EJN European Journal of Neuroscience

FENS Federation of European Neuroscience Societies

European Journal of Neuroscience, pp. 1–13, 2017 doi:10.1111/ejn.13742

Theta band oscillations reflect more than entrainment: behavioral and neural evidence demonstrates an active chunking process

Xiangbin Teng,[1] Xing Tian,[2,3] Keith Doelling[4,5] and David Poeppel[1,4]
[1]Department of Neuroscience, Max-Planck-Institute for Empirical Aesthetics, Frankfurt, Germany
[2]New York University Shanghai, Shanghai, China
[3]NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai, China
[4]Department of Psychology, New York University, New York, NY, USA
[5]Center for Neural Science, New York University, New York, NY, USA

Keywords: auditory perception, auditory system, MEG analysis, oscillation

Abstract

Parsing continuous acoustic streams into perceptual units is fundamental to auditory perception. Previous studies have uncovered a cortical entrainment mechanism in the delta and theta bands (~1–8 Hz) that correlates with formation of perceptual units in speech, music, and other quasi-rhythmic stimuli. Whether cortical oscillations in the delta-theta bands are passively entrained by regular acoustic patterns or play an active role in parsing the acoustic stream is debated. Here, we investigate cortical oscillations using novel stimuli with 1/f modulation spectra. These 1/f signals have no rhythmic structure but contain information over many timescales because of their broadband modulation characteristics. We chose 1/f modulation spectra with varying exponents of f, which simulate the dynamics of environmental noise, speech, vocalizations, and music. While undergoing magnetoencephalography (MEG) recording, participants listened to 1/f stimuli and detected embedded target tones. Tone detection performance varied across stimuli of different exponents and can be explained by local signal-to-noise ratio computed using a temporal window around 200 ms. Furthermore, theta band oscillations, surprisingly, were observed for all stimuli, but robust phase coherence was preferentially displayed by stimuli with exponents 1 and 1.5. We constructed an auditory processing model to quantify acoustic information on various timescales and correlated the model outputs with the neural results. We show that cortical oscillations reflect a chunking of segments, > 200 ms. These results suggest an active auditory segmentation mechanism, complementary to entrainment, operating on a timescale of ~200 ms to organize acoustic information.

Introduction

Cortical oscillations are entrained not only by strictly periodic stimuli but also by quasi-rhythmic structures in sounds, such as the amplitude envelope of speech (Luo & Poeppel, 2007; Kerlin *et al.*, 2010; Cogan & Poeppel, 2011; Peelle *et al.*, 2013; Zion Golumbic *et al.*, 2013; Doelling *et al.*, 2014; Kayser *et al.*, 2015) and music (Doelling & Poeppel, 2015), the frequency modulation envelope (Henry & Obleser, 2012; Herrmann *et al.*, 2013; Henry *et al.*, 2014), and even abstract linguistic structure (Ding *et al.*, 2015). These studies have advanced our understanding of how the auditory system exploits regular temporal structure to organize acoustic information. It is not clearly understood, though, whether cortical oscillations tracking sounds are a result of neural responses passively driven by rhythmic structures or reflect a built-in constructive processing scheme, namely that the auditory system employs a windowing process to actively group acoustic information (Ghitza & Greenberg, 2009; Ding & Simon, 2014).

It has been proposed that cortical oscillations in the auditory system reflect an active parsing mechanism—the auditory system chunks sounds into segments of around 150–300 ms, roughly a cycle of the theta band, for grouping acoustic information (Ghitza & Greenberg, 2009; Schroeder *et al.*, 2010; Ghitza, 2012). A slightly different (but related) view hypothesizes that the auditory system processes sounds using temporal integration windows of multiple sizes concurrently: Within a short temporal window (~30 ms), temporally fine-grained information is processed; a more 'global' acoustic structure is extracted within a larger temporal window (~200 ms)

Correspondence: Xiangbin Teng, as above.
E-mail: xiangbin.teng@gmail.com

X. Tian and K. Doelling contributed equally to this work.

Received 25 April 2017, revised 16 August 2017, accepted 28 September 2017

Edited by Ali Mazaheri

Reviewed by Andrew Dimitrijevic, Sunnybrook Health Sciences Centre, Canada; Simon Hanslmayr, University of Birmingham, UK

The associated peer review process communications can be found in the online version of this article.

(Poeppel, 2003; Giraud & Poeppel, 2012). These frameworks are largely based on studying speech signals that contain these timescales as relatively obvious components: The temporal modulations of speech peak around 4–5 Hz (Ding *et al.*, 2017). However, if such a segmentation scale or integration window exists at the timescale of ~200 ms in the auditory system *intrinsically*, then we should find evidence of its deployment even when the sounds have broadband spectra and are irregularly modulated over a wide range of timescales. In contrast, if cortical oscillations are solely or primarily stimulus-driven, one ought not to find robust oscillatory activity using such irregular sounds.

Natural sounds, such as environmental noise, speech, and some vocalizations, often have broadband modulation spectra that show a 1/f pattern: The modulation spectrum has larger power in the low frequency range and the modulation strength decreases as frequency increases (Voss & Clarke, 1978; Singh & Theunissen, 2003; Theunissen & Elie, 2014). This characteristic of modulation spectra can be delineated using a straight line at a logarithmic scale, with its exponent indicating how sounds are modulated across various timescales. For example, environmental noise has a relatively shallow 1/f modulation spectrum with an exponent of 0.5, while speech has a steeper spectrum with an exponent of f between 1 and 1.5 (Singh & Theunissen, 2003). As 1/f spectra reflect acoustic dynamics across many timescales, and not rhythmic structure centered at a narrow frequency range, 1/f stimuli are well suited to test how the auditory system spontaneously organizes acoustic information across various timescales.

We generated frequency modulated sounds having 1/f modulation spectra with different exponents, to imitate irregular dynamics in natural sounds (Garcia-Lazaro *et al.*, 2006) and inserted a tone of short duration (50 ms) as a detection target. We recorded participants' neurophysiological responses while they listened to the 1/f stimuli and detected the embedded tones. We were interested to see what timescale of acoustic information is used to detect salient changes (i.e. embedded tones) and at what frequencies robust oscillatory activity is evoked by irregular 1/f stimuli. We then used an auditory processing model to quantify acoustic information over different timescales. By employing mutual information analysis, we determine the timescale over which acoustic information is grouped. By designing our experiment in this manner, we are able to investigate the temporal structure imposed by the neural architecture of the auditory system to sample information from the environment.

Materials and methods

Participants

Fifteen participants (age 23–49, one left-handed, eight females) took part in the experiment. Handedness was determined using the Edinburgh Handedness Inventory (Oldfield, 1971). All participants had normal hearing and no neurological deficits. Written informed consent was obtained from every participant before the experiment. The experimental protocol was approved by the New York University Institutional Review Board.

Stimuli and design

We followed the methods used in Garcia-Lazaro *et al.* (2006) to generate similar (but modified) stimuli with modulation spectra of 1/f. A schematic plot of the stimulus generation process is shown in Fig. 1A.

We first generated spectral modulation envelopes with 'random-walk' profiles using an inverse Fourier method. We fit the modulation spectra to have a 1/f shape, with exponents at 0.5, 1, 1.5, and 2 (Fig. 1A left panel) and converted the spectra from the frequency domain back to the temporal domain using inverse Fourier transformation. The phase spectra were obtained from pseudo-random numbers drawn uniformly from the interval $[0, 2\pi]$. Because we fixed stimulus length to 3 s and the sampling rate to 44 100 Hz, we created modulation spectra of 44 100$\times$ 3 points with a frequency range of 0–22 050 Hz. Using different random number seeds for the phase spectra, we were able to generate spectral modulation envelopes (Fig. 1A middle panel) with different dynamics for each exponent. The modulation envelopes were normalized to have unit standard deviation.

Second, we created tone complexes comprising tonal components spaced at third-octave intervals and then used the spectral modulation envelopes generated as above to modulate the tone complexes. We set the fundamental frequency to 200 Hz and limited the frequency range of the stimuli to between 200 and 4000 Hz, well within humans' sensitive hearing range. The frequencies of each tonal component were modulated through the frequency range from 200 Hz to 4 kHz by the envelopes generated in the first step. Modulated tonal components outside this frequency range at one end would reenter it at the opposite end so that the number and spacing of the tonal components within this frequency range was always constant.

We used the same random seed to generate one stimulus for each of four exponents, 0.5, 1, 1.5, and 2 so that all four stimuli have the same phase spectrum but different modulation spectra. During the experiment, we presented these four stimuli 25 times, and we term these four the 'frozen' stimuli. Next, we used distinct random seeds to generate 25 'distinct' stimuli with different phase spectra for each exponent. Each of these was presented once, creating four groups of 'distinct' stimuli. In total, there were eight stimulus groups, comprising four groups of 'frozen' stimuli and four groups of 'distinct' stimuli. In total, 200 stimuli (25 stimuli $\times$ 4 exponents $\times$ 2 stimulus types) were used in the study.

A 1000-Hz pure tone of 50-ms duration was inserted into the 'distinct' stimuli, and the onset of the tone was randomly distributed between 2.2 and 2.7 s. The signal-to-noise ratio of the tone to these distinct stimuli was fixed at -15 dB, because in preliminary testing we determined that a tone at SNR -15 dB can be detected at an adequate rate (i.e. avoiding ceiling or floor effects). We applied a cosine ramp-up function in a window of 30 ms at the onset of all stimuli and normalized the stimuli to ~70 dB SPL (sound pressure level).

Stimulus analysis

To characterize the spectral and temporal modulations in our stimuli, we computed modulation power spectra (MPS) for the four 'frozen' stimuli used in the experiment (Fig. 1B) (Singh & Theunissen, 2003; Elliott & Theunissen, 2009). We first created time–frequency representations of the stimuli using the log amplitude of their spectrograms obtained with Gaussian windows. We then applied the 2D Fourier Transform to the spectrograms and created MPS by taking the amplitude squared as a function of the Fourier pairs of the time and frequency axes. As temporal modulations in our stimuli represent acoustic dynamics across timescales, we averaged the MPS across the spectral modulation dimension to show averaged temporal modulation spectra for each stimulus (Fig. 1C). Figure 1B shows that the prominent spectral modulation centers around 1.7 cycles per

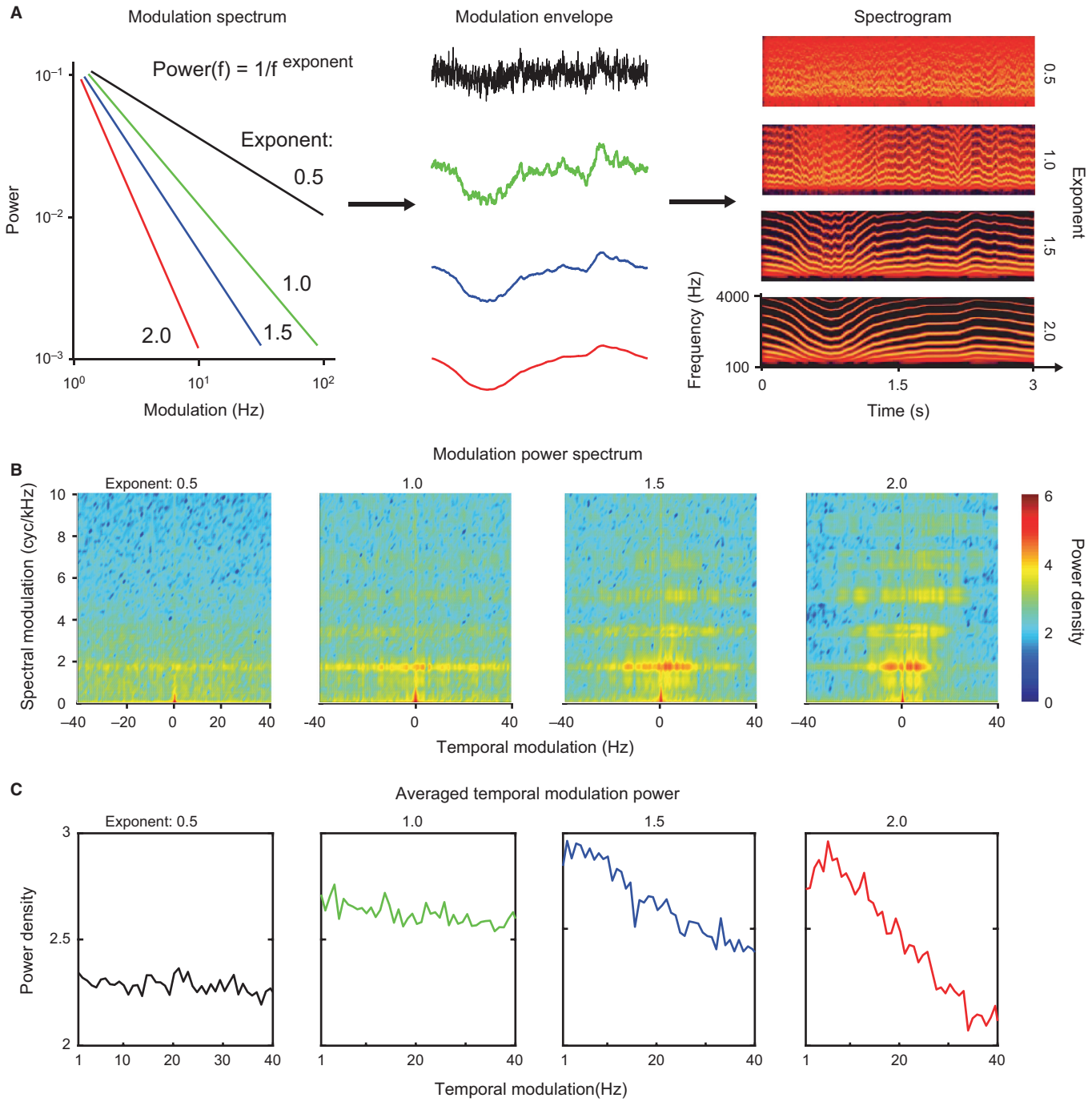

FIG. 1. Stimulus generation and modulation power spectrum. (A) Schematic plot of stimulus generation. Left panel: schematic plot of modulation spectra used to generate modulation envelopes. The color code represents the spectra of different exponents. Black: 0.5; green: 1.0; blue: 1.5; red 2.0. Middle panel: modulation envelopes generated using the four different exponents. (Color code as in the modulation spectra.) Right panel: spectrograms of the four 'frozen' stimuli (see Methods) used in the experiment. Sound files of the stimuli can be accessed here: http://edmond.mpdl.mpg.de/imeji/collection/kZalRMtxa19mlRyG. (B) Modulation power spectra of the four frozen stimuli. The dashed boxes show increased power density at a spectral modulation of around 1.7 cycles per 1000 Hz in the stimuli. (C) The averaged temporal modulation spectrum. The averaged temporal modulation was computed by averaging along the spectral modulation dimension of the modulation power spectrum (in (B). From left to right, the averaged temporal modulation spectrum of each stimulus becomes steeper as the exponent increases. Note that there are no prominent peaks in the averaged temporal modulation spectra that indicate regular modulations centered at a narrow frequency band.

1000 Hz; specifically, at this modulation frequency, there is increased modulation power from exponent 0.5–2. Figure 1C shows that the averaged temporal modulations vary with the modulation spectra we used to generate each stimulus. The stimulus with exponent 0.5 shows a flat averaged temporal modulation spectrum and has low modulation power, whereas the stimulus with exponent 2.0 has the steepest averaged temporal modulation spectrum. Note, importantly, that the averaged temporal modulation spectra of all four stimuli show no peak of power density between 4 and 7 Hz (Fig. 1C).

MEG recording, preprocessing, and protocol

Magnetoencephalography signals were measured with participants in a supine position and in a magnetically shielded room using a 157-channel whole-head axial gradiometer system (KIT, Kanazawa Institute of Technology, Japan). A sampling rate of 1000 Hz was used, with an online 1–200 Hz analog band-pass filter and a notch filter centered around 60 Hz. After the main experiment, participants were presented with 1-kHz tone beeps of 50-ms duration as a localizer to determine their M100 evoked responses (Roberts *et al.*, 2000). 20 channels with the largest M100 responses in both hemispheres (10 channels in each hemisphere) were selected as auditory channels for further analysis for each participant individually.

Magnetoencephalography data analysis was conducted in MATLAB 2015b (The MathWorks, Natick, MA, USA) using the Fieldtrip toolbox 20160106 (Oostenveld *et al.*, 2011) and the wavelet toolbox in MATLAB. Raw MEG data were noise-reduced offline using the time-shifted principle component analysis (de Cheveigné & Simon, 2007). Trials were visually inspected, and those with artifacts such as channel jumps and large fluctuations were discarded. An independent component analysis was then used to correct for artifacts caused by eye blinks, eye movements, heartbeat, and system noise. After preprocessing, 0 to (at most) 5 trials were removed for each exponent of each stimulus type, leaving a minimum of 20 trials per condition. To avoid biased estimation of inter-trial phase coherence, we included exactly 20 trials in the analysis for all exponents of all stimulus types. Each trial was divided into a 5-s epoch, with a 1-s pre-stimulus period and a 4-s post-stimulus period. Each trial was baseline corrected by subtracting the mean of the whole trial prior to further analysis.

During MEG scanning, all stimuli, both 'frozen' and 'distinct', were presented in a pseudo-randomized order for each participant. After each stimulus was presented, participants were required to push one of two buttons to indicate whether they heard a tone in the stimulus. Between 1 and 2 s after participants responded, the next stimulus was presented. The participants were required to keep their eyes open and to fix on a white cross in the center of a black screen. The stimuli were delivered through plastic air tubes connected to foam ear pieces (E-A-R Tone Gold 3A Insert earphones, Aearo Technologies Auditory Systems).

Behavioral data analysis

Behavioral data were analyzed in MATLAB using the Palamedes toolbox 1.5.0 (Prins & Kingdom, 2009). For each exponent, there were 50 stimuli, half of which had a tone embedded. A two-by-two confusion matrix was created for each exponent by treating the trials with the tone embedded as 'target' and the other trials as 'noise'. Correct detection of the tone in the 'target' trials was counted as 'hit', while reports of hearing a tone in the 'noise' trials were counted as 'false alarm'; d-prime values were computed based on hit rates and false alarm rates of each table. A half artificial incorrect trial was added to the table with all correct trials (Macmillan & Creelman, 2004).

Evoked responses to tones

We calculated the root mean square (RMS) of evoked responses to the onset of tones for each 'distinct' group across 20 auditory channels and across 20 trials. Baseline was corrected using the MEG signal from 200-ms pre-onset of the tone in each selected channel. After baseline correction, we averaged RMS across 20 auditory channels.

Evoked responses to stimulus onset

We calculated RMS of evoked responses to the onset of stimulus for each 'frozen' group and each 'distinct' group across 20 auditory channels and across 20 trials. Baseline was corrected using the MEG signal from 200-ms pre-onset of the stimuli in each selected channel. After baseline correction, we averaged RMS across 20 auditory channels.

Local SNR of the embedded tones

The exponents of stimuli result in different modulation profiles and can modulate local SNR of the embedded tones across stimuli. Because the differences of local SNR could potentially explain the behavioral performance of tone detection, we computed the local SNR of the embedded tones using rectangular temporal windows combined with equivalent rectangular bandwidth (ERB) at 1000 Hz (Glasberg & Moore, 1990). We chose five temporal window sizes, 50, 100, 200, 300, and 500 ms, and five bandwidths, 0.25, 0.5, 1, 1.5, and 2 ERB (33, 66, 133, 199, and 265 Hz). Across different bandwidths, we centered the temporal window in the middle of the tone—25 ms after tone onset—and computed power of the 'distinct' stimuli without the tone in this temporal window. Then, to compute local SNR, we divided the power of the tone by the power of the 'distinct' stimuli within the temporal window and the narrow band. We transformed the values of local SNR into decibels by taking a log with base 10 and multiplying by 10.

Phase coherence and power analysis

To extract time–frequency information, single-trial data from each MEG channel were transformed using functions of Morlet wavelets embedded in the Fieldtrip toolbox, with frequencies ranging from 1 to 50 Hz in steps of 1 Hz. As all the stimuli used are 3 s long, to be able to extract low-frequency oscillations (e.g. 1 Hz) and to balance spectral and temporal resolution of time–frequency transformation, window length increased linearly from 1.5 cycles to seven cycles from 1 to 20 Hz and then was kept constant at seven cycles above 20 Hz. Phase and power responses were extracted from the wavelet transform output at each time–frequency point.

The 'inter-trial phase coherence' (ITPC) was calculated for all eight groups of stimuli at each time–frequency point [details can be seen in Lachaux *et al.* (1999)]. ITPC is a measure of consistency of phase-locked neural activity entrained by a stimulus across trials. ITPC of a specific frequency band is thought to reflect cortical entrainment to temporal modulations in sounds (Luo *et al.*, 2013; Ding & Simon, 2014; Doelling *et al.*, 2014; Kayser *et al.*, 2015) and therefore can be used as an index to indicate temporal coding of each stimulus type at certain frequency band here. Although event-evoked responses and ITPC both measure evoked neural responses and are highly correlated (Mazaheri & Picton, 2005), evoked responses show energy that spreads across a broad frequency range (VanRullen *et al.*, 2014) and are limited by event rates (Lakatos *et al.*, 2013). Furthermore, phase reset of ongoing oscillations of certain frequency band is not always correlated with sensory events (Mazaheri & Jensen, 2006, 2010). Therefore, we chose to use ITPC in our current study.

The induced power response was calculated for all eight groups of stimuli and was normalized by dividing the mean power value in the baseline range (−0.6 to −0.1 s) and converted to decibel units.

The ITPC and power response for four groups of 'frozen' stimuli were averaged from 0.25 to 2.8 s post-stimulus to avoid effects of

neural responses evoked by the stimulus onset and offset. We applied the same calculation of ITPC and power response to four groups of 'distinct' stimuli, but used the results as a baseline for the ITPC and power response of the 'frozen' stimuli. The differentiated ITPC (dITPC) and differentiated induced power were obtained by subtracting the ITPC and power response for 'distinct' stimuli out from 'frozen' stimuli for each participant. These two indices reflect phase-locked responses to the repeated temporal structure in the 'frozen' stimuli.

Auditory processing model

The 1/f stimuli have a broadband modulation spectrum and contain information across all timescales. To quantify acoustic information on each timescale and later to examine on what timescale the auditory system groups acoustic information, we constructed an auditory processing model inspired by the concept of cochlear-scaled entropy (Stilp & Kluender, 2010; Stilp *et al.*, 2010) using temporal filters of different sizes. By convolving temporal filters with the envelopes of the stimuli in each cochlear band, we can extract acoustic changes, which represent critical acoustic information on different timescales —and can be seen as an analogue to features in visual stimuli resulting from convolution with Gabor filters (Olshausen & Field, 2004). An illustration of this auditory processing model can be seen in Fig. 4.

First, the stimuli were filtered using a gammatone filterbank of 64 bands. The envelope of each cochlear band was extracted using Hilbert transformation on each band and taking the absolute values (Glasberg & Moore, 1990; Søndergaard & Majdak, 2013). We then convolved the envelope of each band with the temporal filters that we constructed (described below). The values calculated from the convolution were centered on the middle point of the temporal filters and were normalized according to the length of the temporal filter used. We padded 500 ms time points at the beginning and the end of the stimuli. After convolution, we took out padded points and only saved the time points of the original stimuli. We then took a vector norm at each time point across 64 cochlear bands.

The temporal filter was constructed by multiplying a Gaussian temporal window with one period of a sinusoid wave. We chose Gaussian temporal windows of ten sizes: 20, 40, 60, 80, 100, 140, 200, 300, 400, and 500 ms, with the mean centered in the middle of the temporal window and the standard deviation being one-fifth of window length. We then created sinusoid waves from 0 to 2*pi with periods corresponding to each Gaussian temporal window. Then, we multiplied one period of the sinusoid waves with 10 Gaussian temporal windows of corresponding sizes to create 10 temporal filters.

These temporal filters function as a one-dimensional filter that extracts changes in each cochlear band, which can be compared to narrowband spectral–temporal receptive fields often found in inferior colliculus (Escabí *et al.*, 2003; Andoni *et al.*, 2007; Carlson *et al.*, 2012). Within a temporal window in which the envelope fluctuates abruptly, the output of the convolution would give a large value. The calculation of the vector norm summarizes temporal changes across all cochlear bands and generates a value at each time point that represents broadband spectro-temporal changes within this temporal window. This is intended to roughly correspond to auditory processes of cortical areas employing spectral–temporal receptive fields with broadband tuning properties (Theunissen *et al.*, 2000; Machens *et al.*, 2004; Theunissen & Elie, 2014). For example, if the frequency modulation changes abruptly and harmonics sweep across frequency bands within a temporal window, the convolution would generate large values that differ across frequency bands. Taking a vector norm would generate a high value. Therefore, we can quantify acoustic changes along both temporal and spectral domains using output from this model.

The model outputs calculated at each time point indicate the presence of acoustic changes on the timescale corresponding to the temporal filter size. We refer to the model outputs as Acoustic Change Index (ACI). Finally, we downsampled the ACI from 44 100 to 100 Hz to match the sampling rate of the phase series in the MEG signals (100 Hz).

Differentiated mutual information between ACI and phase series

To determine at what timescale acoustic information is extracted by the auditory system, we computed mutual information between phase series of MEG signals and ACI. Mutual information is an index to quantify how much information is shared between two time series and suggests correlation between two series (Cogan & Poeppel, 2011; Gross *et al.*, 2013; Ng *et al.*, 2013; Kayser *et al.*, 2015). We chose to compute mutual information instead of a linear correlation because ACI is an index of real numbers while the phase series is both circular and derived from imaginary numbers. A linear correlation cannot correctly measure the relationship between these two metrics. While ITPC cannot tell us which features in the stimulus drive robust phase coherence, if the phase series in the theta band is found to have high mutual information with ACI of this stimulus at a timescale of 200 ms, we can reasonably conclude that the auditory system extracts acoustic information in this stimulus on a timescale of 200 ms.

We computed the mutual information between the phase series of each frequency (1–50 Hz) collected under the frozen stimuli and ACI of different timescales for each corresponding 'frozen' stimulus type. Then, we computed the mutual information between phase series collected under the 'distinct' stimuli and ACI of different timescales for each corresponding 'frozen' stimulus type. Next, we calculated the differences between the mutual information using trials collected under 'frozen' stimuli and the mutual information using trials under 'distinct' stimuli. By doing this, we subtracted out mutual information contributed by spontaneous phase responses evoked by sounds in general and also normalized the mutual information across frequencies to remove the effects caused by 1/f characteristics of neural signals (He *et al.*, 2010; He, 2014). This differentiated mutual information, resulted from using trials collected under 'distinct' stimuli as a baseline, highlighted the mutual information between the structure of ACI and the phase series of MEG signals. For example, we computed the mutual information between ACI for a frozen stimulus of exponent 1 and 20 phase series collected under this frozen stimulus from MEG signals, and then computed the mutual information between ACI for this frozen stimulus and 20 phase series collected from MEG signals when subjects were listening to 20 'distinct' stimuli of exponent 1. We took a difference between these two values of mutual information and used this difference as the differentiated mutual information.

Mutual information was calculated with the Information Breakdown Toolbox in MATLAB (Pola *et al.*, 2003; Magri *et al.*, 2009). For each frequency of the MEG response, the phase distribution was composed of six equally spaced bins: 0 to pi/3, pi/3 to pi * 2/3, pi * 2/3 to pi, pi to pi * 4/3, pi * 4/3 to pi * 5/3, and pi * 5/3 to pi * 2. The ACI was grouped using eight bins equally spaced from the minimum value to the maximum value. Eight bins were chosen to have enough discrete precision to capture changes in acoustic properties while making sure that each bin has sufficient counts for mutual

information analysis, as the greater number of bins would lead to zero counts in certain bins.

The estimation of mutual information is subject to bias caused by finite sampling of the probability distributions because limited data were supplied in this study. Therefore, a quadratic extrapolation embedded in the Information Breakdown Toolbox was applied to correct bias. Mutual information is computed on the data set of each condition. A quadratic function is then fit to the data points, and the actual mutual information is taken to be the zero-crossing value. This new value reflects the estimated mutual information for an infinite number of trials and greatly reduces the finite sampling bias (Montemurro *et al.*, 2007; Panzeri *et al.*, 2007). The mutual information value of each frequency was calculated for each subject and for each channel across trials before averaging.

Results

Tone detection performance increases with exponent though SNR is constant

Behavioral results

Subjects detected tones inserted into the 'distinct' stimuli. The behavioral results (Fig. 2A) demonstrate that participants' sensitivity to tones (d-prime) increased to sounds with increasing exponent, although, importantly, the SNR is the same across all stimuli. The behavioral performance in detecting tones was examined using a repeated-measures one-way ANOVA (RMANOVA) with the main factor of Exponent. There is a significant main effect of Exponent ($F_{3,42} = 34.07$, $P < 0.001$, $\eta_p^2 = 0.709$), and a linear trend test showed a significant upward trend ($F_{1,14} = 59.19$, $P < 0.001$, $\eta_p^2 = 0.809$).

RMS of tone evoked responses

As the listeners' performance on detecting tones varied across stimuli with different exponents of the 1/f stimuli, we examined whether the evoked responses elicited by the tones also show an effect of Exponent (Fig. 2B, upper panel). We calculated the RMS of the MEG signal elicited by tones, averaged over 20 auditory channels, and conducted, on each time point from the onset point of tones to 250 ms after tone onset, a one-way RMANOVA with Exponent as the main factor. After adjusted FDR correction, we found a significant main effect of Exponent from 120 to 175 ms ($P < 0.01$) after tone onset. To investigate this further, we averaged across this window and found a significant linear trend ($F_{1,14} = 25.16$, $P < 0.001$, $\eta_p^2 = 0.642$). The result is shown in Fig. 2B (lower panel). The RMS results correspond to behavioral results and demonstrate that exponents do modulate detection of tones. The behavioral results are not likely caused by response bias.

Local SNRs

The behavioral results and RMS results demonstrate that tone detection varies with exponent. Although the global SNR was matched across the stimuli with different exponents, local SNR varies with exponent and could cause differences in tone detection performance across different stimuli. Therefore, we computed local SNR using rectangular temporal windows of different sizes combined with ERBs of different bandwidths across all 25 trials for each of four exponents. The trial number marched the trails used in behavioral analysis. Pearson's correlation between behavioral results and the local SNRs across four exponents was then calculated to assess whether the local SNR can explain tone detection performance.

We found high correlation coefficients (> 0.8) between behavioral results and local SNR computed using all combinations of temporal window sizes and ERB bandwidths (Fig. 2D). To rule out spurious correlations, we established a significant threshold using a shuffling procedure. We first shuffled the labels of the four types of 'distinct' stimuli and generated a new set of stimuli, and then computed local SNR of each type of stimuli. We then correlated the local SNR with behavioral results to get a correlation coefficient for each combination of temporal window size and ERB bandwidth. We repeated this shuffling procedure 1000 times and used a right-sided alpha level of 99% as the significance threshold level. Significant correlations between behavioral results and local SNR were found for the temporal window sizes between 140 and 250 ms, combined with ERB bandwidths from 1 to 2. We plotted local SNR against tone detection performance on each exponent separately for the significant peak correlation computed using each ERB bandwidth (Fig. 2E). These results show that tone detection performance can be explained by the local SNR modulated by exponents. The acoustic structure of the stimuli becomes sparser with larger exponents, and therefore, local SNR increases, which facilitates tone detection in the stimuli. Most importantly, the local SNR computed using the temporal window of around 200 ms can best capture the behavioral variance. This suggests that a temporal window of ~200 ms is used by the auditory system to group acoustic information and extract salient changes in acoustic streams.

Exponent modulates onset responses and differentiated inter-trial phase coherence in the delta and theta bands

RMS of onset responses

As the acoustic structures of the 1/f stimuli vary with exponents, we examined whether the onset responses to the stimuli also show an effect of Exponent (Fig. 3A). We calculated the RMS of the MEG signal elicited by eight stimulus groups (four 'frozen' groups and four 'distinct' groups), averaged over 20 auditory channels, and conducted a one-way RMANOVA with Exponent as the main factor, separately for the 'frozen' stimuli and the 'distinct' stimuli. The one-way RMANOVA was conducted on each time point from the onset point to 250 ms after stimulus onset. After adjusted FDR correction, we found a significant main effect of Exponent for the 'frozen' stimuli from 90 ms to 115 ms and for the 'distinct' stimuli from 95 ms to 105 ms and from 115 ms to 130 ms ($P < 0.01$). The RMS results demonstrate that onset responses increase with exponents. As onset responses are sensitive to acoustic structures of sounds and are modulated by spectral complexity (Shahin *et al.*, 2007), the results here are likely caused by the spectral sparsity—as the exponent increases, spectral modulation of the stimuli becomes more centered (Fig. 1) and, therefore, spectral sparsity increases with exponents.

Differentiated inter-trial phase coherence

The dITPC, the difference of phase coherence between the 'frozen' stimuli and the 'distinct' stimuli on the independently defined auditory channels, was calculated from 1 to 50 Hz. The results are shown in Fig. 3B. We observed robust phase coherence in the theta band (5–7 Hz) for stimuli of all four exponents and some degree of selectivity for the stimuli with exponents 1 and 1.5. In the delta band (1 Hz), there was a preference in phase coherence for the

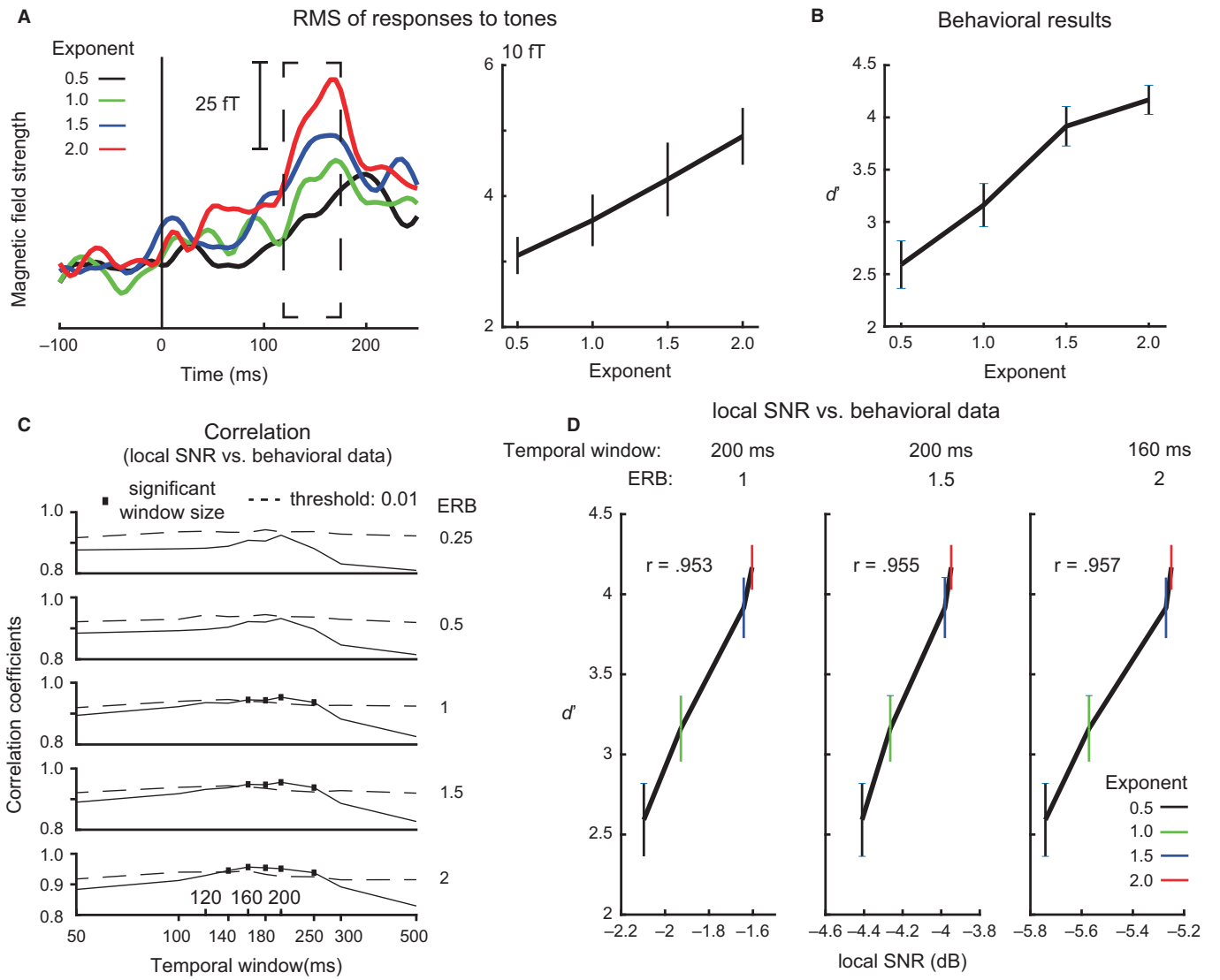

FIG. 2. (A) Root mean square (RMS) of MEG waveform responses to tones. Left panel: RMS of evoked fields to tones from 200 ms before tone onset to 250 ms after tone onset. (Color code as in Fig. 1) The dashed box indicates the time range from 120 to 175 ms that show significant effects of the Exponent ($P < 0.05$, one-way RMANOVA). Right panel: RMS averaged from 120 ms to 175 ms. The behavioral and the RMS results show the same upward trend along exponent values. (B) Behavioral results of tone detection. (C) Correlations between tone detection performance and local SNR. ERB indicates bandwidth used to compute local SNR. ERB refers to one equivalent rectangular bandwidth (133 Hz) of the narrowband centered on 1000 Hz. *Y*-axis indicates values of correlation coefficients. *X*-axis shows different temporal window sizes. The dashed lines indicate significant thresholds ($P < 0.01$, one-sided), and the square highlights significant correlation results. Local SNRs computed using the temporal window ranging between 140 and 250 ms and the bandwidths larger than 1 ERB can explain the differences in tone detection rate across the different stimuli. (D) local SNR plotted again behavioral data for the highest correlation of each bandwidth. From left to right, the bandwidth is 1, 1.5, and 2 ERB. Temporal window indicates the temporal window used to compute the high correlation of each bandwidth. *X*-axis indicates local SNR. The color of error bars codes for different exponents. This result suggests that the auditory system uses a temporal window of ~200 ms to chunk the acoustic stream for separation of targets from background sounds. Lines represent mean, and error bars represent ± SEM.

stimuli with exponent 2. The topographies of dITPC for four exponents in the delta and theta bands are shown in Fig. 3C.

To measure the effects of exponent on dITPC across frequencies, we conducted a one-way RMANOVA with Exponent as the main factor from 1 to 50 Hz. After adjusted FDR correction, this revealed a main effect of Exponent from 5 to 7 Hz ($P < 0.05$), which is in the theta band range, and a main effect at 1 Hz ($P < 0.05$), which is in the delta band. We then averaged dITPC within two frequency ranges separately and conducted a two-way RMANOVA with factors of Exponent and Frequency band (delta: 1 Hz; and theta: 5–7 Hz) on dITPC. We found a main effect of Exponent ($F_{3,42} = 5.24$, $P = 0.004$, $\eta_p^2 = 0.273$) and an interaction between Exponent and Frequency band (Greenhouse–Geisser corrected: $F_{3,42} = 11.64$, $P < 0.001$, $\eta_p^2 = 0.454$). A one-way RMANOVA with a factor of Exponent conducted separately for each frequency band shows a main effect both in the delta band (Greenhouse–Geisser corrected: $F_{3,42} = 7.69$, $P < 0.001$, $\eta_p^2 = 0.355$) and in the theta band ($F_{3,42} = 8.28$, $P < 0.001$, $\eta_p^2 = 0.372$). A *post hoc* paired *t*-test conducted in the theta band (5–7 Hz) showed that dITPC of stimuli with exponents 1 and 1.5 are significantly larger than the stimuli with exponent 0.5 [exponent 1: t (14) = 4.27, $P = 0.006$, $d = 2.28$; exponent 1.5: t (14) = 4.08, $P = 0.006$, $d = 2.18$] after Bonferroni

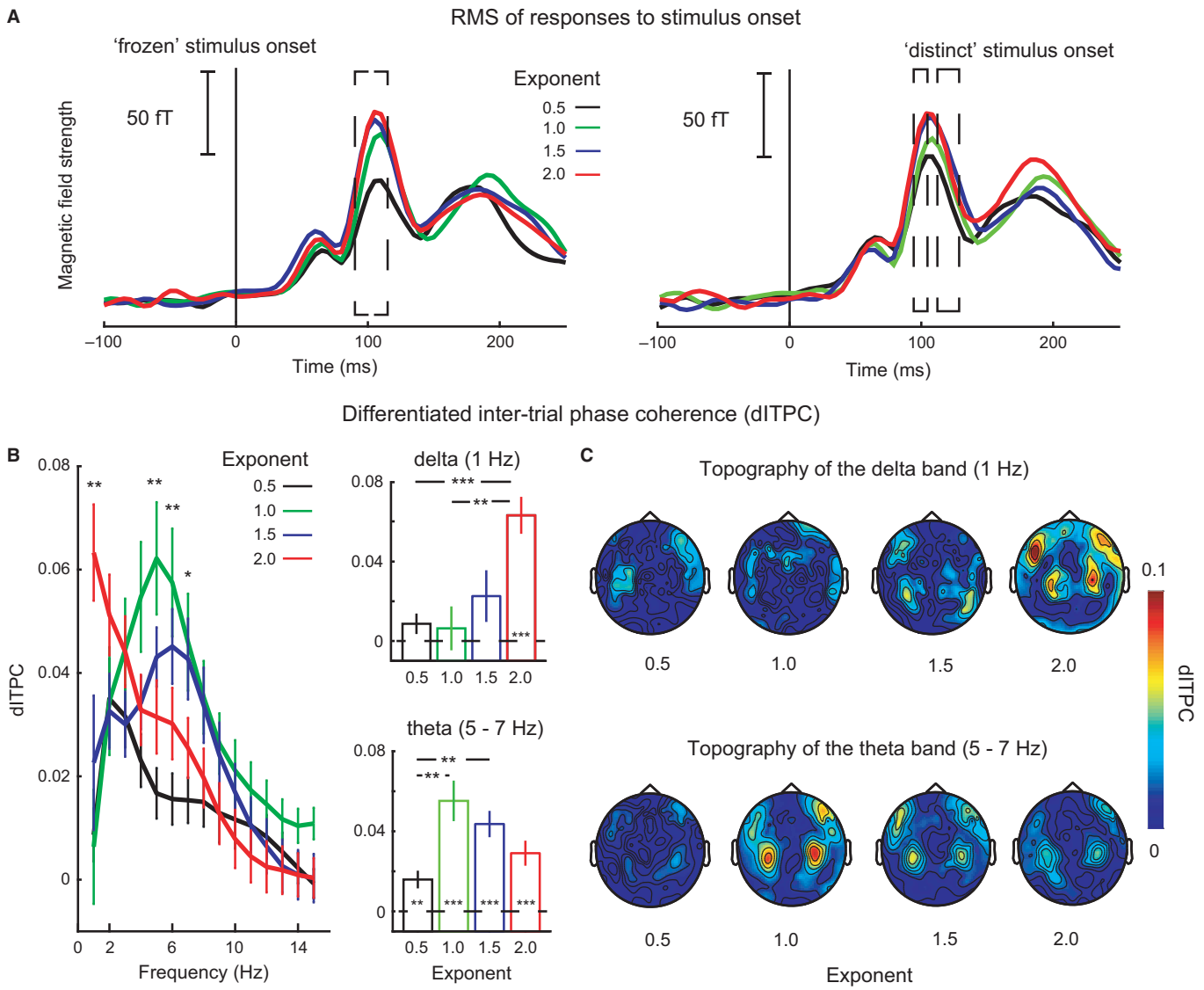

FIG. 3. RMS of responses to stimulus onset and differentiated inter-trial phase coherence (dITPC). (A) RMS of response to the 'frozen' and 'distinct' stimuli. Left panel: RMS of responses to the 'frozen' stimuli. Right panel: RMS of responses to the 'distinct' stimuli (color code as in Fig. 1). The dashed box of the left panel indicates the time range from 90 to 115 ms that shows significant effects of the Exponent for the 'frozen' stimuli ($P < 0.05$, one-way RMANOVA). The dashed boxes of the right panel indicate the time ranges from 95 to 105 ms and from 115 to 130 ms that show significant effects of the Exponent for the 'distinct' stimuli. (B) dITPC on auditory channels. Left panel: dITPC from 1 to 15 Hz (color code as in Fig. 1). We found significant main effects of Exponent in two frequency bands: delta (1 Hz) and theta (5–7 Hz). Right panels: averaged dITPC within frequency bands. Asterisks inside bars indicate that dITPC is significantly above zero. The theta band activity tracks all stimuli, even if there is no regular temporal structure present; delta band activity is more narrowly responsive to stimuli with exponent 2. (C) Topographies of dITPC in the delta and theta bands. Typical auditory response topographies can be seen in both bands. This underscores that the robust phase coherence results originate from auditory processing regions. Error bars represent ± SEM.

correction. Comparisons of dITPC for stimuli of exponents 1 and 1.5 with stimuli of exponent 2 were significant but did not survive correction for multiple comparisons [exponent 1: t (14) = 2.31, $P = 0.036$, $d = 1.23$; exponent 1.5: t (14) = 2.34, $P = 0.035$, $d = 1.25$]. In the delta band (1 Hz), the paired t-test shown that dITPC of stimuli with exponent 2 is significantly larger than the stimuli with exponent 0.5 [t (14) = 7.12, $P < 0.001$, $d = 3.81$] and exponent 1.0 [t (14) = 4.16, $P = 0.006$, $d = 2.22$].

Because dITPC reflects the difference of phase coherence between the 'frozen' stimuli and the 'distinct' stimuli, a one-sample t-test against zero on dITPC of each stimulus type in each band tests whether there is robust phase coherence across trials evoked by the 'frozen' stimuli. In the delta band (Fig. 3B, right upper panel), we found dITPC was significantly above zero when the exponent is 2 [t (14) = 6.80, $P < 0.001$, $d = 3.63$]. In the theta band (Fig. 3B, right lower panel), we found significant dITPC above zeros across all exponents [Exponent 0.5: t (14) = 3.59, $P = 0.024$, $d = 1.92$; Exponent 1.0: t (14) = 5.45, $P < 0.001$, $d = 2.91$; Exponent 1.5: t (14) = 6.63, $P < 0.001$, $d = 3.54$; Exponent 2.0: t (14) = 4.64, $P < 0.001$, $d = 2.48$]. Bonferroni correction was applied in each band.

In summary, the results show that all four types of 'frozen' stimuli evoked robust phase coherence in the theta band. This supports the hypothesis that phase coherence observed in the theta band (and in many studies) is not solely a result of stimulus-driven entrainment, as no regular temporal modulation exists in the stimuli.

The stimuli with exponent 1 and 1.5 revealed higher phase coherence values than the stimuli with exponent 0.5. This phase coherence pattern in the theta band showed a similar pattern to findings in ferrets using single-unit recording (Garcia-Lazaro *et al.*, 2006). Our results further show that this coding preference comes from the theta band, which indicates an underlying auditory process on a timescale of ~150–250 ms. The auditory processing on a timescale of 150–250 ms, reflected by robust phase coherence in the theta band, may be critical and is possibly the reason for the preference found in Garcia-Lazaro *et al.* (2006).

Surprisingly, we observed in the delta band that the stimuli of exponent 2 evoked robust phase coherence. The differences in dITPC patterns between the theta and delta bands indicate that the auditory system independently tunes to information on the timescales corresponding to the theta and delta bands (Cogan & Poeppel, 2011).

Differentiated Induced Power shows no effect

We examined effects of exponents on induced power from 1 to 50 Hz by conducting a one-way RMANOVA with Exponent as the main factor. We found no significant effect on Exponent from 1 to 50 Hz after adjusted FDR correction ($P > 0.05$). This suggests that the power response does not differentially code temporal information critically, which is also consistent with previous studies (Cogan & Poeppel, 2011; Luo & Poeppel, 2012; Ng *et al.*, 2013; Doelling *et al.*, 2014; Kayser *et al.*, 2015).

Raw power shows no effect and does not bias ITPC estimation

We examined effects of the exponents on raw power (without baseline correction) from 1 to 50 Hz by conducting a one-way RMANOVA with Exponent as the main factor. We did such tests on the 'frozen' stimuli and the 'distinct' stimuli, separately. We found no significant effect of Exponent for the 'frozen' stimuli from 1 to 50 Hz after adjusted FDR correction ($P > 0.05$). Similarly, we found no significant effect of Exponent for the 'distinct' stimuli from 1 to 50 Hz after adjusted FDR correction ($P > 0.05$). This suggests that the power is homogenous across different exponents, and therefore, estimation of ITPC for stimuli of different exponents should not be biased by the power.

Differentiated mutual information between phase and ACI on distinct scales

Next, we used a mutual information approach to quantify at what timescale the acoustic information in the stimuli robustly entrained cortical oscillations in the delta (1 Hz) and theta (5–7 Hz) bands. In Fig. 4A–F, we illustrate how ACI was generated by the auditory processing model. The waveforms of stimuli were filtered through the Gammatone filter bank of 64 bands (Fig. 4B), and cochleograms were generated for the 'frozen' stimuli (Fig. 4C). We convolved each band of the cochleogram with temporal filters of various lengths (Fig. 4D) and created a convolved cochleogram for each filter length (Fig. 4E). Vector norm was applied on the convolved cochleogram, which resulted in ACI. An example of ACI computed using a filter length of 200 ms was shown in Fig. 4F. The ACI of each 'frozen' stimulus was used to compute mutual information. From the mutual information results, we found that the delta band oscillation was unaffected by the temporal filter size, whereas in the theta band the mutual information showed an effect of the filter size starting from 200 ms (Fig. 4G).

A three-way Frequency band × Exponent × Filter size RMANOVA was conducted on differentiated mutual information. We found a significant main effect of Exponent (Greenhouse–Geisser corrected: $F_{4,42} = 5.22$, $P = 0.004$, $\eta_p^2 = 0.272$). We also found significant interaction effects between Frequency band and Exponent (Greenhouse–Geisser corrected: $F_{3,42} = 8.42$, $P < 0.001$, $\eta_p^2 = 0.387$), between Exponent and Filter size ($F_{27,378} = 1.58$, $P = 0.036$, $\eta_p^2 = 0.101$), and between Frequency band and Exponent and Filter size ($F_{27,378} = 2.42$, $P < 0.001$, $\eta_p^2 = 0.147$).

We then conducted a two-way Filter Size × Exponent RMANOVA in the delta band. We found a significant main effect of Exponent (Greenhouse–Geisser corrected: $F_{3,42} = 6.84$, $P = 0.007$, $\eta_p^2 = 0.328$) but not of Filter size (Greenhouse–Geisser corrected: F (9,126) = 0.59, $P = 0.802$, $\eta_p^2 = 0.041$). The interaction was not significant ($F_{27,378} = 1.34$, $P = 0.123$, $\eta_p^2 = 0.087$).

In the theta band, we conducted a two-way Filter Size × Exponent RMANOVA and found significant main effects of Exponent (Greenhouse–Geisser corrected: $F_{3,42} = 10.14$, $P < .001$, $\eta_p^2 = 0.420$) and of Filter size (Greenhouse–Geisser corrected: $F_{9,126} = 7.19$, $P < .001$, $\eta_p^2 = 0.339$). The interaction between Exponent and Filter size is also significant ($F_{27,378} = 5.10$, $P < .001$, $\eta_p^2 = 0.267$). To test which filter size differentiates among stimulus types, we conducted a one-way RMANOVA on each filter size with Exponent as main factor. After Bonferroni correction, we found main effects of Exponent on the filter sizes: 200 ms (Greenhouse–Geisser corrected: $F_{3,42} = 7.69$, $P = .048$, $\eta_p^2 = 0.354$), 400 ms (Greenhouse–Geisser corrected: $F_{3,42} = 12.74$, $P = .006$, $\eta_p^2 = 0.476$), and 500 ms (Greenhouse–Geisser corrected: $F_{3,42} = 9.87$, $P = .010$, $\eta_p^2 = 0.414$). Then, we examined what stimulus type was affected by filter size by conducting a one-way RMANOVA with Filter size as a main factor. We found a main effect on Filter size for the stimuli with exponent 0.5 (Greenhouse–Geisser corrected: $F_{9,126} = 6.82$, $P = .008$, $\eta_p^2 = 0.291$), exponent 1.0 (Greenhouse–Geisser corrected: $F_{9,126} = 5.76$, $P = .036$, $\eta_p^2 = 0.328$), and exponent 1.5 (Greenhouse–Geisser corrected: $F_{9,126} = 7.45$, $P < .001$, $\eta_p^2 = 0.347$) but not for the stimuli with exponent 2 ($F_{9,126} = 2.22$, $P = .100$, $\eta_p^2 = 0.137$).

To summarize the results of the differentiated mutual information analysis, we found that, although the stimuli in our experiment have a 1/f modulation spectrum and show no dominant temporal modulation frequencies or regular temporal patterns, the phase patterns of the theta band cortical oscillations were captured by the ACI extracting temporal information larger than 200 ms. On the other hand, cortical oscillations in the delta band are not captured by ACI computed on the timescales < 500 ms.

The finding that the delta band is unaffected by the filter size is probably because the delta band oscillations tune to acoustic changes on a long timescale (e.g. > 1 s). The acoustic information represented by ACI is on a timescale smaller than 500 ms, which does not contribute to the global change of the stimuli extracted by a large temporal window (e.g. > 1 s). Another explanation is that the delta band tunes to high-level information in the stimuli that our model fails to reveal.

That theta phase shows greater MI with ACI of timescales larger than 200 ms is consistent with the results of phase coherence in the theta band (Fig. 4). A reasonable hypothesis is that the auditory system uses a default temporal window of > 200 ms to chunk the continuous acoustic stream (VanRullen & Koch, 2003; Ghitza & Greenberg, 2009; Ghitza, 2012; Giraud & Poeppel, 2012; VanRullen, 2016). This temporal window size is reflected in the dominant theta oscillations found in both our results and in other studies (Ding & Simon, 2013; Luo *et al.*, 2013; Andrillon *et al.*, 2016). In stimuli without overt rhythmic structure, the auditory system

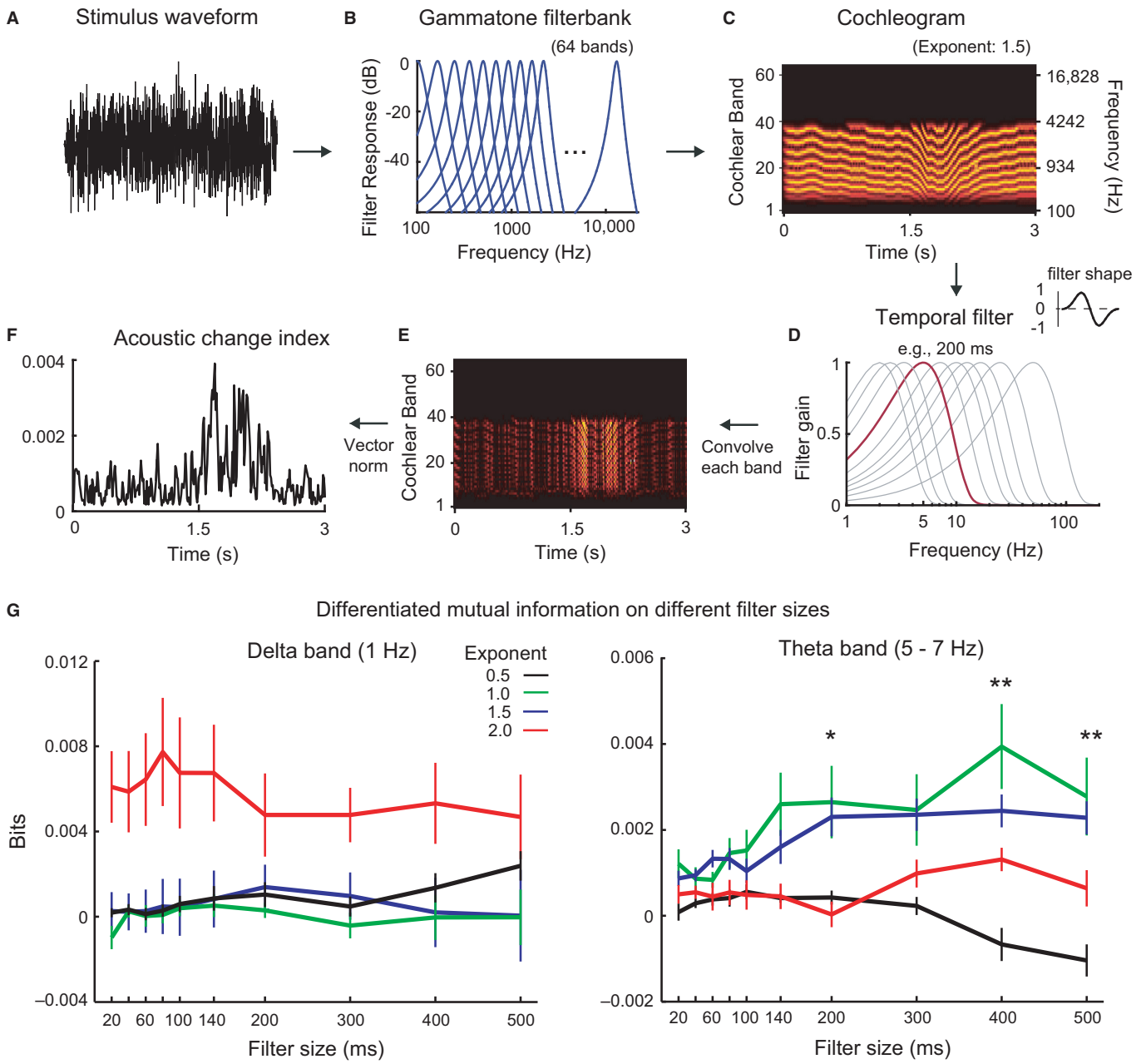

FIG. 4. Illustration of the auditory processing model and differentiated mutual information results. (A) The waveform of a stimulus of exponent 1.5. (B) A schematic plot of the Gammatone filter bank of 64 bands used in the model. (C) The cochleogram of the stimulus. The left *y*-axis represents the number of the cochlear channels and the right *y*-axis represents the center frequency of each channel. (D) Frequency responses of temporal filters. The frequency responses of all temporal filters were plotted. Beside the plot, the filter shape on temporal dimension is shown. We highlighted the frequency response of the temporal filter of 200 ms, which is used here for further illustrations. (E) Convolution of the temporal filter with each band in the cochleogram. After convolving with the temporal filter, large values appear at time points where the modulation changes abruptly (depending on the temporal filter size). (F) Acoustic change index (ACI) resulted from taking the vector norm of (E). (G) Differentiated mutual information in the delta and theta bands across different filter sizes. *X*-axis: filter size of the temporal filter used in the auditory processing model. *Y*-axis: bits of differentiated mutual information. The color code is as in Fig. 1. In the delta band, the differentiated mutual information does not vary with the filter size. Differentiated mutual information in the theta band varied with filter size (asterisks indicate main effect of Exponent). The mutual information results demonstrate that the auditory processing model with temporal windows of > 200 ms is consistent with neural auditory processing extracting temporal regularities from irregular sounds. Lines and bars represent mean while error bars represent ± SEM.

actively parses the acoustic stream and extracts acoustic changes using a temporal window corresponding to roughly a cycle of a theta band oscillation. Because this chunking process has a limited time constant, that is, 150–s300 ms, only the acoustic changes on a timescale of larger than 200 ms are captured within acoustic dynamics across all timescales in our stimuli. Our auditory processing model using a temporal filter with different sizes simulated this hypothesized chunking process. The model used temporal windows to chunk acoustic streams and computed acoustic changes within each temporal window. The ACI on the timescale of ~200 ms, therefore, reflects the acoustic information extracted by this chunking process in the auditory system.

In summary, the results suggest that the auditory system uses a temporal windowing process to chunk acoustic information and extracts acoustic changes from irregular stimuli, and this temporal window is larger than 200 ms. The preference of the auditory system for stimuli with exponents 1 and 1.5, shown in our results in the theta band and in Garcia-Lazaro *et al.* (2006), is likely a result of this chunking process.

Discussion

We investigated neurophysiological responses to stimuli with 1/f modulation spectra and tested how listeners detect embedded tones. We found that cortical oscillations in the theta band track the irregular temporal structure and show a preference to 1/f stimuli with exponents 1 and 1.5, which roughly correspond to signals with the modulation spectrum of speech. The delta band oscillations are entrained by stimuli with exponent 2, which has the slowest temporal modulation. The fact that we find robust phase coherence in theta band in the absence of regular dynamics suggests that theta oscillations are not a simple consequence of the acoustic input but rather may represent the temporal structure of internal neural processing. By computing mutual information between the model outputs and the phase series in the delta and theta bands, we found that phase coherence in the theta band can be best explained by acoustic changes captured by temporal windows at least as large as 200 ms. Further supporting this finding, the local SNR computed using a temporal window of 200 ms predicts the tone detection rates and confirms the mechanism by which the auditory system uses a temporal window (~ 200 ms) to group acoustic information and extract salient acoustic changes.

Robust phase coherence in the theta band is not solely stimulus driven

As the 1/f stimuli were not specifically modulated between 5 and 7 Hz to drive theta band oscillations (Fig. 1C), the robust theta oscillatory activity, therefore, must partly originate from an intrinsic auditory processing mechanism. In most of studies using rhythmic stimuli, the observed cortical entrainment in the theta band could be due to the fact that the regular temporal structure overlaps with the timescale of this architecturally intrinsic and probably innate grouping mechanism. Robust phase tracking in the theta band seems to be ubiquitously evoked by sounds. It has been shown, for example, that repeated noise induces phase coherence in the theta band, and the magnitude of phase coherence correlates with behavioral performance (Luo *et al.*, 2013; Andrillon *et al.*, 2016). In such studies, there is no regular temporal structure in sounds centered in the theta band that entrains the theta band oscillations.

One reason for theta band tracking of sounds of various temporal structures, regular and irregular, is possibly that the theta band oscillations play an active role in perceptual grouping of acoustic information, rather than being passive, stimulus-driven neural activities (Ghitza & Greenberg, 2009; Schroeder *et al.*, 2010; Ghitza, 2012; Riecke *et al.*, 2015). Our auditory processing model simulates this chunking process across timescales, and the mutual information results between the model outputs and phase series in the theta band differentiate stimuli of different exponents on a timescale larger than 200 ms and echoes the results of dITPC. Therefore, the robust phase coherence can be explained by the chunking process simulated by our auditory processing model (Fig. 4A–F).

The active chunking process is probably a trade-off between integrating a long period of acoustic information for precise analysis and making timely perceptual decisions. Although acoustic streams are continuous, the auditory system cannot integrate acoustic information over an arbitrary long period because of limited information capacity of the auditory system and of requirements for humans to make fast perceptual decisions. This chunking process of ~200 ms divides continuous acoustic streams into discrete perceptual units, so that further auditory analysis could be conducted timely within a ~200-ms temporal window for humans to make immediate perceptual decisions.

Preferential tuning to exponents 1 and 1.5 due to chunking

Our results of dITPC in the theta band indicate a preference of the auditory system for stimuli with exponents 1 and 1.5, which replicate the response pattern found in Garcia-Lazaro *et al.* (2006) using single-unit recording in ferret primate auditory cortex. Furthermore, the mutual information results (Fig. 4) suggest that this preference is likely caused by the chunking process with temporal windows larger than 200 ms in the auditory system. Although all of the 1/f stimuli have dynamics across all timescales, the chunking process mainly extracts dynamics on a timescale corresponding to the theta band. The stimuli with exponents < 1 and larger than 1.5 are either modulated too rapidly or too slowly, so that the dynamics on the timescale of the theta band range has less 'chunking potency' than in the stimuli with exponents 1 and 1.5.

Tone detection results explained by local SNR confirms a chunking process of ~200 ms

We found that although the long-term SNR of tones is the same across all four types of stimuli, the detection rates differ because of local SNR modulated by the exponents. These results are illuminated by the informational masking literature, which suggests that the structure of background sounds (maskers) matters when listeners try to detect a target (Brungart, 2001; Kidd *et al.*, 2007). The key finding here is that participant behavior is modulated by the structure of background maskers in the same 200-ms window. This suggests the auditory system is extracting 200-ms windows for temporal analysis. This finding supports our interpretation of the neural data, discussed above, that the auditory system groups acoustic information on a timescale of ~200 ms and further suggests that this chunking process is probably fundamental for further auditory analysis; that is to say, the separation of targets from background sounds is probably built on this chunking process.

Delta band oscillations are invariant to acoustic details on timescales < 500 ms

We found that only the stimuli with exponent 2 evoked robust phase coherence in the delta band, which supplements the findings by Garcia-Lazaro *et al.* (2006). The mutual information results in the delta band do not vary with the filter sizes used in the auditory processing model. This surprising result further suggests that the delta band oscillations are probably not sensitive to low-level acoustic details, but probably to a high-level perceptual cues, such as linguistic structure in speech (Ding *et al.*, 2015), and attention-related rhythmic processing (Lakatos *et al.*, 2008; Schroeder *et al.*, 2010).

Memory and attention as potential confounds

As the 'frozen' stimuli were repeatedly presented to the participants while each of the 'distinct' stimuli was only presented once, one might surmise that the participants are able to memorize the 'frozen' stimuli. Previous studies have shown, though, that it is challenging,

and perhaps even impossible, for humans to memorize acoustic local details of sounds textures of more than 200 ms long (McDermott *et al.*, 2013; Teng *et al.*, 2016), well short of the 3-s length of our stimuli. (Obviously, speech or music can be encoded and recalled.) As the 'frozen' and the 'distinct' stimuli were comparable in terms of long-term acoustic properties, such as spectral modulation and spectrum, the participants had to remember the acoustic details to be able to tell apart the 'frozen' stimuli from the 'distinct' stimuli with corresponding exponents. It would be very challenging indeed for the participants to differentiate one 'frozen' stimulus out of 25 'distinct' stimuli with similar long-term acoustic properties. If the participants could successfully identify each 'frozen' stimulus, we would not expect memory to be affected by the exponent of frequency modulation, as we have found here. Therefore, we conjecture that memory does not contribute significantly to our results.

With regard to attention, as we only presented the target tones in the 'distinct' stimuli, so it would be possible that the participants could choose to only attend to the distinct stimuli to detect the tone. If the participants could distinguish the 'frozen' and 'distinct' stimuli by memorizing the 'frozen' stimuli and figure out that tones are contained only in each of the distinct stimuli (we did not tell the participants this information), we would expect that the tone detection performance should be similar across all exponents, as the participants could simply choose the 'distinct' stimuli as the target. But we did find a difference of tone detection across different exponents, and this difference, importantly, can be explained by our acoustic analysis on local SNRs (Fig. 2).

Therefore, although it is true that memory and attention are always relevant considerations, the effects caused by memory and attention are unlikely to form the explanatory basis of our main results.

Conclusion: active chunking on a timescale of ~200 ms in the auditory system

Our results demonstrate an active chunking scheme in the auditory system (Poeppel, 2003; Ghitza & Greenberg, 2009; Panzeri *et al.*, 2010; Ghitza, 2012; Giraud & Poeppel, 2012; VanRullen, 2016): On the timescale of ~200 ms, the auditory system actively groups acoustic information to parse a continuous acoustic stream into segments. The robust phase coherence in the theta band is not solely driven by external stimuli but also reflects active chunking. This chunking scheme is prevalent in auditory processing of sounds of various dynamics and may serve as a fundamental step for further perceptual analysis.

Acknowledgements

We thank Jeff Walker for his technical support and Adeen Flinker for his help with analyzing the stimuli.

Funding

This work was supported by the National Institutes of Health (2R01DC05660 to DP); the Major Projects Program of the Shanghai Municipal Science; Technology Commission (15JC1400104 and 17JC1404104 to XT); National Natural Science Foundation of China (31500914 to XT); and Program of Introducing Talents of Discipline to Universities (Base B16018 to XT).

Conflict of interest

The authors declare no competing financial interests.

Author contributions

XTeng designed the experiment, collected and analyzed data, and drafted the manuscript. XTian, KD, and DP interpreted the data, edited the manuscript, and provided critical revisions. DP supervised the project. All authors approved the final version.

Data accessibility

All behavioral data and part of neurophysiology data underlying this article can be accessed on Edmond http://edmond.mpdl.mpg.de/imeji/collection/kZalRMtxa19mlRyG and used under the Creative Commons Attribution license. Complete neurophysiology data are available under request.

Abbreviations

ACI, acoustic change index; dB, decibel; dITPC, differentiated ITPC; ERB, equivalent rectangular bandwidth; ITPC, Inter-trial phase coherence; MEG, magnetoencephalography; MI, mutual information; MPS, modulation power spectra; RMS, root mean square; SNR, signal-to-noise ratio; SPL, sound pressure level.

References

Andoni, S., Li, N. & Pollak, G.D. (2007) Spectrotemporal receptive fields in the inferior colliculus revealing selectivity for spectral motion in conspecific vocalizations. *J. Neurosci.*, **27**, 4882–4893.

Andrillon, T., Kouider, S., Agus, T. & Pressnitzer, D. (2016) Perceptual learning of acoustic noise generates memory-evoked potentials. *Curr. Biol.*, **25**, 2823–2829.

Brungart, D.S. (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.*, **109**, 1101–1109.

Carlson, N.L., Ming, V.L. & Deweese, M.R. (2012) Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput. Biol.*, **8**, 1–15.

de Cheveigné, A. & Simon, J.Z. (2007) Denoising based on time-shift PCA. *J. Neurosci. Meth.*, **165**, 297–305.

Cogan, G.B. & Poeppel, D. (2011) A mutual information analysis of neural coding of speech by low-frequency MEG phase information. *J. Neurophysiol.*, **106**, 554–563.

Ding, N. & Simon, J.Z. (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.*, **33**, 5728–5735.

Ding, N. & Simon, J.Z. (2014) Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci*, **8**, 1–7.

Ding, N., Melloni, L., Zhang, H., Tian, X. & Poeppel, D. (2015) Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci.*, **19**, 158–164.

Ding, N., Patel, A.D., Chen, L., Butler, H., Luo, C. & Poeppel, D. (2017) Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.*

Doelling, K.B. & Poeppel, D. (2015) Cortical entrainment to music and its modulation by expertise. *Proc. Natl. Acad. Sci. USA*, **112**, E6233–E6242.

Doelling, K.B., Arnal, L.H., Ghitza, O. & Poeppel, D. (2014) Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, **85**, (Part 2 IS)761–768.

Elliott, T.M. & Theunissen, F.E. (2009) The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.*, **5**, 1–14.

Escabí, M.A., Miller, L.M., Read, H.L. & Schreiner, C.E. (2003) Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J. Neurosci.*, **23**, 11489–11504.

Garcia-Lazaro, J.A., Ahmed, B. & Schnupp, J.W.H. (2006) Tuning to natural stimulus dynamics in primary auditory cortex. *Curr. Biol.*, **16**, 264–271.

Ghitza, O. (2012) On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Front Psychol.*, **3**, 238.

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**, 113–126.

Giraud, A.-L. & Poeppel, D. (2012) Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.*, **15**, 511–517.

Glasberg, B.R. & Moore, B.C.J. (1990) Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, **47**, 103–138.
Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P. & Garrod, S. (2013) Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol.*, **11**, e1001752.
He, B.J. (2014) Scale-free brain activity: past, present, and future. *Trends Cogn. Sci.*, **18**, 1–8.
He, B.J., Zempel, J.M., Snyder, A.Z. & Raichle, M.E. (2010) The temporal structures and functional significance of scale-free brain activity. *Neuron*, **66**, 353–369.
Henry, M.J. & Obleser, J. (2012) Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc. Natl. Acad. Sci. USA*, **109**, 20095–20100.
Henry, M.J., Herrmann, B. & Obleser, J. (2014) Entrained neural oscillations in multiple frequency bands comodulate behavior. *Proc. Natl. Acad. Sci. USA*, **111**, 14935–14940.
Herrmann, B., Henry, M.J., Grigutsch, M. & Obleser, J. (2013) Oscillatory phase dynamics in neural entrainment underpin illusory percepts of time. *J. Neurosci.*, **33**, 15799–15809.
Kayser, S.J., Ince, R.A.A., Gross, J. & Kayser, C. (2015) Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *J. Neurosci.*, **35**, 14691–14701.
Kerlin, J.R., Shahin, A.J. & Miller, L.M. (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J. Neurosci.*, **30**, 620–628.
Kidd, G. Jr, Mason, C.R., Richards, V.M., Gallun, F.J. & Durlach, N.I. (2007). Informational Masking. In Yost, W.A., Popper, A.N. & Fay, R.R. (Eds), *Auditory Perception of Sound Sources, Springer Handbook of Auditory Research*. Springer, Boston, MA, pp. 143–189.
Lachaux, J.-P., Rodriguez, E., Martinerie, J. & Varela, F.J. (1999) Measuring phase synchrony in brain signals. *Hum. Brain Mapp.*, **8**, 194–208.
Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I. & Schroeder, C.E. (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, **320**, 110–113.
Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C. & Schroeder, C.E. (2013) The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, **77**, 750–761.
Luo, H. & Poeppel, D. (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, **54**, 1001–1010.
Luo, H. & Poeppel, D. (2012) Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. *Front Psychol.*, **3**, 170.
Luo, H., Tian, X., Song, K., Zhou, K. & Poeppel, D. (2013) Neural response phase tracks how listeners learn new acoustic representations. *Curr. Biol.*, **23**, 968–974.
Machens, C.K., Wehr, M.S. & Zador, A.M. (2004) Linearity of cortical receptive fields measured with natural sounds. *J. Neurosci.*, **24**, 1089–1100.
Macmillan, N.A. & Creelman, C.D. (2004) *Detection Theory: A User's Guide*. Hove, UK: Psychology Press.
Magri, C., Whittingstall, K., Singh, V., Logothetis, N.K. & Panzeri, S. (2009) A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. *BMC Neurosci.*, **10**, 81.
Mazaheri, A. & Jensen, O. (2006) Posterior α activity is not phase-reset by visual stimuli. *Proc. Natl. Acad. Sci. USA*, **103**, 2948–2952.
Mazaheri, A. & Jensen, O. (2010) Rhythmic pulsing: linking ongoing brain activity with evoked responses. *Front. Hum. Neurosci.*, **4**, 177.
Mazaheri, A. & Picton, T.W. (2005) EEG spectral dynamics during discrimination of auditory and visual targets. *Cognitive Brain Res.*, **24**, 81–96.
McDermott, J.H., Schemitsch, M. & Simoncelli, E.P. (2013) Summary statistics in auditory perception. *Nat. Neurosci.*, **16**, 493–498.
Montemurro, M.A., Senatore, R. & Panzeri, S. (2007) Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Comput.*, **19**, 2913–2957.
Ng, B.S.W., Logothetis, N.K. & Kayser, C. (2013) EEG phase patterns reflect the selectivity of neural firing. *Cereb. Cortex*, **23**, 389–398.
Oldfield, R.C. (1971) The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, **9**, 97–113.
Olshausen, B.A. & Field, D.J. (2004) Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.*, **14**, 481–487.
Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. (2011) Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.*, **2011**, 1–9.
Panzeri, S., Senatore, R., Montemurro, M.A. & Petersen, R.S. (2007) Correcting for the sampling bias problem in spike train information measures. *J. Neurophysiol.*, **98**, 1064–1072.
Panzeri, S., Brunel, N., Logothetis, N.K. & Kayser, C. (2010) Sensory neural codes using multiplexed temporal scales. *Trends Neurosci.*, **33**, 111–120.
Peelle, J.E., Gross, J. & Davis, M.H. (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex*, **23**, 1378–1387.
Poeppel, D. (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time". *Speech Commun.*, **41**, 245–255.
Pola, G., Thiele, A., Hoffmann, K.P. & Panzeri, S. (2003) An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network*, **14**, 35–60.
Prins, N. & Kingdom, F.A.A. (2009) Palamedes: Matlab routines for analyzing psychophysical data. Available http://www.palamedestoolbox.org.
Riecke, L., Sack, A.T. & Schroeder, C.E. (2015) Endogenous delta/theta sound-brain phase entrainment accelerates the buildup of auditory streaming. *Curr. Biol.*, **25**, 3196–3201.
Roberts, T.P., Ferrari, P., Stufflebeam, S.M. & Poeppel, D. (2000) Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights toward perception. *J. Clin. Neurophysiol.*, **17**, 114–129.
Schroeder, C.E., Wilson, D.A., Radman, T., Scharfman, H. & Lakatos, P. (2010) Dynamics of Active Sensing and perceptual selection. *Curr. Opin. Neurobiol.*, **20**, 172–176.
Shahin, A.J., Roberts, L.E., Miller, L.M., McDonald, K.L. & Alain, C. (2007) Sensitivity of EEG and MEG to the N1 and P2 auditory evoked responses modulated by spectral complexity of sounds. *Brain Topogr.*, **20**, 55–61.
Singh, N.C. & Theunissen, F.E. (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.*, **114**, 3394–3411.
Søndergaard, P.L. & Majdak, P. (2013). The Auditory Modeling Toolbox. In Blauert, J. (Ed), *The Technology of Binaural Listening*. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 33–56.
Stilp, C.E. & Kluender, K.R. (2010) Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proc. Natl. Acad. Sci. USA*, **107**, 12387–12392.
Stilp, C.E., Kiefte, M., Alexander, J.M. & Kluender, K.R. (2010) Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences a. *J. Acoust. Soc. Am.*, **128**, 2112–2126.
Teng, X., Tian, X. & Poeppel, D. (2016) Testing multi-scale processing in the auditory system. *Sci. Rep.*, **6**, 34390EP.
Theunissen, F.E. & Elie, J.E. (2014) Neural processing of natural sounds. *Nat. Rev. Neurosci.*, **15**, 355–366.
Theunissen, F.E., Sen, K. & Doupe, A.J. (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci.*, **20**, 2315–2331.
VanRullen, R. (2016) Perceptual cycles. *Trends Cogn. Sci.*, **20**, 1–13.
VanRullen, R. & Koch, C. (2003) Is perception discrete or continuous? *Trends Cogn. Sci.*, **7**, 207–213.
VanRullen, R., Zoefel, B. & Ilhan, B.A. (2014) On the cyclic nature of perception in vision versus audition. *Philos. Tr. R. Soc. Lond. B Biol. Sci.*, **369**, 20130214.
Voss, R.F. & Clarke, J. (1978) "1/f noise" in music: music from 1/f noise. *J. Acoust. Soc. Am.*, **63**, 258–263.
Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R.G. *et al.* (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, **77**, 980–991.