nonlinear interaction, Pitch, Timbre, Intensity.

INTRODUCTION

As everyday acoustic environments challenge the auditory system with a deluge of sensory information, the brain has to selectively focus its limited resources on the most relevant events that are crucial both for survival and awareness of our changing surrounds. Salience is an attribute of events that inherently reflects their perceptual relevance and as such guides exogenous attention to important information in the stimulus (Itti and Koch, 2001). The siren of an emergency vehicle or an offbeat segment in an orchestral piece are events that undoubtedly stand out perceptually and attract our attention to unique moments in the acoustic scene. What makes these events salient is the fact that they differ relative from other sounds in the scene, hence deviating from an internal model of the sensory world. This interpretation posits that the brain maintains an internal representation of the stimulus which is used to reconcile with incoming information causing any deviation to pop-out (Friston, 2010). This *contrast* principle is at the core of current theories of how salience computation operates in the brain, and appears to apply with a great deal of accuracy across different sensory modalities (Moskowitz and Gerbers, 1974; Wolfe and Horowitz, 2004; Kayser et al., 2005; Walsh et al., 2016).

In audition, the contrast theory incorporates the dynamic nature of sound as an inherent component of salience (Kaya and Elhilali, 2014). Specifically, the auditory system builds on its deviance detection mechanisms to flag presence of oddball elements in the stimulus hence guiding the brain's response to unexpected events. When listening to our acoustic environment, the auditory system infers patterns in sound sources that help build an internal model of the world (Winkler et al., 2009; Huang and Elhilali, 2017). This model extracts coherent regularities in the sensory space by leveraging Gestalt structures in the stimulus that shape internal representations of putative sound objects in the scene (Bregman, 1990). These primitive regularities are encoded in the auditory system with increasingly adaptive representations wherein faithful encoding of acoustic features at the peripheral level evolve to progressively become more sensitive to deviant

*Corresponding author.
E-mail address: mounya@jhu.edu (M. Elhilali).

https://doi.org/10.1016/j.neuroscience.2020.05.018
0306-4522/© 2020 The Author(s). Published by Elsevier Ltd on behalf of IBRO.

patterns (Ulanovsky, 2004; Antunes and Malmierca, 2014; Nelken, 2014).

Understanding underpinnings of auditory salience is intimately tied to defining the structure of the acoustic space over which this internal model of regularities is developed. Building on our understanding of sensory processing along the auditory pathway, we know that sounds undergo a series of transformation from the periphery to the central auditory system, wherein features such as spectral content, spectral shape profile (e.g. bandwidth) and temporal dynamics are extracted along separate feature maps (Schreiner, 1992; Schreiner, 1995; Versnel et al., 1995; Kowalski et al., 1996). The rich tuning of cortical neurons suggests that acoustic stimuli are mapped onto a high-dimensional space over which the structures guiding our internal model can be built (Chakrabarty and Elhilali, 2019). Cortical representations reflect not only the underlying feature characteristics of the stimulus, but also complex interactions that guide coherent perception of integrated objects. Nonlinear interactions across acoustic features are abound in cortical data and suggest synergistic integration across features (Atencio et al., 2008; Bizley et al., 2009; Sloas et al., 2016). How these nonlinear interactions manifest themselves to shape our perception of salient events is unclear, especially that salience is guided both by attributes of a sound event as well as its context (Huang and Elhilali, 2017).

In a previous behavioral report, manipulating deviance of sound tokens along multiple acoustic attributes revealed strong nonlinear interactions that guide judgments of salience elicited with dynamic scenes (Kaya and Elhilali, 2014). In this Kaya study, listeners were presented with different complex stimuli including speech, musical melodies or nature sounds where deviance of an embedded token was manipulated along a collective space spanning pitch, timbre and intensity. Regardless of sound category, an interaction effect was reported across these acoustic attributes consistently for speech, music and nature sounds suggesting an underlying interdependent representation of auditory feature maps that not only guides their encoding in the brain, but also modulates their contrast against an internal model of the scenes to guide judgments of salience of specific events.

Interdependent effects of acoustic attributes guide our perception in a variety of tasks. Earlier reports of interaction between attributes such as pitch, timbre and intensity have revealed profound nonlinear interdependencies that shape judgments of detection, discrimination or even sound classification (Moore and Hearing, 1995; Allen et al., 2013). Melara and Marks have argued of an interpretation based on an ‘‘interactive multichannel model” of auditory processing (Melara and Marks, 1990); though functional imaging data suggests no clear anatomical distinctions between cortical networks engaged in building internal models of the stimulus along different acoustic channels (Allen et al., 2017).

In order to further probe the underpinnings of these interactions in the context of a salience paradigm, we record electrophysiological responses from human listeners presented with dynamic musical melodies whose acoustic structure is controlled by a statistical distribution along various dimensions spanning pitch, timbre and sound intensity. Occasional salient notes violate the statistical distribution of the melody along one or many of these dimensions, hence diverging from the internal model elicited by the melody itself. Theories based on deviance detection suggests that shifts of the statistical structure of the input will likely elicit changes in neural responses. One of the questions explored in the current work is how the joint manipulation of multiple acoustic dimensions manifests itself in these cortical responses, and to what extent do observed interdependencies in behavioral judgments arise from such underlying neural responses (Kaya and Elhilali, 2014)? The current work also examines what aspects of neural responses are most modulated by such interactions in response to salient sounds. Given the tight link between salience perception and exogenous attention, it is an open question how such form of attention manifests itself and how its markers relate to well-known effects of endogenous or top-down attention on brain responses to complex sounds.

Recent work on top-down attention using natural speech, animal vocalizations, and ambient sounds demonstrates that neural activity fluctuates in a pattern matching that of the attended stimulus, driving the power of oscillations at the stimulus rate or low-frequency oscillations and modulating this power by attention (Lakatos et al., 2008; Besle et al., 2011; Zion Golumbic et al., 2013; Jaeger et al., 2018). This enhanced representation of the attended stimulus has been used to track auditory attention using envelope decoding paradigms (O'Sullivan et al., 2015; Fuglsang et al., 2017), also see review (Wong et al., 2018; Alickovic et al., 2019). While these studies have successfully extracted stimulus-specific information from neural recordings to natural continuous sound environments, they have all employed experimental paradigms under the influence of top-down attention. In the current study, we explore whether exogenous attention reveals parallel responses in terms of changes in fidelity of encoding or oscillatory activity in response to salient sound tokens. By employing dynamic scenes that manipulate salience along different attributes, we can specifically probe how neural markers of salience are influenced by specific acoustic dimensions. Subjects attention is directed away from the sounds and engaged in a demanding visual task, hence controlling their attentional focus away from the acoustic scene except for occasional salient tokens that attract their attention exogenously.

EXPERIMENTAL PROCEDURES

Participants

Thirteen subjects (7 female) with normal vision and hearing and no history of neurological disorders participated in the experiment in a soundproof booth after giving informed consent and were compensated for their participation. All procedures were approved by the Johns Hopkins Institutional Review Board.

Stimuli and experimental paradigm

Subjects performed an active visual task while auditory stimuli were concurrently presented, and subjects were instructed to ignore the sounds. *Auditory stimuli* closely followed the design used in (Kaya and Elhilali, 2014). Each sound stimulus was 5 s long and consisted of regularly spaced, temporally overlapping musical notes each 1.2 s long, with a new note starting every 300 ms. Individual notes were extracted from the RWC Musical Instrument Sound Database (Goto et al., 2003) for 3 instruments: Pianoforte (Normal, Mezzo), Acoustic Guitar (Al Aire, Mezzo) and Clavichord (Normal, Forte) at 44.1 kHz; and were amplitude normalized relative to their peak with 0.1 s onset and offset sinusoidal ramps. The repetition rate and instruments were specifically selected to sound pleasing and flow naturally in order to resemble musical melodies. The 3 instruments were chosen to balance a number of considerations: high contrast in timbre along factors such as spectral flux, irregularity and temporal attack, as reported in earlier timbre studies (McAdams et al., 1995). The temporal envelope of these 3 instruments allowed a better control of intensity relative to amplitude peak because the instruments contained a sufficiently prominent steady-state activity.

Notes in each 5 s sequence were played by the same instrument (denoted Timbre-bg or $\mathbf{T}_b$, i.e. timbre of the background scene), controlled an average intensity at a comfortable hearing level, and maintained a pitch within ±2 semitones of a nominal pitch value around A3 (220 Hz). In "test" trials, one note (selected at random in the middle of the melody anywhere between 2.4 s and 3.8 s from onset of the melody) was chosen as "salient", and had acoustic attributes that differed from the melody in that trial: different timbre (new instrument, denoted Timbre-sal or $\mathbf{T}_s$), higher pitch **P** (2 or 6 semitones higher) and higher intensity **I** (2 or 6 dB higher) relative to the background scene. Salient notes were manipulated in a factorial design to test *all* combination of variations along all 3 acoustic dimensions (pitch, timbre and intensity). Due to the difficulty of defining timbre on a scale, we characterized timbre differences categorically by testing all 9 combinations of the 3 instruments for melody notes (Timbre-bg) and salient notes (Timbre-sal). This resulted in $3 * 3 * 2 * 2 = 36$ trials to test every possible feature deviation (i.e. 3 background timbres or instruments $\mathbf{T}_b$, 3 salient note instruments $\mathbf{T}_s$, 2 pitch deviations **P** and 2 intensity deviations **I**). Each feature deviation was repeated 10 times with different dynamic backgrounds and salient onset times, for a total of 360 salient trials.

In addition, "control" trials were constructed in a similar fashion, but without any salient notes. The attributes (pitch, timbre, intensity) of notes in control trials were carefully chosen to embed each of the salient notes without making it salient given its context. For example, if a clavichord note was previously presented as salient in a melody of guitar notes, that same clavichord note with the same intensity and pitch was now embedded in a clavichord "control" melody with overlapping range of intensity and pitch values making this same note non-salient in the context of control trials. In each control trial, this specific note was controlled to appear at two randomly selected positions no earlier than 2.4 s from the start of the trial, with a minimum of 900 ms between the two occurrences. Five control trials were constructed for each of the 12 salient notes, resulting in 60 control trials and 420 experiment trials in total. The order of trials was randomized for each subject.

Visual stimuli consisted of digits and capital letters presented on a black screen where subjects were instructed to report digits. Each target was uniformly chosen at random from the numbers 0–9, while each non-target was uniformly chosen at random from the letters A-Z. Subjects were instructed to enter any numbers they observed after each trial in the order of appearance, using a numeric keypad. The next trial was initiated by the subject after entering their response up. Within each trial, 56 characters were presented in sequence, with one presented every 90 ms. The visual task was divided into two difficulty conditions. The low-load condition consisted of white numbers in contrast to gray letters, with all characters remaining on-screen for 90 ms. In the high-load condition, both targets and non-targets were the same shade of gray, and all characters were presented for only 20 ms. Trials were presented in 12 blocks, with blocks alternating between the two load conditions. Presentation order of low and high-load conditions was counter-balanced across subjects.

In most visual trials, two targets were presented at random points in the trial. To avoid confounds with salient events in the sound stream, one target ("early") occurred within the first 2.4 s of the trial, while the other ("late") occurred after 4.2 s. The first and last two characters were always non-targets. Target positions were uniformly chosen at random within their respective ranges. In 20% of trials, only one target was presented, with an equal chance of it being either early or late. Finally, for 30 trials, one of the visual targets was positioned between 2.4 and 4.2 s from the start of the trial, while still being at least 1.3 s away from any salient auditory stimuli. This adjustment ensured that subjects paid attention to the visual stimulus throughout each trial.

Neural data acquisition

Electroencephalography (EEG) recordings were performed with a Biosemi Active Two system with 128 electrodes, plus left and right mastoids acquired at 2048 Hz. Four additional electrodes recorded eye and facial artifacts from EOG (electrooculography) locations, and a final electrode was placed on the nose to serve as reference. The nose electrode was used only to examine ERP components, particularly mismatch negativity at mastoids (Mahajan et al., 2017). The average mastoid reference was used for all further analyses.

Initial processing of signals was performed with the Fieldtrip software package for MATLAB (Oostenveld et al., 2011). Trials were epoched with 0.5 s of buffer time before and after each trial capturing fixation segments, referenced to the average of the left and right mastoids, downsampled to 256 Hz, and filtered between 0.5–100 Hz. To remove muscle and eye artifacts from the signals, we used independent component analysis (ICA) as implemented by FieldTrip. ICA components were removed if their amplitude was greater than the mean plus 4 standard deviations for more than 5 trials. The resulting filtered signals were visually confirmed to be free of prominent eye blinks and large amplitude deviants.

Neural data analysis

The stimulus paradigm presented the same physical note as salient (in the context of ''test'' trials) or as control (in the context of ''control'' trials). All neural data analyses compared salient notes to control notes (same note when non-salient), and analyses were divided by salience level for each feature tested (pitch, timbre, intensity).

Neural power analysis: Time–frequency power analysis of each experimental trial was computed with the matching pursuit algorithm using a discrete cosine transform dictionary (Mallat and Zhang, 1993). For precise spectral resolution, neural responses from salient notes and matching control notes were extracted in segments spanning two notes (i.e. 0.6 s post note onset). Extracted segments were concatenated across trials, and the power of the Discrete Fourier Transform (DFT) was obtained at each frequency bin of interest. Concatenating the signals was necessary to increase the spectral resolution of the frequency analysis. While this process could create an edge effect at exactly 1.67 Hz (1/0.6 s) and possibly its integer multiples, post hoc analyses and visual inspection confirmed that no significant artifacts resulted from the concatenation procedure. Furthermore, the same potential effects would affect salient and control trials equally.

The power of the Discrete Fourier Transform (DFT) of the signal at the sample closest to 3.33 Hz (1/0.3 s) was divided by the average power at 2.33–4.33 Hz, with the power at 3.33 Hz excluded. The neural power of salient notes at the stimulus rhythm was defined as the normalized power at 3.33 Hz averaged over the top 15 channels with the strongest response. The power at other adjacent and further frequencies (3.2, 3.4, 6, 12, 20, 30, 38, 40) was also obtained from the same spectrum. Channels were allowed to vary between subjects to allow for inter-participant variability, following the procedure used in (Elhilali et al., 2009). The same analysis was performed for salient notes as well as identical control notes. Qualitatively similar results were obtained by including a larger number of channels, though the noise floor increased as well.

Phase-coherence: The neural response to each test and control trial was decomposed into multiple narrowband signals by spectrally filtering responses of each channel individually along the following bands '*B*': Delta 1–3 Hz, Theta 4–8 Hz, Alpha 9–15 Hz, Beta 16–30 Hz, Gamma 31–100 Hz. The instantaneous phase of the Hilbert transform was then obtained for each *B*-narrowband signal at trial *i*, yielding the quantity $\{\theta_i^B(t)\}$ (King, 2009). Signal segments corresponding to salient notes (note onset-300 ms) and reciprocal control notes were obtained, and any segments near melody onset (0–2.4sec) and offset (0.8 - end) were excluded to avoid narrowband filter boundary effects. The phase-coherence across trials $\{c_\theta^B\}$ was computed for each frequency band and each segment separately. This inter-trial coherence quantity is a measure of alignment *in phase* across responses to the same note across many repetitions (trials), for a given spectral band *B*, integrated over time *t*. It is defined as:

$$c^B = \sum_{i=1}^{N} \int e^{j*\theta_i^B(t)} dt$$

which quantifies the magnitude of the average instantaneous phase, integrated over time and averaged across all trials.

ERP analysis: EEG trials were bandpass filtered between 0.7 and 25 Hz. Responses from frontal electrodes (Fz and 21 surrounding electrodes) and central electrodes (Cz and 23 surrounding electrodes) were analyzed (Shuai and Elhilali, 2014). Segments corresponding to salient notes and corresponding control notes were extracted separately for each channel. First, difference waveforms at the left mastoid, right mastoid, and Fz channels were analyzed. These channels were selected based on the MMN literature to confirm maximum MMN amplitude at Fz and polarity reversal at the mastoids (Schroger, 1998). Significant negative peaks were confirmed for all subjects at Fz by paired t-tests comparing the MMN time window point-by-point to 0, polarity reversals at the mastoids were confirmed visually. Next, trials were re-referenced to the average mastoids, and baseline corrected using the 100msec prior to each trial. Difference waveforms were re-computed for all subjects across central and frontal electrodes (though qualitatively similar results were obtained for individual subjects, albeit at a lower SNR). For each average waveform, peaks were extracted over various windows of interest: P1 (positive peak) at 25–75 ms, N1 (negative peak) at 75–120 ms, MMN (negative peak) at 120–180 ms, P3a (positive peak) at 225–275 ms.

Spectrotemporal receptive fields: The cortical activity giving rise to EEG signals was modeled by spectrotemporal receptive fields (*STRF*). This function infers a processing filter that acts on a transformation of the auditory stimulus along time and frequency. Specifically, the neural response $r(t)$ is modeled as a result of processing the time–frequency spectrogram of the stimulus $S(f,t)$ by this kernel *STRF* then integrated across time lags and frequencies, plus any additional background cortical activity and noise denoted by $\epsilon(t)$. The *STRF* model is then described as:

$$r(t) = \sum_f \int STRF(f,\tau)s(f,t-\tau)d\tau + \epsilon(t).$$

Estimation of the *STRF* was performed by boosting (David et al., 2007), implemented by a simple iterative algorithm that converges to an unbiased estimate. A brief description of the algorithm is as follows. The *STRF* (size $F \times T$) was initialized to zero, and a small step size was defined as δ. For each time–frequency point in the *STRF* (every element in the matrix), the *STRF* was incremented by δ and $-\delta$ giving a pool of $F * T * 2$ possible *STRF* increments. The increment that provided the smallest mean-squared error was selected for the current iteration. This process was repeated until none of the *STRF*s in the possible increment pool improve the mean-squared error. Next the step size was reduced to $\delta/2$ and continuing the same process, with 4 step size reductions in total.

*STRF*s were estimated for salient and control segments separately and were defined for a 300 ms window, reflecting the frequency of new notes. Twofold cross validation was used to validate *STRF*s during estimation: Trials were divided into two groups with equal number of factorial repetitions in each group. A *STRF* was estimated for one group, and used to obtain an estimated neural response for the other group which is then correlated with the actual response. *STRF*s with a correlation of less than 0.05 were eliminated to remove estimates with low predictive power, and the remaining *STRF*s were averaged as the final STRF estimate for that condition. Using higher fold estimates did not give significantly different results for the overall case across all salient notes. The number of folds was limited to two given the limited number of trials that allowed an analysis of STRFs for each salient feature category (pitch, intensity, timbre). Feature STRFs were analyzed by using the salient segments that corresponded to each level of the feature at hand in separate estimations. All STRFs were averaged over data in channel Fz (C21 on the Biosemi map) and 4 surrounding channels (C22, C20, C12, C25).

Statistical analysis

The cross-factorial experimental design allowed an analysis of individual features (Pitch, Intensity, Timbre-sal, and Timbre-bg) as well as combined features using within-subject ANOVAs. All results were corrected for multiple comparisons using Holm-Bonferroni correction to confirm statistical significance (Snedecor and Cochran, 1989). Results post correction are reported here. Residuals were checked for normality using the Shapiro–Wilk test ($p < 0.05$), as well as visual inspection of QQ plots) and Mauchlys test of Sphericity was used to check for sphericity ($p > 0.05$).

RESULTS

Subjects performed a rapid serial visual presentation (RSVP) task by identifying numbers within a sequence of characters (Haroush et al., 2010) (Fig. 1A-top). Participants were closely engaged in this task (overall target detection accuracy 70.3 ± 6.9%, and their performance was modulated by task difficulty (accuracy 75.2% for the low-load and 65.4% for the high-load task).

Concurrently, sound melodies were played diotically in the background and subjects were asked to completely ignore them. Acoustic stimuli consisted of sequences of musical notes with temporal regularity. Fig. 1A shows a schematic of such melody composed of violin notes with varying intensities and pitches and an unexpected salient piano note. This depiction is an example of a "test" trial, which included an occasional salient note that did not fit the statistical structure of the melody (e.g. piano among violins). Salient notes varied along pitch, timbre and intensity in a crossed-factorial design (Fisher, 1935). In contrast, "control" trials were melodies from the same instrument whose notes also statistically varied along pitch and intensity but did not deviate from a constrained distribution, hence containing no pop-out notes. The same musical notes that were salient in "test" trials were also embedded in "control" trials; but the statistical distribution of these "control" trials was manipulated to make these notes non-salient. The same piano note in Fig. 1A would be salient in a "test" trial among violins; but would not be salient in "control" trial among other piano notes of similar pitch and intensity. Employing the exact physical note when salient vs. control allows to factor out any effects of the exact acoustic attributes of the note itself. All acoustic stimuli (test and control trials) consisted of melodies with temporally overlapping notes, though the entire tune had an temporal regularity with a period of 300 ms (Fig. 1B).

While visual targets and auditory salient notes were not aligned in the experimental design, we probed distraction effects due to the presence of occasional salient notes in the ignored melodies. Visual trials coinciding with "test" trials contained a subset of targets that occurred prior to salient notes while others occurred after the salient note. When comparing effect of salient and control melodies on visual target detection, there was no notable difference in detection accuracy for targets occurring prior to salient notes (unpaired t-test, t (13) = 0.86, p = 0.41); while detection was significantly reduced for targets occurring after salient notes relative to control trials (unpaired t-test, t(13) = 3.27, p = 0.006).

Though ignored, the auditory melody induced a strong neural response with a clear activation around 3.33 Hz, likely driven by the rhythmic pattern in the melody. A time-frequency profile of neural responses shows energy around 3.33 Hz is particularly prominent after onset of the salient note (Fig. 1C). To quantify the observed change in neural power, the spectral energy averaged over the region [3–3.5] Hz was compared before and after the onset of salient notes and confirmed to be statistically significant (paired t-test, $t(13) = 11.3, p < 10^{-7}$). Part of this increase in neural power is likely due to acoustic changes in the physical nature of the salient note when compared to notes in the melody before the change. It is also not spectrally precise because of the broad frequency resolution of the matching pursuit algorithm used to derive the time–frequency profile shown in Fig. 1C. We therefore focused subsequent analyses on comparing the *identical* note when salient in "test" trials and when non-salient in "control" trials, hence eliminating any effects due to the acoustic attributes of the note itself (see Methods for details).

Using the discrete Fourier transform (DFT), we looked closely at entrainment of cortical responses exactly to 3.33 Hz, as well as other frequencies. Comparing the same note when salient versus not (identical note, test vs. control trials), the neural power in response to the stimulus rhythm significantly increased only at the melody rate (paired t-test, $p < 10^{-4}$. Fig. 2A, A, top

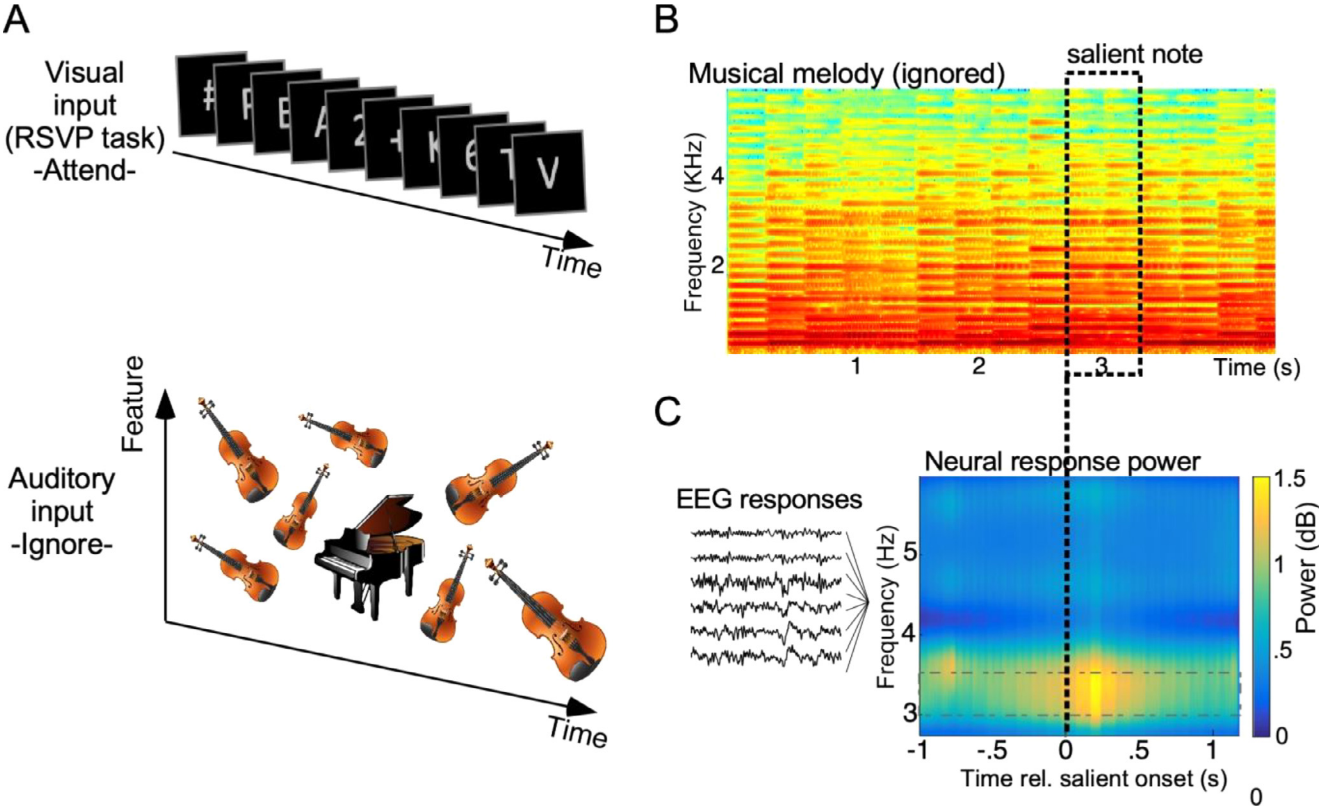

Fig. 1. Schematic of experimental paradigm. **(A)** Participants are asked to attend to a screen and perform a rapid serial visual presentation (RSVP) task (top panel). Concurrently, a melody plays in the background and subjects are asked to ignore it (bottom panel). In "test" trials (shown), the melody has an occasional salient note that did not fit the statistical structure of the melody. In "control" trials (not shown), no salient note was present.**(B)** Power spectral density of a sample melody. Notes forming the rich melodic scene overlap temporally but still form a regular rhythm at 3.33 Hz. Only one note in the melody deviates from the statistical structure of the surround. **(C)** Grand average EEG power shows a significant enhancement upon the presentation of the salient note. The enhancement is particularly pronounced around 3–3.5 Hz, a range including the stimulus rate 3.33 Hz (1/0.3 s).

row), even though both test and control trials have the same underlying 3.33 Hz rhythm. No such increase was found for neural responses at close frequencies (exactly 3.2, 3.4 Hz), nor in higher frequencies. Looking closely at effects of acoustic attributes of the salient note, stronger salience in a particular feature resulted in stronger neural power relative to a weaker salience (Fig. 2A, second and third rows). Specifically, both pitch ($F(1,13) = 37.0, p = 3.9 \times 10^{-5}$) and intensity ($F(1,13) = 35.58, p = 4.7 \times 10^{-5}$) resulted in greater modulation of neural power only at the melodic rhythm. Deviance along timbre also revealed significant differential neural power enhancement only at the stimulus rate (Fig. 2A, fourth row); with guitar deviants driving a larger increase in neural power compared to piano and clavichord notes (Timbre-bg: $F(2,26) = 6.13$, $p = 0.0065$, Timbre-sal $F(2,26) = 15.5, p = 3.7 \times 10^{-5}$). These effects are in line with reported variations of acoustic profiles of clavichord, piano, and guitar, indicating stronger differences in the guitar spectral flux, spectral irregularity as well as temporal attack time relative to the other two instruments (McAdams et al., 1995).

Given the factorial design of the paradigm concurrently probing combinations of features, changes in neural power in response to the rhythm can be examined *across* acoustic dimensions. Results of a within-subjects ANOVA are given in Table 1 (Neural power column). The analysis shows a sweeping range of strong effects and significant interactions across features. Worth noting are main effects of pitch, intensity and timbre (all with significance levels $p < 10^{-4}$). In addition, we note numerous nonlinear interactions across many features including 3-way and 4-way interactions. Specifically, pitch appears to interact strongly with intensity and timbre (both salient and background) in addition to a 3-way interaction between pitch, salient and background timbres. The results also reveal a statistically significant 4-way interaction between all 4 factors (pitch x intensity x salient-timbre x background-timbre). Many effects reported in Table 1 are in line with similar interactions previously reported in behavioral experiments using the same acoustic stimuli (Kaya and Elhilali, 2014); while other interactions are only observed here in neural power responses (e.g. 4-way interaction between pitch * intensity * salient-timbre * background-Timbre).

The precision of neural power effects is striking and reminiscent of effects reported with top-down attentional engagement (Bidet-Caulet et al., 2007; Elhilali et al.,

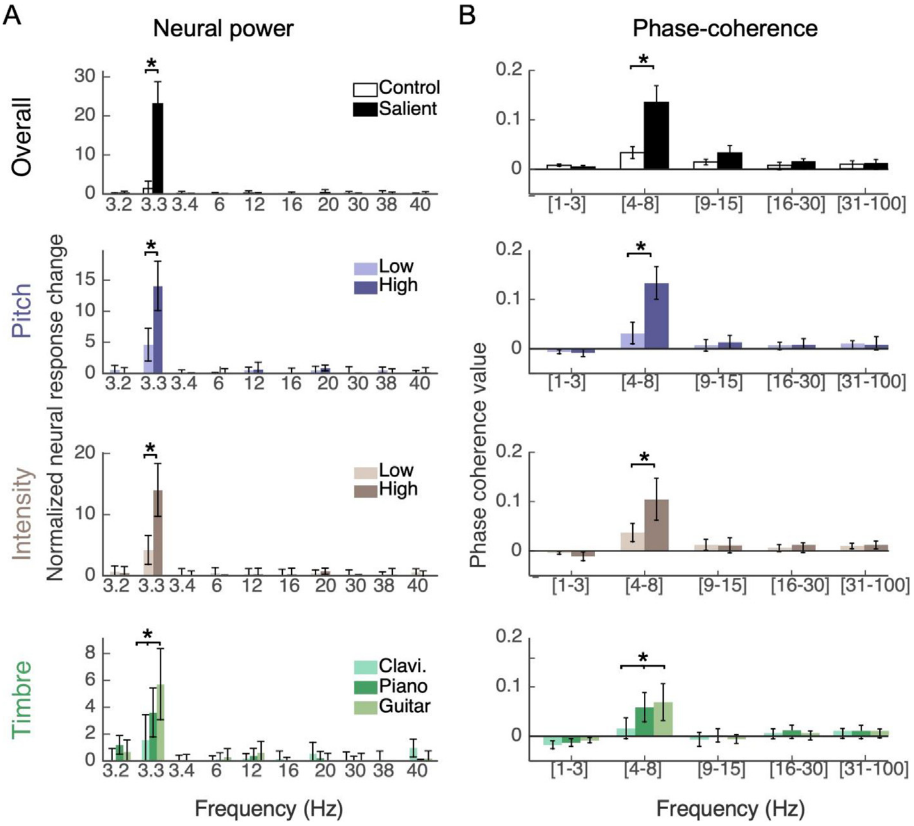

Fig. 2. **(A)** Analysis of neral power at different frequency bins for the salient notes versus identical notes in control trials. Top panel shows power across all notes, while next 3 rows compare power at various degree of salience in pitch, intensity and different instruments (timbre). **(B)** Analysis of cross-trial phase coherence of Hilbert envelopes at different frequency bands of overall salient notes (top) as well as different levels of salience in pitch, intensity and timbre.

2009; Ding and Simon, 2014). However, it is important to examine how much of this modulation can be explained from expected peaks in evoked response potentials (ERP), such as the mismatch negativity (MMN). As depicted in Fig. 3A, comparing the response of salient and control trials revealed MMN and P3a evoked components with significant amplitude effects around these two time windows (paired t-test: $t(13) = -1.4, p = 3.1 \times 10^{-6}$ for MMN and $t(13) = 2.2, p = 1.0 \times 10^{-6}$ for P3a at fronto-central sites), but no significant latency effects. No notable differences in the P50 (paired t-test: $t(13) = 0.3, p = 0.23$) or N1 ($t(13) = 0.3, p = 0.17$) time windows at any channel were noted (Fig. 3B top). Both MMN and P3a components were further modulated by the increase of salience along pitch or intensity (paired t-test: $p < 10^{-3}$ for both components and both features; Fig. 3B, second and third row). The timbre of salient notes showed a significant change in the magnitude of the P3a component (one-way repeated-subjects ANOVA: $p < 10^{-2}$) but no significant modulation of the MMN component.

Next, the dynamic and continuous nature of the experimental stimulus allowed the estimation of spectrotemporal neural filters that process sensory input that are modeled after spectro-temporal receptive fields (STRFs) that have been used to characterize the tuning properties of neurons in auditory cortex (Elhilali et al., 2013). Unlike time-locked ERP analysis, the STRF approach reveals how energy patterns at any point in the stimulus affect the neural response with a delay of *t*, without necessarily aligning to the trial "onset". This STRF profile could be interpreted as an extension of the ERP analysis that reveals additional spectral details about the brain's response to the stimulus melodies. Our derivation of STRFs contrasted neural tuning during control notes against changes in the neural filter due to presence of salient notes (Fig. 4A). The tuning profile showed more pronounced response patterns in salient filters, likely in line with stronger overall responses as reported earlier. Of particular interest were filter characteristics in areas corresponding to time windows of neural responses that showed significant changes in the evoked ERP: the 120–170 ms MMN time window which revealed a deeper negative response, and the 220–270 ms P3a time window which showed a stronger positive response. Looking closely at specific acoustic features, an increase in pitch (Fig. 4B) and intensity (Fig. 4C) salience levels also resulted in a similar stronger response; though pitch salience also induced a broader spectral spread than intensity. Different instruments also gave rise to varying spectro-temporal patterns, possibly indicating different neural processing for each instrument (Fig. 4D). These variations are consistent with greater conspicuity of guitar spectrotemporal structure relative to piano and clavichord notes, particularly in terms of irregularity of spectral spread and sharp temporal dynamics caused by the plucking of guitar strings (Giordano and McAdams, 2010; Peeters et al., 130; Patil et al., 2012).

To complement the neural power analysis, we investigated effects of salient notes on phase-profiles of neural responses. A measure of inter-trial phase-coherence was used to quantify similarity of neural phase patterns across trial repetitions. Again when comparing salient notes with the same notes in control context, phase-coherence was overall strongest in the theta band, where salient notes evoked significantly higher phase-coherence (Fig. 2B, B, top row). Phase-coherence across salient notes also increased based on salience strength. No significant phase effects were observed in the delta or beta ranges. A higher pitch or intensity resulted in stronger coherence compared to a low pitch or intensity difference (Pitch:

Table 1. Feature effects on EEG measures of salience. **P** refers to Pitch, **I** refers to intensity, $\mathbf{T}_s$ refers to the timbre (instrument) of the salient note and $\mathbf{T}_b$ refers to the instrument of scene preceding the salient note. The table shows the F-statistic of within-subject ANOVA along with p (the significance value) and effect size [$\eta^2 p$]. Bolded values indicate significant interactions ($p < 0.01$) after Holm-Bonferroni correction for multiple tests

Effects	F (p) [$\eta^2 p$]	
	Neural power	Theta Coherence
Pitch (P)	**37.00 (3.9 × 10^{-5})** **[0.026]**	**81.28 (5.9 × 10^{-7})** **[0.080]**
Intensity (I)	**35.58 (4.7 × 10^{-5})** **[0.022]**	**23.23 (3.3 × 10^{-4})** **[0.073]**
Timbre-bg (T_b)	**6.13 (6.5 × 10^{-3})** **[0.016]**	**9.85 (6.5 × 10^{-4})** **[0.029]**
Timbre-sal (T_s)	**15.50 (3.7 × 10^{-5})** **[0.018]**	**12.98 (1.2 × 10^{-4})** **[0.045]**
P, I	**28.08 (1.4 × 10^{-4})** **[0.014]**	**29.63 (1.1 × 10^{-4})** **[0.020]**
P, T_b	**8.04 (1.9 × 10^{-3})** **[0.006]**	1.17 (0.32) [0.005]
P, T_s	**6.37 (5.5 × 10^{-3})** **[0.010]**	**9.29 (9.0 × 10^{-4})** **[0.015]**
I, T_b	2.72 (0.08) [0.002]	0.72 (0.49) [0.006]
I, T_s	**8.97 (1.0 × 10^{-3})** **[0.014]**	**6.27 (5.9 × 10^{-3})** **[0.015]**
T_b,T_s	**3.96 (7.0 × 10^{-3})** **[0.012]**	**28.99 (1.1 × 10^{-12})** **[0.060]**
P, I, T_b	0.57 (0.57) [0.002]	**10.42 (4.7 × 10^{-4})** **[0.007]**
P, I, T_s	3.94 (0.03) [0.004]	**8.65 (1.3 × 10^{-3})** **[0.009]**
P, T_b, T_s	**4.91 (1.9 × 10^{-3})** **[0.012]**	**4.05 (6.2 × 10^{-3})** **[0.010]**
I, T_b, T_s	3.30 (.02) [0.004]	0.66 (0.63) [0.002]
P, I, T_b**,** T_s	**4.14 (5.4 × 10^{-3})** **[0.018]**	**4.31** **(4.3 × 10^{-3})[0.011]**

$F(1,13) = 81.28, p = 5.9 \times 10^{-7}$, Intensity: $F(1,13) = 23.23, p = 3.3 \times 10^{-4}$, Fig. 2B, second and third rows). Different salient note timbres also elicited significantly different amounts of phase-coherence (Timbre-bg: $F(2,26) = 9.85, p = 6.5 \times 10^{-4}$, Timbre-sal: $F(2,26) = 12.98, p = 1.2 \times 10^{-4}$, Fig. 2B, B, bottom row). An assessment of interaction effects of phase-coherence in the Theta band also revealed sweeping effects, with many interactions consistent with those observed in neural power, notably an interaction between pitch and intensity, pitch and salient-timbre as well as intensity and salient-timbre (Table 1, right column). Also of note are systematic 3-way interaction between pitch and all other factors (intensity, salient-timbre and background-timbre).

Fig. 5(A) summarizes the nonlinear interactions across acoustic features observed in *both* response power and phase-coherence detailed in Table 1. Effects that are common across both neural measures are shown in black, revealing consistent nonlinear synergy across acoustic features. Pitch appears to interact strongly with all other attributes used in this paradigm, but all other attributes interdependently modulate brain responses of other attributes either in a 2-way, 3-way or even 4-way interaction. Interestingly, these interdependencies are closely aligned with nonlinear interactions obtained from the previous behavioral experiments using the same stimuli (Kaya and Elhilali, 2014) (Fig. 5(B) replicates the published effects for comparison).

Because the experimental design manipulates multiple acoustic features simultaneously, we probed the neural correlates of salience as a function of overall acoustic salience. The greater the number of salient features, the greater the effect on the neural response in neural power and theta phase-coherence (Fig. 6). The figure varies a systematic increase in the number of salient features (x-axis), with change in the neural response (y-axis). A slope quantifying the linear fit of this increase confirms significantly positive increases for neural power (95% bootstrap intervals [0,4.34]) and theta-band phase-coherence (95% bootstrap intervals 3.42, 7.18]). No significant increases are noted for delta-band and beta-band phase-coherence (95% bootstrap intervals [-1.92, 5.57] and [-3.53, 5.79] respectively).

DISCUSSION

This study examines neural markers of auditory salience using complex natural melodies. Specifically, the results show that the long-term statistical structure of sounds shapes the neural and perceptual salience of each note in the melody, much like spatial context dictates salience of visual items beyond their local visual properties (Wolfe and Horowitz, 2004; Nothdurft, 2005; Itti and Koch, 2001). In this work, brain responses are shown to be sensitive to the acoustic context of sounds by tracking the dynamic changes in pitch, timbre and intensity of musical sequences. The presence of salient notes that stand out from their context significantly enhances the rhythm's neural power and cross-trial theta phase alignment of salient events; and causes them to distract subjects from the task at hand, even in another modality (visual task). The degree of modulation of neural responses is closely linked to the acoustic structure of salient notes given their context (Fig. 6)); and reflects a nonlinear integration of variability across a high-dimensional acoustic feature space. For instance, a deviance in the melodic pitch line induces neural changes that are closely influenced by the musical timbre and overall intensity of the melody. While such interactions have been previously reported in behavioral studies (Melara and Marks, 1990; Kaya and Elhilali, 2014), the close alignment between these perceptual effects and neural responses in the context of salience suggests the presence of interdependent encoding of these attributes in the auditory system and provides a neural constraint on such nonlinear interactions that explain perception of salient sound objects. Neurophysiological studies have provided support for such nonlinear integration of acoustic features in overlapping neural circuits (Bizley et al., 2009; Allen et al., 2017); a concept which lays the ground-

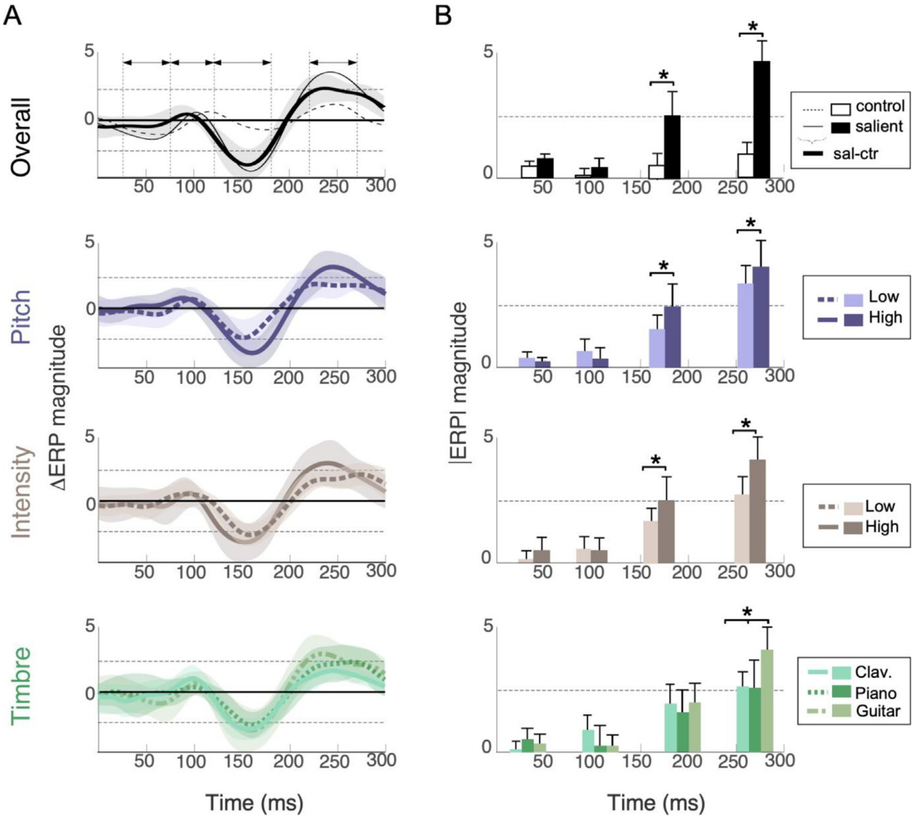

Fig. 3. **(A)** Profile of evoked responses relative to note onset. Each plot shows the difference in response (ΔERP) between a salient note and the identical note in control trials. The top plot also shows original neural waveform of salient note (thin solid line line) and control note (dashed line) before subtraction, as well as difference response (thick solid line). Top plot shows the ΔERP for *all* salient notes, while the next rows show a breakdown of neural responses for different attributes of salient notes (pitch, intensity and timbre) at different levels for pitch and intensity (high salience -solid lines- or low salience -dashed lines). The timbre response contrasts the response of 3 instruments. Shaded areas in all plots correspond to 5-th and 95-th percentile confidence intervals. Horizontal arrows in top plot show windows of interest for statistical analysis of ERP effects: 25–75 ms (for P50), 75–120 ms (for N1), 120–180 ms (for MMN), 225–275 ms (for P3a). **(B)** comparison of absolute value of ERP peaks over 4 windows of interest contrasting salient notes vs. control notes with different acoustic attributes.

work for an integrated encoding of auditory objects in terms of their high-order attributes (Nelken and Bar-Yosef, 2008). Here, we show that such integrated encoding is itself shaped by the long-term statistical structure of the context of the acoustic scene, in line with a wide-range of contextual feedback effects that shape nonlinear neural responses in auditory cortex (Bartlett and Wang, 2005; Asari and Zador, 2009; Mesgarani et al., 2009; Angeloni and Geffen, 2018).

Changes in the neural response to salient notes are specifically observed in neural power and phase-alignment to the auditory rhythm, even with subjects' attention directed away from the auditory stimulus. The enhancement of neural power complements previously reported "gain" effects that have mostly been attributed to top-down attention (Hillyard et al., 1998) and interpreted as facilitating the readout of attended sensory information, effectively modulating the signal-to-noise ratio of sensory encoding in favor of the attended target. A large body of work has shown that directing attention *towards a target of interest* does induce clear neural entrainment to the rate or envelope of the attended auditory streams, hence enhancing its representation (Elhilali et al., 2009; Kerlin et al., 2010). In fact, studies simulating the "cocktail party effect" with multiple competing speakers reveal that neural oscillations entrain to the envelope of the attended speaker (Ding and Simon, 2012; Mesgarani and Chang, 2012; O'Sullivan et al., 2015; Fuglsang et al., 2017). Of particular interest to the present work is the observation that representations of unattended acoustic objects are nonetheless maintained in early sensory areas even if in an unsegregated fashion (Ding and Simon, 2012; Puvvada and Simon, 2017). Here, we observe that even ignored sounds can induce similar gain changes when these events are conspicuous enough relative to their context, effectively engaging attentional processes in a bottom-up fashion. The melodic rhythm used in the current study falls within the slow modulation range typical for natural sounds (e.g. speech) and is commensurate with rates that single-neurons and local field potentials in early auditory cortex are known to phase-lock to (Wang et al., 2008; Kayser et al., 2009; Chandrasekaran et al., 2010). While it is unclear whether the observed enhancement in neural power is a direct result of contextual modulations of these local neural computations or whether it reflects cognitive networks typically engaged in top-down attentional tasks, the nature of the stimulus and observed behavioral effects suggest an engagement of both: the complex nature of salient stimuli likely evokes large neural circuits or multiple neural centers spanning multiple acoustic feature maps, and the observed distraction effects on a visual task also posit an engagement of association or cognitive areas likely spanning parietal and frontal networks in agreement with broad circuits reported to be engaged during involuntary attention (Watkins et al., 2007; Salmi et al., 2009; Ahveninen et al., 2013). The reported presence of a P3a evoked component that is itself modulated by note salience further supports the engagement of involuntary attentional mechanisms that likely extends to neural circuits beyond sensory cortex (Escera and Corral, 2007; Soltani and Knight, 2000).

Complementing the steady-state "gain" effects, the study also reports an enhancement of inter-trial phase-coherence in the theta range whose effect size is

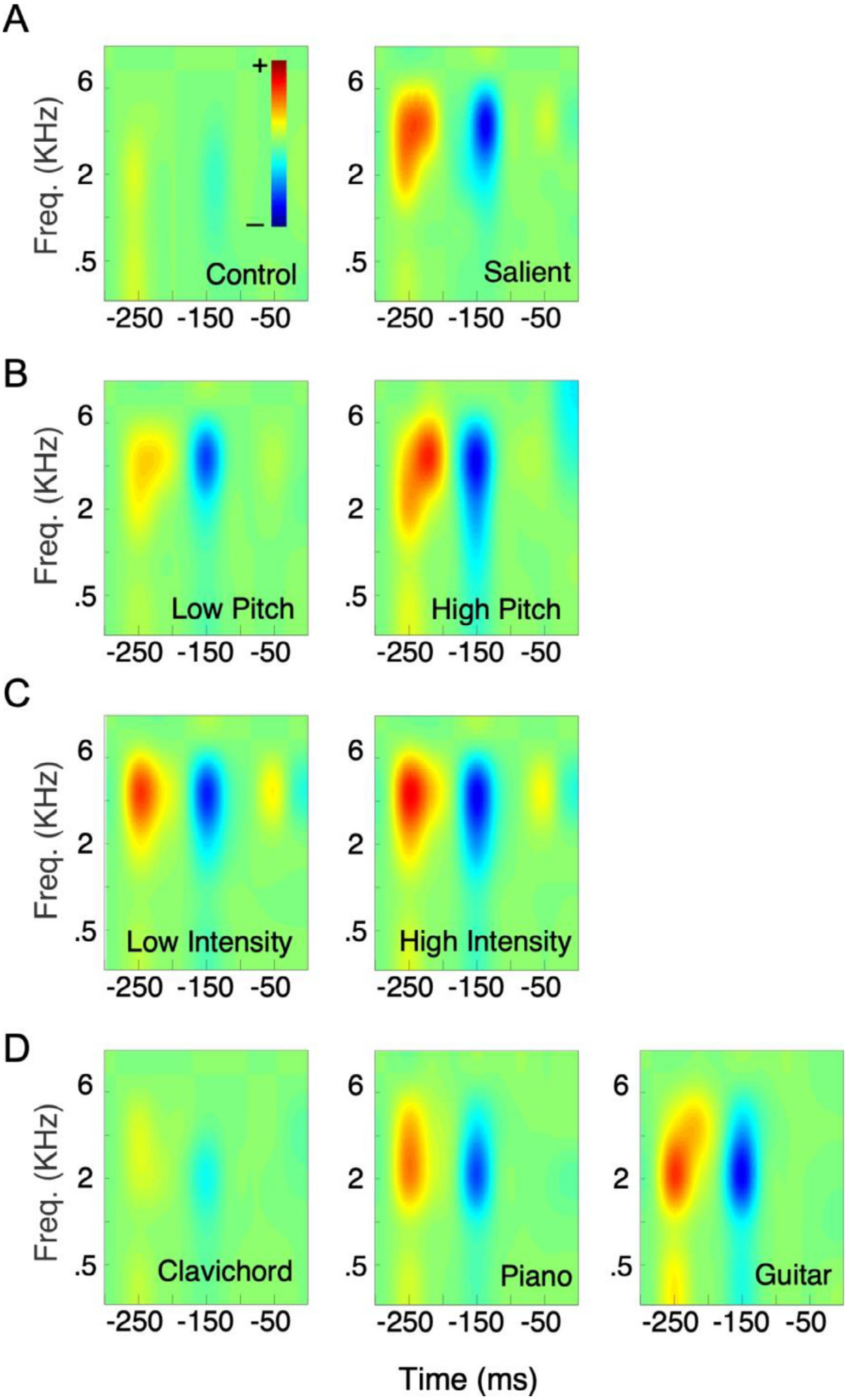

Fig. 4. Estimated STRF averaged across subjects, computed for overall control vs. salient notes **(A)** as well as specific levels of tested features (**(B)** pitch, **(C)** intensity, and **(D)** timbres).

strongly regulated by the degree of salience. Enhanced entrainment to low-frequency cortical oscillations has been posited as a mechanism that boosts or stabilizes the neural representation of attended objects relative to distractors in the environment (Henry and Obleser, 2012; Ng et al., 2012). As a correlate of temporal consistency of brain responses across trials, inter-trial phase coherence measures temporal fidelity in specific oscillation ranges. Modulations in the theta band specifically have been tied to shared attentional paradigms whereby a theta rhythmic sampling operation allows less target-relevant stimuli to be sampled, resulting in a more ecologically essential examination of the environment; thus, these modulations can be a marker of divided attention (Landau et al., 2015; Keller et al., 2017; Spyropoulos et al., 2018; Teng et al., 2018). In the present study, not only does the strength of phase-coherence follow closely the salience of the conspicuous note, but the strong parallels between neural and behavioral nonlinear interactions (Fig. 5) proffer a link between perceptual detection and temporal fidelity of the underlying neural representation of salient events in a dynamic ambient scene.

It is important to note that all changes in neural responses due to presence of salient notes cannot be explained by the absolute values of acoustic features of the deviant instances in the melody. Firstly, the entire melodic piece is highly dynamic, exhibiting a great deal of acoustic variability (e.g. a typical pitch interval of a sequence spans the range [G3-B3]); these changes induce temporal variability in the neural response. Changes reported here are beyond this inherent variability. Secondly, all analyses in the current study compare neural responses to the same note when salient vs. not. It is important to emphasize that the global acoustic profile of a melody (rather than local acoustics) is what dictates the salience of a particular sound event. A piano note is not surprising among pianos, but would be among violins. As such, neural responses are clearly being modulated by the longer-term acoustic profile of the melody and the conspicuous acoustic change of certain notes *given their preceding context*. Such salient changes induce profound effects on brain responses that can be

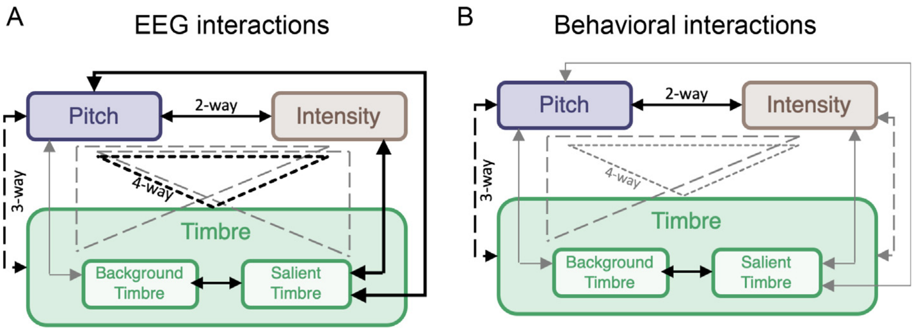

Fig. 5. **(A)** Summary of interaction weights based on neural power to stimulus rhythm and phase coherence results as outlined in Table 1. Solid lines indicate 2-way, dashed lines 3-way and dotted lines 4-way interactions. Effects that emerge for both measures are shown black, and those that are found for at least one measure are shown gray. **(B)** Reproduction of Fig. 4 from (Kaya and Elhilali, 2014) (with permission) which summarizes interaction effects observed in human behavioral responses. Solid lines indicate 2-way, dashed lines 3-way and dotted lines 4-way interactions. Black lines indicate effects that emerge from the behavioral results for the stimulus in this work. Gray lines indicate effects that emerge from behavioral results for speech and nature stimuli tested with the same experimental design as music stimuli.

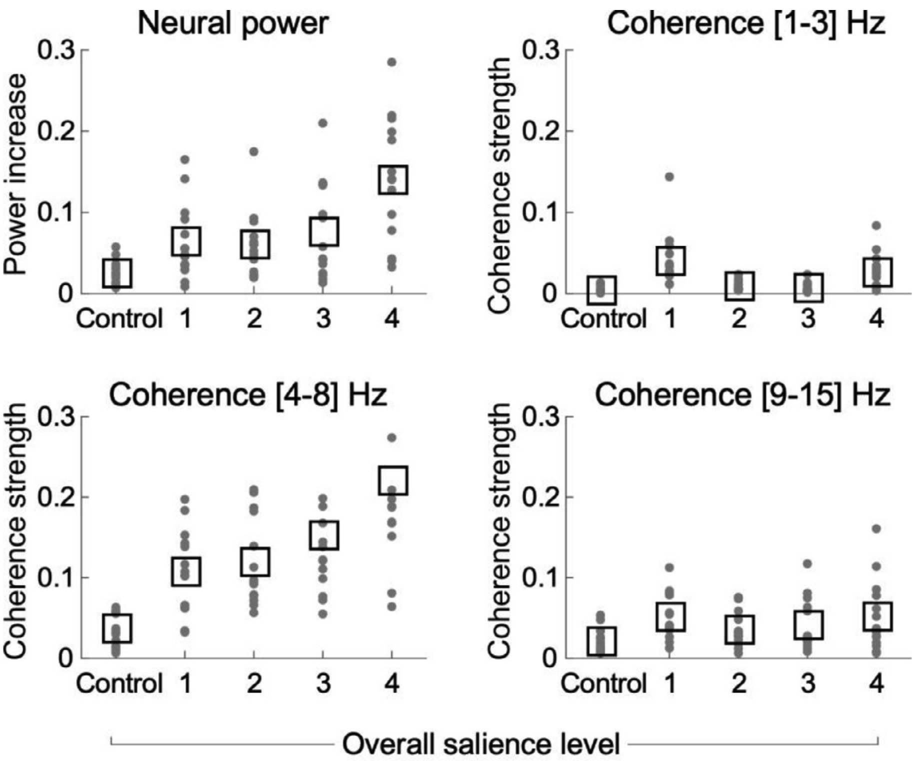

Fig. 6. Distribution of neural effects as a function of overall salience. Both neural power and cross-trial coherence increase with greater stimulus salience. The x-axis denotes the number of features in which the salient note has a high level of difference from control notes, with controls having 0 level of salience. Salience level 1 corresponds to notes with lowest change in acoustic attributes and no change in timbre (i.e., no difference in timbre, 2 dB difference in intensity, and 2 semitones difference in pitch). Changes in timbre (different instrument for salient note) are labeled as higher salience level, with level 4 corresponding to all salient notes with highest change in acoustic attributes and a change between clavichord and piano.

interpreted as markers of auditory salience in the context of complex dynamic scenes.

The use of complex melodic structures in this work is crucial in shedding light on strong nonlinear interactions in neural processing of salient sound events. While effects reported here are heavily tied to acoustic changes in the stimulus, the presence of a mismatch component followed by an early P3a component provides further support that entrainment effects are indeed associated with engagement of attentional networks. The emergence of a deviance MMN component despite the dynamic nature of the background strongly suggests that the auditory system collects statistics about the ongoing environment, thereby forming internal representations about the regularity in the melody. Violation of these regularities is clearly marked by a mismatch component and further engages attentional processes as reflected by the P3a component (Escera and Corral, 2007; Muller-Gass et al., 2007). The presence of both components in this paradigm is in line with existing hypotheses positing a distributed architecture spanning the pre-attentive and attentional cerebral generators and reflecting that the complex nature of salient notes in the melody indeed engages listeners' attention in a stimulus-driven fashion (Escera et al., 2000; Garrido et al., 2009). An interesting question remains regarding the link between these ERP components and neural oscillations. Generally, ERP components, including MMN and P3a, are hypothesized to be a result of either transient bursts of activity across neurons or neural groups time-locked to the stimulus superimposed on "irrelevant" background neural oscillations, or realignment of the phase of ongoing oscillations (phase-resetting) (David et al., 2005; Sauseng et al., 2007). Previous work has observed MMN responses with increased phase-coherence in the theta band with no increase in power, and suggested that MMN is at least partially brought forth by phase-resetting (Klimesch et al., 2004; Fuentemilla et al., 2008; Hsiao et al., 2009). Our study presents similar coherence and ERP results; though it remains an open question whether the two markers reflect different processes. We can speculate of a distinction between these effects by noting that significant ERP amplitude increases are limited to time ranges of the negative and positive components around 150 ms and 250 ms, and that time–frequency analysis by matching pursuit reveals increased effect of target rhythm on a trial-by-trial basis, making evoked responses an unlikely mechanism for the observed entrainment effects.

In a similar vein, it is interesting to consider the distinction between ERP and STRF results. ERPs are obtained by time-locked averages of neural signals, thus extracting the positive or negative signal deflections that occur at the same time across epochs. The STRF, on the other hand, finds a sparse set of filter coefficients that best explain every instance in the epoch as a function of the past 300 ms of input sound (Ding and Simon, 2012; Elhilali et al., 2013). Given the rhythmic nature of the stimulus, the temporal profile derived the STRFs appear to reflect slow temporal dynamics in the acoustic input prominently and reveal strong inhibition and excitation corresponding to time windows of significant ERP components (MMN and P3a, respectively). Crucially, the STRFs reveal that the spectral span of the neural transfer function are also heavily modulated by degree of salience.

Overall, the findings of this study open new avenues to investigate bottom-up auditory attention without relying on active subject responses in the auditory domain, thus eliminating top-down confounds. Results suggest a unified framework where both bottom-up and top-down auditory attention modulate the phase of the ongoing neural activity to organize scene perception. The entrainment measures employed in this study can further be used for natural scenes to decode salience responses from EEG or MEG recordings, allowing the construction of a ground-truth salience dataset for the auditory domain as an analog to eye-tracking data in vision. Naturally, the use of musical melodies offers a great springboard to explore the role of contextual statistics in shaping salience perception and its manifestation in brain responses. Statistical properties of music not only guide encoding of expectations of musical scales (Choi et al., 2014), but also modulate expectations of melodic components that extend beyond local acoustic attributes of the notes (Liberto et al., 2020).

ACKNOWLEDGMENTS

This research was supported by NIH U01AG058532 and R01HL133043, ONR N000141912014, N000141712736, N000141912689 and NSF 1734744.

REFERENCES

Ahveninen J, Huang S, Belliveau JW, Chang W-T, Hämäläinen M (2013) Dynamic oscillatory processes governing cued orienting and allocation of auditory attention. J Cogn Neurosci 25 (11):1926–1943.

Alickovic Emina, Lunner Thomas, Gustafsson Fredrik, Ljung Lennart (2019) A tutorial on auditory attention identification methods. Front Neurosci. https://doi.org/10.3389/fnins.2019.00153. ISSN 1662-453X.

Allen, EJ, Oxenham AJ. Interactions of Pitch and Timbre: How Changes in One Dimension Affect Discrimination of the Other. In Association for Research in Otolaryngology, Midwinter Meeting (abstract), vol. 36; 2013. .

Allen EJ, Burton PC, Olman CA, Oxenham AJ (2017) Representations of pitch and timbre variation in human auditory cortex. J Neurosci 37(5):1284–1293.

Angeloni C, Geffen M (2018) Contextual modulation of sound processing in the auditory cortex. Curr Opin Neurobiol 49:8–15.

Antunes FM, Malmierca MS (2014) An overview of stimulus-specific adaptation in the auditory thalamus. Brain Topogr 27(4):480–499.

Asari H, Zador AM (2009) Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. J Neurophysiol 102(5):2638–2656.

Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative nonlinearities in auditory cortical neurons. Neuron 58(6):956–966.

Bartlett EL, Wang X (2005) Long-lasting modulation by stimulus context in primate auditory cortex. J Neurophysiol 94(1): 83–104.

Besle J, Schevon CA, Mehta AD, Lakatos P, Goodman RR, McKhann GM, Emerson RG, Schroeder CE (2011) Tuning of the human neocortex to the temporal dynamics of attended events. J Neurosci 31(9):3176–3185.

Bidet-Caulet A, Fischer C, Besle J, Aguera P-E, Giard M-H, Bertrand O (2007) Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. J Neurosci 27(35):9252–9261.

Bizley JK, Walker KMM, Silverman BW, King AJ, Schnupp JWH (2009) Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J Neurosci 29(7):2064–2075.

Bregman AS (1990) Auditory scene analysis: the perceptual organization of sound. Cambridge Mass.: MIT Press.

Chakrabarty D, Elhilali M (2019) A Gestalt inference model for auditory scene segregation. PLoS Comput Biol 15(1):e1006711.

Chandrasekaran C, Turesson HK, Brown CH, Ghazanfar AA (2010) The influence of natural scene dynamics on auditory cortical activity. J Neurosci 30(42):13919–13931.

Choi I, Bharadwaj HM, Bressler S, Loui P, Lee K, Shinn-Cunningham BG (2014) Automatic processing of abstract musical tonality. Front Hum Neurosci 8:12.

David O, Harrison L, Friston KJ (2005) Modelling event-related responses in the brain. NeuroImage 25(3):756–770.

David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectro-temporal receptive fields with natural stimuli. Network-Comp Neural 18(3):191–212.

Di Liberto GM, Pelofi C, Bianco R, Patel P, Mehta AD, Herrero JL, Cheveigné AD, Shamma S, Mesgarani N (2020) Cortical encoding of melodic expectations in human temporal cortex. Elife 9:e51784.

Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. P Natl Acad Sci 109(29):11854–11859.

Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. Front Hum Neurosci 8:5.

Elhilali M, Shamma SA, Simon JZ, Fritz JB (2013) A linear systems view to the concept of STRF. In: Depireux D, Elhilali M, editors. Handbook of Modern Techniques in Auditory Cortex. Nova Science Pub Inc. p. 33–60.

Elhilali M, Xiang J, Shamma SA, Simon JZ (2009) Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol 7(6):e1000129.

Escera C, Corral M (2007) Role of mismatch negativity and novelty-p3 in involuntary auditory attention. J Psychophysiol 21(3-4):251–264.

Escera C, Alho K, Schröger E, Winkler I (2000) Involuntary attention and distractibility as evaluated with event-related brain potentials. Audiol Neurootol 5(3–4):151–166.

Fisher RAS (1935) The design of experiments. 8th ed. New York: Hafner Pub. Co.

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11(2):127–138.

Fuentemilla L, Marco-Pallarés J, Münte TF, Grau C (2008) Theta EEG oscillatory activity and auditory change detection. Brain Res 1220:93–101.

Fuglsang S, Dau T, Hjortkjær J (2017) Noise-robust cortical tracking of attended speech in real-world acoustic scenes. NeuroImage 156:435–444.

Garrido MI, Kilner JM, Stephan KE, Friston KJ (2009) The mismatch negativity: a review of underlying mechanisms. Clin Neurophysiol 120(3):453.

Giordano BL, McAdams S (2010) Sound source mechanics and musical timbre perception: evidence from previous studies. Music Percept 28(2):155–168.

Goto M, Hashiguchi H, Nishimura T, Oka R (2003) RWC music database: music genre database and musical instrument sound database. Proc Inter Symp Music Info Retriev:229–230.

Haroush K, Hochstein S, Deouell LY (2010) Momentary fluctuations in allocation of attention: cross-modal effects of visual task load on auditory discrimination. J Cog Neurosci 22(7):1440–1451.

Henry MJ, Obleser J (2012) Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. P Natl Acad Sci 109(49):20095–20100.

Hillyard SA, Vogel EK, Luck SJ (1998) Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philos T R S B 353(1373):1257–1270.

Hsiao FJ, Wu ZA, Ho LT, Lin YY (2009) Theta oscillation during auditory change detection: an MEG study. Biol Psychol 81 (1):58–66.

Huang N, Elhilali M (2017) Auditory salience using natural soundscapes. J Acoust Soc Am 141(3):2163.

Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194–203.

Jaeger M, Bleichner MG, Bauer AKR, Mirkovic B, Debener S (2018) Did you listen to the beat? Auditory steady-state responses in the human electroencephalogram at 4 and 7 Hz modulation rates reflect selective attention. Brain Topogr 31(5):811.

Kaya EM, Elhilali M (2014) Investigating bottom-up auditory attention. Front Hum Neurosci 8:327.

Kayser C, Petkov CI, Lippert M, Logothetis NK (2005) Mechanisms for allocating auditory attention: an auditory saliency map. Curr Biol 15(21):1943–1947.

Kayser C, Montemurro MA, Logothetis NK, Panzeri S (2009) Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. Neuron 61(4):597–608.

Keller AS, Payne L, Sekuler R (2017) Characterizing the roles of alpha and theta oscillations in multisensory attention. Neuropsychologia 99:48–63.

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a cocktail party. J Neurosci 30(2):620–628.

King FW (2009) Hilbert transforms. Cambridge: Cambridge University Press.

Klimesch W, Schabus M, Doppelmayr M, Gruber W, Sauseng P (2004) Evoked oscillations and early components of event-related potentials: an analysis. Int J Bifurcat Chaos 14(2):705–718.

Kowalski N, Depireux DA, Shamma SA (1996) Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J Neurophys 76 (5):3503–3523.

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320(5872):110–113.

Landau AN, Schreyer HM, van Pelt S, Fries P (2015) Distributed attention is implemented through theta-rhythmic gamma modulation. Curr Biol 25(17):2332–2337.

Mahajan Y, Peter V, Sharma M (2017) Effect of EEG referencing methods on auditory mismatch negativity. Front Neurosci 11:10.

Mallat SG, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. IEEE T Signal Process 41(12):3397–3415.

McAdams S, Winsberg S, Donnadieu S, Soete GD, Krimphoff J (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58(3):177–192.

Melara RD, Marks LE (1990) Interaction among auditory dimensions: timbre, pitch, and loudness. Percept Psychophys 48(2):169–178.

Mesgarani N, David S, Shamma S (2009) Influence of context and behavior on the population code in primary auditory cortex. J Neurophys 102:3329–3333.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485 (7397):233–236.

Moore BCJ. Hearing. Academic Press, second ed.; 1995. .

Moskowitz HR, Gerbers CL (1974) Dimensional salience of odors. Ann NY Acad Sci 237(1):1–16.

Muller-Gass A, Macdonald M, Schröger E, Sculthorpe L, Campbell K (2007) Evidence for the auditory P3a reflecting an automatic process: Elicitation during highly-focused continuous visual attention. Brain Res 1170:71–78.

Nelken I (2014) Stimulus-specific adaptation and deviance detection in the auditory system: experiments and models. Biol Cybern 108 (5):655–663.

Nelken I, Bar-Yosef O (2008) Neurons and objects: the case of auditory cortex. Front Neurosci 2(1):107–113.

Ng BSW, Schroeder T, Kayser C (2012) A precluding but not ensuring role of entrained low-frequency oscillations for auditory. Perception. J Neurosci 32(35):12268–12276.

Nothdurft H-C (2005) Salience of feature contrast. In neurobiology of attention. Elsevier. p. 233–239.

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intel Neurosc 156869(12):2011.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25(7): 1697–1706.

Patil K, Pressnitzer D, Shamma S, Elhilali M (2012) Music in our ears: the biological bases of musical timbre perception. PLoS Comput Biol 8(11):e1002759.

Peeters G, Giordano BL, Susini P, Misdariis N, McAdams S (130) The timbre toolbox: extracting audio descriptors from musical signals. J Acoust Soc Am 130(5):2902–2916.

Puvvada KC, Simon JZ (2017) Cortical representations of speech in a multitalker auditory scene. J Neurosci 37(38):9189–9196.

Salmi J, Rinne T, Koistinen S, Salonen O, Alho K (2009) Brain networks of bottom-up triggered and top-down controlled shifting of auditory attention. Brain Res 1286:155–164.

Sauseng P, Klimesch W, Gruber W, Hanslmayr S, Freunberger R, Doppelmayr M (2007) Are event-related potential components generated by phase resetting of brain oscillations? A critical discussion. Neuroscience 146(4):1435–1444.

Schreiner CE (1992) Functional organization of the auditory cortex: maps and mechanisms. Curr Opin Neurobiol 2(4):516.

Schreiner CE (1995) Order and disorder in auditory cortical maps. Curr Opin Neurobiol 5(4):489–496.

Schroger E (1998) Measurement and interpretation of the mismatch negativity. Behav Res Meth Ins C 30(1):131–145.

Shuai L, Elhilali M (2014) Task-dependent neural representations of salient events in dynamic auditory scenes. Front Neurosci 8:203.

Sloas DC, Zhuo R, Xue H, Chambers AR, Kolaczyk E, Polley DB, Sen K (2016) Interactions across multiple stimulus dimensions in primary auditory Cortex. Eneuro 3(4). 0124–16.

Snedecor G, Cochran W (1989) Statistical methods. Ames: Iowa State University Press.

Soltani M, Knight RT (2000) Neural Origins of the P300. Crit Rev Neurobiol 14(3–4):26.

Spyropoulos G, Bosman CA, Fries P (2018) A theta rhythm in macaque visual cortex and its attentional modulation. P Nat Acad Sci 115(24):E5614–E5623.

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. Eur J Neurosci 48 (8):2770–2782.

Ulanovsky N (2004) Multiple time scales of adaptation in auditory cortex neurons. J Neurosci 24(46):10440–10453.

Versnel H, Kowalski N, Shamma SA (1995) Ripple analysis in ferret primary auditory cortex. III. Topographic distribution of ripple response parameters. J Audit Neurosci 1:271–286.

Walsh L, Critchlow J, Beck B, Cataldo A, de Boer L, Haggard P (2016) Salience-driven overestimation of total somatosensory stimulation. Cognition 154:118–129.

Wang X, Lu T, Bendor D, Bartlett E (2008) Neural coding of temporal information in auditory thalamus and cortex. Neuroscience 157:484–493.

Watkins S, Dalton P, Lavie N, Rees G (2007) Brain mechanisms mediating auditory attentional capture in humans. Cereb Cortex 17(7):1697–1700.

Winkler I, Denham SL, Nelken I (2009) Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn Sci 13(12):532–540.

Wolfe JM, Horowitz TS (2004) What attributes guide the deployment of visual attention and how do they do it? Nat Rev Neurosci 5 (6):495–501.

Daniel Wong, Søren Fuglsang, Jens Hjortkjær, Enea Ceolini, Malcolm Slaney, Alain de Cheveigné (2018) A comparison of regularization methods in forward and backward models for auditory attention decoding. Front Neurosci 12(AUG):531. https://doi.org/10.3389/fnins.2018.00531.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a cocktail party. Neuron 77(5):980–991.

(Received 22 November 2019, Accepted 10 May 2020)
(Available online 21 May 2020)