NeuroImage 66 (2013) 318–332

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Processing of speech and non-speech sounds in the supratemporal plane: Auditory input preference does not predict sensitivity to statistical structure

P. Tremblay [a,b,*], M. Baroni [c,d], U. Hasson [c,e]

[a] Université Laval, Rehabilitation Department, Québec City, Qc., Canada
[b] Centre de Recherche de l'Institut Universitaire en santé mentale de Québec (CRIUSMQ), Québec City, Qc., Canada
[c] Center for Mind/Brain Sciences (CIMeC), University of Trento, via delle Regole, 1010, 38060, Mattarello (TN), Italy
[d] Department of Information Science, University of Trento, via delle Regole, 1010, 38060, Mattarello (TN), Italy
[e] Department of Psychology and Cognitive Sciences, University of Trento, via delle Regole, 1010, 38060, Mattarello (TN), Italy

ARTICLE INFO

Article history:
Accepted 15 October 2012
Available online 29 October 2012

Keywords:
Statistical regularities
Supratemporal plane
Language
Speech processing

ABSTRACT

The supratemporal plane contains several functionally heterogeneous subregions that respond strongly to speech. Much of the prior work on the issue of speech processing in the supratemporal plane has focused on neural responses to single speech vs. non-speech sounds rather than focusing on higher-level computations that are required to process more complex auditory sequences. Here we examined how information is integrated over time for speech and non-speech sounds by quantifying the BOLD fMRI response to stochastic (non-deterministic) sequences of speech and non-speech naturalistic sounds that varied in their statistical structure (from random to highly structured sequences) during passive listening. Behaviorally, the participants were accurate in segmenting speech and non-speech sequences, though they were more accurate for speech. Several supratemporal regions showed increased activation magnitude for speech sequences (preference), but, importantly, this did not predict sensitivity to statistical structure: (i) several areas showing a speech preference were sensitive to statistical structure in both speech and non-speech sequences, and (ii) several regions that responded to both speech and non-speech sounds showed distinct responses to statistical structure in speech and non-speech sequences. While the behavioral findings highlight the tight relation between statistical structure and segmentation processes, the neuroimaging results suggest that the supratemporal plane mediates complex statistical processing for both speech and non-speech sequences and emphasize the importance of studying the neurocomputations associated with auditory sequence processing. These findings identify new partitions of functionally distinct areas in the supratemporal plane that cannot be evoked by single stimuli. The findings demonstrate the importance of going beyond input preference to examine the neural computations implemented in the superior temporal plane.

Introduction

The human brain, particularly in the supratemporal plane, is uniquely sensitive to the complex acoustical properties that are the hallmark of the speech signal. The supratemporal plane is indeed sensitive to many features of the speech signal, such as vowel formant frequencies (Formisano et al., 2008; Naatanen et al., 1997; Obleser et al., 2003, 2006; Poeppel et al., 1997; Reil, 1809), and consonants' spectro-temporal composition (Obleser et al., 2007; Raizada and Poldrack, 2007). The supratemporal plane can also distinguish between phonological and non-phonological acoustical differences (Formisano et al., 2008; Jacquemot et al., 2003; Obleser et al., 2010; Staeren et al., 2009). This remarkable sensitivity has led some researchers to argue for some level of speech-specialization in non-primary areas of the supratemporal plane.

Much of this prior work has asked whether different types of sounds are processed in dedicated (sometimes referred to as 'specialized') brain areas in the supratemporal plane focusing on non-primary auditory cortex. However, a different approach to studying speech and non-speech processing in the supratemporal plane is to quantify the extent to which subregions of the supratemporal plane are specialized in terms of the type of neurocomputations they implement, rather than in terms of the kind of stimulus they respond to more strongly ('prefer'). While the speech signal is undoubtedly a very unique input, its individual components share similarities with other types of signal. For instance, musical instruments and animal vocalizations have complex spectral features. Moreover, sequential organization, which is a hallmark of speech, is shared by other types of sensory signal such as bird songs and music. For speech, sequential organization is based on complex language-specific (i.e. phonotactic) constraints, which are manifested in language-specific *statistics*. Specifically, syllables with

* Corresponding author at: Université Laval, Département de Réadaptation, Centre de Recherche de l'Institut Universitaire en santé mentale de Québec (CRIUSMQ), 2601 rue de la Canardière, office F-2445, Québec (Québec), Canada G1J 2G3.
E-mail address: Pascale.tremblay@fmed.ulaval.ca (P. Tremblay).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2012.10.055

high transition probabilities (TP) – the probability of transitioning from one state or one syllable to another in a single step – tend to form words whereas those with lower TPs mark word boundaries. Because fluent speech does not contain invariant acoustical cues marking word boundaries (Cole and Jakimik, 1980; Lehiste and Shockey, 1972), access to statistical information, in particular TP, is critical for learning how to segment continuous speech into its constituent units (i.e. syllables and words), especially during early childhood where words boundaries are unknown. Behavioral experiments have shown that infants, but also adults, are sensitive to TP in speech (Newport and Aslin, 2004; Pelucchi et al., 2009a,2009b; Pena et al., 2002; Saffran et al., 1996, 1999), as well as in non-speech auditory and visual stimuli (Fiser and Aslin, 2001, 2002; Kirkham et al., 2002; Saffran et al., 1999, 2007). In adulthood, statistical information can be used to predict upcoming syllables and words. To illustrate, if a given syllable "x" can only be followed by syllable "y", then it follows that hearing "x" offers useful information for predicting "y". This information can therefore be used to disambiguate speech in degraded listening situations, or in the face of an unfamiliar accent. Formal architectures that describe how predictive codes can be implemented at a neural level have been gaining prominence in various domains of perceptual learning (Friston and Kiebel, 2009; Rao and Ballard, 1999).

Despite the richness of the statistical information present in speech, and the fact that humans exhibit sensitivity to this information throughout the lifespan, there has not been much neurobiological research on the neural underpinning of statistical information processing. Little is known about temporal regions sensitive to statistical information in the speech signal, and no study as directly compared statistical information processing for speech and non-speech inputs. Understanding this issue is the main motivation of the current study. Prior magnetoencephalographic (MEG) studies demonstrated sensitivity to statistical information in the right temporal-parietal junction, including the most posterior part of the supratemporal plane in a tone sequence (Furl et al., 2011) when statistical information is quantified via TP. Consistent with this finding, Overath et al. (2007) found that the bilateral planum temporale (PT) tracks statistical properties in tone series when statistics are quantified via Sample Entropy (Overath et al., 2007). Specifically for speech, enhanced activation in the posterior supratemporal plane has been found in children and adults processing sequences of random syllables vs. sequences with fixed syllable triads[1] (McNealy et al., 2006, 2010). Recently, fMRI studies have shown (i) sensitivity to TP structure in sequences of pure tones in the parietal operculum was found, with higher activity for both random and highly structured series than for mid-ordered series (Tobia et al., 2012a), and (ii) sensitivity to perceived changes in sequence regularity in a region of left superior temporal gyrus (STG) (Tobia et al., 2012b).

These prior findings suggest that sensitivity to statistical information in speech and non-speech inputs is present at the level of the supratemporal plane. A fundamental yet unanswered issue is whether statistical information for auditory inputs is processed in a domain-general or domain-specific manner. To address this issue we examined the neural basis of statistical information processing in speech and non-speech complex natural sound sequences using functional MRI. Departing from prior work, we did not include deterministic sequences or fixed combinations ("words") in our series, but manipulated the statistical structures of sequences so that these ranged from random (i.e. there is no statistical structure) to highly predictable though not deterministic (i.e. sequences for which statistical information can be used to build a probabilistic internal representation of the sequence structure). Our first hypothesis was that core supratemporal regions including the transverse temporal gyrus (TTG) and PT would respond to both speech and non-speech sounds. However, we expected only PT to track the statistical structure of auditory inputs (based on Overath et al., 2007). Our second prediction was that BOLD (Blood Oxygen-Level Dependent) signal would vary as a function of the level of statistical structure for both speech and non-speech reflecting domain-general statistical information processing mechanisms. However, because statistics can only be computed on the basis of an a-priori internal representation of the unique elements present in a signal established through segmentation, we also expected that statistical processing of speech sequences would enjoy a significant advantage since its constituent units (syllables in current study) are known a-priori. This advantage could take the form of a lower response magnitude in a subset of supratemporal areas, or more spatially localized activation patterns for speech reflecting reduced processing effort. In contrast, when processing less familiar signals such as a non-native language or sequences of environmental sounds, segmentation may be more demanding since processing sequences consisting of unknown elements may necessitate identifying the independent elements first. To address this issue, in addition to examining differences in BOLD signal, we also conducted two behavioral experiments (i) to examine the discriminability of both speech and non-speech sounds, and (ii) to quantify the participants' ability to segment speech and non-speech auditory sequences. We predicted that the speech sequences would be more easily segmented because of their highly familiar nature; a prediction that was verified. From a functional perspective, addressing these issues would answer the question of whether speech enjoys a unique status with respect to statistical processing in the supratemporal plane. From a neurobiological perspective, this design allowed us to identify functional regions in the supratemporal plane that differ in the extent to which they are sensitive to statistical structure in auditory input.

Materials and methods

Participants

The participants were 20 healthy right-handed (Oldfield, 1971) native Italian speakers (9 females; 24 ± 4.5 years, education: 17.1 ± 3.64 years), with normal self-reported hearing, and no history of language or neurological/neuropsychological disorders. The study was approved by the Human research ethics committee of the University of Trento in Italy.

Stimuli

Two types of complex acoustic stimuli were created: speech and non-speech, which were equalized in duration (225 ms), sampling rate (44 kHz), envelope (225 ms, ±15 ms fade in, ±15 ms fade out) and root mean square (RMS) intensity.

Speech stimuli

The speech stimuli consisted of 70 non-meaningful Italian consonant-vowel combinations (CV) (e.g. [ba]). A list of all syllables is provided as Supplementary material S1. These syllables were chosen based on their distributional properties, which were extracted from itWaC, a large corpus (~1.5 billion words) of Italian Web pages (Baroni et al., 2009). The itWaC corpus was automatically transcribed and syllabified using a phonetic lexicon containing transcriptions of about 400,000 Italian words (100 words commonly occurring in itWaC but not in the original lexicon were added) (Cosi and Avesani, 2001). Words not present in this lexicon were discarded. Word-final consonants (and consonant clusters) that were followed by a word-initial vowel were syllabified as onsets with the vowel (e.g. /un amico/ was syllabified as [u.na.mi.ko] consistent with Italian phonotactics). The resulting transcribed corpus contained more than 3 billion syllables; syllable frequency information was extracted from this corpus to choose syllables to be used in this experiment. A

[1] This manipulation can be considered an extreme variant of manipulation of statistical features since the fixed-syllable condition is one associated with transition constraints whereas the random one is not.

total of 5405 unique CV syllables were coded in the database. Of the syllables that were chosen to form our corpus, the mean rank order was 5165, and the median 5290, representing the 5% most frequent syllables. The selected syllables were 3 orders of magnitude more frequent than the mean: the mean/median log10 frequency for these syllables was 6.58 and 6.61 respectively while the mean/median for the entire dataset was 3.55 and 3.36 respectively. Hence, the selected syllables were highly frequent, non-meaningful, Italian consonant vowel (CV) syllables. These syllables were composed of combinations of five vowels (/a, e, i, o, u/) and twenty-four consonants.[2]

A native male Italian speaker from the North of Italy pronounced these syllables in a sound-attenuated booth. Each syllable was produced at least five times each, always within a carrier sentence ("adesso dico [ba]"; translation: "now I say [ba]"). The best token of each syllable was used in the experiment, such that each syllable in the study was represented with a single token (following the procedure used by Buiatti et al., 2009; McNealy et al., 2006). The syllables were recorded at 44 kH using a unidirectional microphone connected to a professional amplifier, saved directly to disk using Sound Studio 3.5.4 (Felt Tip Software, NY, USA), edited offline using Wave Pad Sound Editor 4.53 (NHC Software, Canberra, Australia) to have a mean duration of 225 ms, and normalized for RMS intensity.

Non-speech stimuli

The non-speech stimuli were 70 unique bird sounds created from the recordings of thirty seven (37) different birds (including raven, parrot, duck, heron, falcon, and starling; see Supplementary material S2 for the complete listing). We chose bird sounds because they form a natural class of sounds, and because, like speech, they are acoustically rich exhibiting a formant structure (see Supplementary material S3 for an illustration of the bird sounds), and characterized by fast modulations of spectral power over time. The bird stimuli were created from a high quality digital collection of bird sounds recorded at 44 kHz, and commercially available on ITunes (The Ultimate Sound Effects Collection: Birds; 2010 by HDsoundFX). All sounds were edited using Wave Pad Sound Editor to have a mean duration of 225 ms and normalized for RMS intensity. 75% of the total energy (power) contained in the bird sounds was contained between 1 and 5 kHz, with roughly equal contribution of each (1 kHz) frequency bin (1–2 kHz, 2–3 kHz, 3–4 kHz and 4–5 kHz).

Spectro-temporal characteristics of the acoustic stimuli

In order to characterize the speech and non-speech sounds in terms of their spectro-temporal characteristics, we computed the spectral entropy of each sound, a measure that reflects the relative complexity (entropy) of a sound's frequency spectrum on a scale of 0 to 1. To illustrate, contrary to speech, pure tones have a very focal distribution of energy with a spectral entropy around 0. This may account for observed speech vs. pure-tone difference found in prior work (Benson et al., 2001). In contrast, more complex "noisy" or broadband sounds (in which energy is evenly distributed across all frequencies) have a higher entropy value around 1. For each sound used in this experiment, spectral entropy was calculated using the spectral entropy function of the Seewave R package (Sueur et al., 2008). For both speech and non-speech, spectral entropy was high, with a mean ± SD spectral entropy of .70 ± .069 for speech, and a mean entropy of .79 ± .076 for non-speech. For the speech sounds, on average, over 99% of the total spectral energy was in the 0–2 kHz frequency range. For the non-speech sounds, roughly 30% of the total spectral energy was in this range, indicating that, though non spectrally identical, there was significant overlap in the spectral composition of the speech and non-speech sounds. 83% of the total spectral energy in the non-speech sounds was located between 0 and 5 kHz. An example of a speech and a non-speech sequence (waveforms and spectrograms) is provided as Supplementary material S3.

Construction of the sound sequences and formal validation of order manipulation

The stimuli (speech, non-speech) were arranged in 126 unique sequences containing combinations of four different sounds (either four syllables or four bird sounds, taken from the pool of 70 speech and 70 non-speech sounds) repeated ±8 times each according to transition constraints as detailed below. Each sequence was 8.8 s long and consisted of 32 sounds presented at a rate of 3.6 Hz, consistent with normal speech production rate (Kent, 2000; Rosen, 1992). Sounds within a sequence were separated by 50 ms of silence. To optimize within-sequence syllable discriminability, each speech sequence included syllables consisting of 4 different consonants and 4 different vowels.[3] The speech sequences were verified by a two native Italian speakers to ensure that they did not contain meaningful syllable combinations. To optimize discriminability for the non-speech sounds, no two sounds from the same bird species were presented within a sequence. In order to ensure that the speech and non-speech sounds had similar discriminability, a behavioral experiment was conducted which is described in Section Behavioral auditory discrimination study.

The statistical structure (SS) of the sequences, which was determined by the TP matrices, was manipulated experimentally. The sequences ranged from random to highly structured in 3 levels (low SS, mid SS, and high SS), each associated with a different level of Markov Entropy (ME). ME is a measure of unpredictability: the more predictable a sequence is, the lower its ME value. TP matrices are presented in Supplementary material S4. As can be seen in Table S4a, in the low structure condition, each item was equally likely to appear at any point independently of the previously presented item (mean ± SD entropy = 1.84 ± .03; range: 1.8–1.9[4]). It was therefore impossible for the participants to form expectations. As shown Table S4b and S4c, in the mid structure (mean ± SD entropy = 1.51 ± .018; range 1.48–1.53) and high structure sequences (mean ± SD entropy = .79 ± .012; range .78–.81), the TP matrices were more constrained, which allowed the participants to form expectations about upcoming sounds. The overall proportion of self-repetitions (the diagonal of the matrices presented in Table S2a) was set at 25%. This was done to control for repetition suppression effects (Hasson et al., 2006).

To ensure that the speech sequences only differed in terms of experimentally-determined TP (and not in terms of naturalistic, Italian TP), we extracted from the itWaC corpus the mean TP for all syllable combinations presented within our sequences for each condition (low, mid, and high statistical structure). The average Italian within-sequence TPs P(syl1 | syl2) were low in all three conditions (high structure = .00451 ± .01 SD; mid structure = .004562 ± .005 SD; random = .0056 ± .006 SD). Low TP indicates rarely occurring syllables combinations, which was expected since we did not use words. Importantly, they did not differ across conditions (p values for all 3 contrasts > .58). This experimental setup resulted in a 3 × 2 design with Statistical Structure (3 levels) and auditory Category (2 levels) as within-subject factors, for a total of 6 conditions and 126 experimental trials. Two sequences were excluded from the analyses to balance the number of self-repetitions across conditions, leaving 123 trials for each participant. The experimental trials were split between 3 runs of approximately 10 min each.

To ensure that subjects could acquire the statistical features of the mid and high structure sequences, and to determine the point at which these sequences diverged in their relative departure from randomness, a formal analysis of these sequences was conducted that quantified the rate at which the transition structure of these

[2] ([b, d, z, d͡z, d͡ʒ, f, g, k, j, m, n, ɲ, l, ʎ, p, r, s, ʃ, t͡s, t͡ʃ, v, w, z]).

[3] e.g. /be-jo-mu-zi/.

[4] Note that the maximum entropy for 4-item sequences, i.e., the random case, is 2 and can only be achieved in very long series.

sequences departed from the null (uniform, random) transition distribution through the progressions of each sequence. This was done by implementing a moving window analysis on each sequence that (a) determined the transition probability structure existing up to each point in the sequence, and (b) quantified the degree to which that transition structure diverged from what would be expected in a random distribution, using the Kullback–Leibler (KL) divergence measure (Kullback and Leibler, 1951). This analysis is described in Supplementary material S5. Importantly, the results showed that, as was expected, mid and highly structured sequences differed in terms of their learning potential from a random distribution.

Presentation of the stimuli

Each sound sequence was presented only once, through a high quality, digital passive noise-attenuation MRI-compatible stereo headset (SereneSound, Resonance Technology Inc). Within each of the three runs, 42 experimental trials, each 8.8 s (corresponding to the duration of the sound sequences) were interleaved with 49 "jittered" short silence (rest) intervals. These rest intervals had a mean duration of 4.2 s, and accounted for a total time of 206 s (per run). Within each run, the order of the conditions and the number of rest trials were optimized (randomized) using OPTseq2 (http://surfer.nmr.mgh.harvard.edu/optseq/).

In-scanner behavioral task

Before beginning the experiment, the participants were introduced to the syllables and birds sounds. This was done to avoid any surprise effect, as well as to familiarize the participants with the speaker's voice and accent. Task instructions were pre-recorded by the same speaker who recorded the stimuli and played back to all participants before beginning the fMRI session. The participants were asked to passively attend to the auditory sequences while monitoring a static polygon shape presented on the screen via back projection. This visual monitoring task required them to press a button on a response box each time the polygon started rotating in a clockwise fashion ("catch trials"), which ensured that the participants maintained vigilance but did not necessitate attention to the auditory sequences or the SS manipulation. There were a total of 24 catch trials, representing approximately 20% of all experimental trials. Responses to catch trials were taken into account in the fMRI analyses (i.e. modeled as a regressor of no interest), but these are not reported here as they have no theoretical relevance.

Post-scan behavioral task

At the end of the fMRI session, the participants were presented with all 123 auditory sequences again and were asked to rate their perceived statistical structure using a 7-point scare, at self-paced rate.[5] In addition, in about 40% of all trials, they were also asked to indicate the number of perceived sounds. The order of presentation of the six conditions was fully randomized. This portion of the study took about 45 min.

Image acquisition

A 4 T 8-coil Bruker system was used to acquire high-resolution anatomical and functional data for each participant. Structural scans were acquired with a 3D T1-weighted MPRAGE sequence (TR/TE = 2700/4 ms, flip angle = 7°, isotropic voxel size = 1 mm^3, matrix = 256 × 224; 176 sagittal slices). Two structural volumes were obtained for all but two participants and averaged to allow more accurate image processing. Single-shot EPI BOLD functional images were acquired using the point-spread-function distortion correction method (Zaitsev et al., 2004). Each functional EPI run began with six dummy scans to allow the magnetization to stabilize to a steady state. Eight hundred and seventy functional images were acquired across 3 runs for this experiment (TR/TE = 2200/33 ms, 37 interleaved slices parallel to AC/PC, voxel size = 3 × 3 × 3.45, gap = .2 mm; matrix = 64 × 64; 1914 s of the scan time).

Data analysis

Behavioral data analyses

Behavioral data was obtained in the post-scan stage as described in Section Post-scan behavioral task. Two dependent behavioral measures were obtained for each sequence — perceived statistical structure and perceived numerosity (number of unique sounds). Both measures were analyzed and compared for speech and non-speech across all levels of statistical structure levels using a 2 × 3 repeated measure ANOVA with Category, and statistical structure as the within-subject factors. Trend analyses were also conducted to examine whether these measures changed as a function of increased statistical structure. In addition to these analyses, we also examined the relationship between perceived numerosity and perceived statistical structure using a mediation analysis implemented using the INDIRECT algorithm (Hayes, 2008; Preacher and Hayes, 2004). Mediation analyses determine whether mediator variables affect the relationship between a dependent (Y) and an independent (X) variable, thereby serving to clarify the nature of the relationship between independent and dependent variables. This analysis was used to test (i) the potential causal relationship between auditory category (X) and the perception of statistical structure (Y), and (ii) whether this relationship is mediated by numerosity perception (M).

fMRI time-series analyses

All functional time-series were motion-corrected, time-shifted, de-spiked and mean-normalized using AFNI (Cox, 1996). In addition, we censored time points occurring during excessive motion, defined as >1 mm (Johnstone et al., 2006). For each participant we first regressed the mean, linear, and quadratic trend components as well as the 6 motion parameters (x, y, z and roll, pitch and yaw) separately for each experimental run. The resulting cleaned time-series were projected onto the 2-dimensional surfaces where all subsequent processing took place. For each participant, the anatomical images were aligned to the functional volumes automatically (Saad et al., 2009) and alignment was verified and manually adjusted when necessary. A surface representation of the participant's anatomy was then created using FreeSurfer (Dale et al., 1999; Fischl et al., 1999) by inflating each hemisphere of the anatomical volumes to a surface representation and aligning it to a template of average curvature. SUMA was used to import the surface representations into the AFNI 3D space and to project the pre-processed time-series from the 3-dimensional volumes onto the 2-dimensional surfaces. The time-series were smoothed on the surface to achieve a target smoothing value of 6 mm using a Gaussian full width half maximum (FWHM) filter. Smoothing on the surface as opposed to the volume ensures that white matter values are not included, and that functional data located in anatomically distant locations on the cortical surface are not averaged across sulci (Argall et al., 2006). For each participant, we created a set of 7 regressors, one for each of the experimental condition and one for the catch trials. A finite impulse response linear least squares model established a fit to each time point of this hemodynamic response function (HRF) for each of these conditions using AFNI's tent basis function. This model-free deconvolution method allows the shape of the hemodynamic response to vary for each condition rather than assuming a single response profile for all conditions (Meltzer et al., 2008). The interval

[5] Participants were asked to judge the structure of the sequences, that is, how ordered or regular the sequences were. For instance, they were told that sequences containing repeating, predictable patterns (the following example PA-TA-KA-PA-TA-KA-PA-TA-KA was given) should be rated high (close to seven), and random, unpredictable sequences should be rated low (close to zero).

Table 1
Definition of the supratemporal regions of interest.

Region	Short name	Definition
Planum polare	PP	Unedited FreeSurfer ROI. PP is bounded medially by the circular sulcus of the insula, caudally by TTG, laterally by STG and rostrally by the temporal pole.
Superior temporal gyrus	STG	The FreeSurfer STG ROI, which runs from the rostral edge of the STS to the supramarginal gyrus; it is bounded medially by the lateral fissure; we manually divided this region into thirds of roughly the same size along a rostro-caudal axis.
Transverse temporal gyrus	TTG	The FreeSurfer TTG ROI, which is bounded rostrally by the rostral extent of the transverse temporal sulcus, caudally by the caudal portion of the insular cortex, laterally by the superior temporal gyrus and medially by the lateral fissure, was manually divided into halves of roughly the same size along a medial-lateral axis.
Transverse temporal sulcus	TTS	The FreeSurfer TTS ROI, located immediately anterior to PT and posterior to TTG, was manually divided into halves of roughly the same size along a medial-lateral axis.
Planum temporale	PT	The FreeSurfer PT ROI, defined as the part of the superior temporal plane immediately posterior to the transverse temporal sulcus, bounded medially by the Sylvian fissure, and posteriorly by the supramarginal gyrus; we manually divided this region into three segments of roughly the same size along a rostro-caudal axis.
Caudal segment of the Sylvian fissure	SF	The Freesurfer posterior Sylvian fissure ROI runs from the lower end of the central sulcus to the end of the posterior ascending ramus (Dahl et al., 2006). The FreeSurfer posterior SF ROI was manually subdivided into two segments (anterior, posterior) of roughly the same size.

that we modeled subsumed the entire trial duration (i.e. 8.8 s), beginning at stimulus onset and continuing at 2.2-s intervals for 17.6 s. Typically, shorter intervals are modeled in fMRI analyses, but in the present study, it was necessary to cover a long interval to capture the BOLD signal associated with statistical information processing, which can only emerge after several seconds of stimuli presentation, since statistical features by definition develop over time (as shown in the formal analysis, Section Construction of the sound sequences and formal validation of order manipulation). All analyses (whole brain and ROIs) focused on the 2.2 to 13.2 s interval that followed sequence onset. The first time-point (0–2.2 s) was not included in the analysis because sensitivity to statistics cannot emerge instantaneously. The last 2 time-points (13.2–17.6 s) were excluded because they corresponded to the tail of the HRF and undershoot periods. An example of a typical HRF is provided in Fig. 2.

Group-level voxel wise analyses

First, whole-brain group analyses were performed using SUMA on the participants' beta values resulting from the first level analysis. A 3-way repeated measure ANOVA was conducted with Category (speech, non-speech), Statistical structure (low, mid, high) and Time (starting at 2.2 s post-stimulus onset to 13.2 s in five steps) as the within-subject factors. All group analyses were corrected for multiple comparisons using a Monte Carlo simulation procedure on surface data, which implements the cluster-size threshold procedure as a protection against Type I error (Forman et al., 1995). The simulations determined that a family-wise error (FWE) rate of $p<.05$ is achieved with a minimum cluster size of 79 contiguous surface nodes (rmm 4 mm), with each node significant at $p<.001$. From this analysis we also identified areas sensitive to both speech and non-speech (*speech ∩ non-speech.*) via a conjunction mask (Nichols et al., 2005) of brain activity from the whole-brain contrasts (corrected for multiple comparisons).

Anatomical ROI analysis

Anatomical definitions for the ROIs. The core of the analyses focused on characterizing the functional profiles of a set of 13 a priori selected bilateral anatomical regions of interest (ROI) covering the supratemporal plane. ROIs were anatomically defined on each individual's cortical surface representation using an automated parcellation scheme (Desikan et al., 2006; Fischl et al., 2004). This procedure uses a probabilistic labeling algorithm that incorporates the anatomical conventions of Duvernoy (1991) and thus is based on macroanatomical landmarks, not on cytoarchitectonic maps. The anatomical accuracy of this method is high, approaching that of manual parcellation (Desikan et al., 2006; Fischl et al., 2002, 2004). This initial parcellation was augmented manually with further subdivisions to increase the spatial resolution of the ROI analysis. The ROIs used in the current study included: (1) planum polare (PP), (2) superior temporal gyrus (STG), (3) transverse temporal gyrus (TTG), (4) transverse temporal sulcus (TTS), (5) planum temporale (PT), and (6) caudal segment of the Sylvian fissure (SF); these ROIs are described in Table 1 (see also Supplementary material S6 for an illustration).

ROI-level statistical analyses. For each ROI and each participant, we first extracted the estimated percentage of BOLD signal change (beta weights) for each of the 6 conditions. This resulted in a per-participant table of 6 (condition) by 5 (time-points) reflecting activity ranging from 2.2 s post-stimulus onset to 13.2 s (exclusively) post stimulus onset; all statistical analyses were conducted on these tables. For each ROI we conducted a Greenhouse–Geisser corrected[6] 3-way ANOVA, with repeated measurements on Category (speech, non-speech), Statistical structure (low, mid, and high), and Time (5 time points).

The first step in the analysis involved examination of Category (C) effects collapsing over statistical structure, which provides a basic understanding of auditory Category preference in the supratemporal plane. To this end, we identified ROIs showing an FDR-corrected ($q=.05$, $i=13$ per hemisphere) significant main effect of Category, as well as ROIs showing a significant FDR-corrected Category×Time (CT) interaction. Two kinds of response profiles can result in a main effect of Category. The first, which we refer to as a *relative* advantage (or 'preference'), corresponds to the case where a ROI is significantly active for both speech and non-speech, with stronger activation magnitude for one category over the other. The second response pattern, which we refer to as an *absolute* preference, indicates that a ROI is significantly active for only one category and further shows a stronger response to this category than the other. To determine the type of response pattern driving category-sensitivity in supratemporal ROIs, we examined, using a set of FDR-corrected one-sample comparisons, whether activation level was significantly different from zero for each Category. For ROIs showing a Category×Time (CT) interaction, this procedure was conducted on the time-point showing the maximal effect of auditory category.

The second step in the analysis was to identify ROIs sensitive to the statistical structure manipulation, and particularly, whether sensitivity to statistical structure varied as function of Category (speech vs. non-speech). Absence of modulation of Statistical structure by Category suggests general, category-independent processing. In those regions that showed statistically significant Category× Statistical structure (CSS) interaction (FDR corrected ($q=.05$, $i=13$ per hemisphere)), we examined the general response profiles for the three SS levels (low, mid, and high), using a set of orthogonal complex contrasts (trends) performed separately for the speech and

[6] This correction was applied whenever the sphericity assumption was violated.

the non-speech sequences (Rosenthal et al., 2000). Trend analyses establish the function that best characterizes the relationship between a dependent variable (brain activity) and the levels of an independent variable (SS). Here we examined two orthogonal trends, the linear (1 0–1), and the quadratic trends (1–2 1). For ROIs showing a 3-way Category × SS × Time (CSST) interaction, the same trend analysis was conducted, focusing on the time point exhibiting the maximal SS effect within each auditory domain.

Behavioral auditory discrimination study

In order to ensure that the combinations of sounds (both speech and non-speech) used in the main experiment were equally discriminable, we conducted an independent behavioral auditory discrimination experiment (AX same/difference judgment for pairs of sound) with a separate group of participants consisting of 11 native Italian speakers (5 males; mean age: 24 ± 9.5 years; education: 13.9 ± 1.97 years), all right-handed (Oldfield, 1971) with self-reported normal hearing. A total of 756 pairs of different sounds were created using PRAAT (Boersma and Weenink, 2011), including 378 pairs of bird sounds and 378 pairs of syllables. Each syllable-pair or bird-sound-pair that was presented during the main experiment (fMRI) was included in the behavioral study. In addition, 756 pairs of identical sounds (378 birds, 378 syllables) were included to control for potential response biases and allow sensitivity calculations. To avoid long testing sessions and associated fatigue effects, this material set was split into two lists, and each participant presented with one of those. A self-paced AX discrimination paradigm was used in which pairs of sound were presented using PRAAT in a sound-attenuated booth through high quality headphones. Sounds within a trial were presented at the same rate used in the sequences in the main experiment, that is, they were separated by 50 ms silence. Participants performed same/different judgments by pressing on a computer mouse button. Their responses were recorded using PRATT.

Results

Behavioral results

Behavioral auditory discrimination study

Accuracy (mean ± SD) in the discrimination task approached ceiling with 99.63 ± .64% correct responses for speech and 99.39 ± .89% correct responses for non-speech pairs. Crucially, for both speech and non-speech, participants almost never rated "different" pairs as the same: the accuracy in these trials was 99.9% for speech, and 99.86% for non-speech, indicating that the differences observed in processing speech vs. non-speech sequences in the main experiment cannot be attributed to an intrinsic perceptual advantage for discriminating speech over non-speech sounds.

In-scanner performance on auxiliary visual target detection task

Participants' performance on the visual monitoring task during the fMRI study was highly accurate with 96.7% of correct responses, indicating that participants were alert during the study. In total 15 mistakes over 456 trials were committed. One participant (#19) failed to respond and was therefore excluded from the fMRI analyses.

Performance in the post-scan behavioral task

In the postscan behavioral study, participants were presented again with all the sequences they heard during the fMRI study, and were asked to rate the perceived regularity of each sequence (explicit rating of statistical structure). In 40% of the sequences, they were also asked to provide a rating of how many unique sounds were embedded in the series they heard ("numerosity ratings"). Analyses of variance were performed on both these ratings.

Ratings of numerosity offer a unique insight into how participants were able to segment the speech and non-speech sequences into constituents. For these, a 2 × 2 ANOVA with Category (speech, non-speech) and Statistical Structure (low, medium, high) revealed a significant main effect of Category ($F_{(1,19)} = 5.09$, $p = .036$), a significant main effect of statistical structure ($F_{(2,38)} = 5.36$, $p = .009$), and no interaction ($F_{(2,38)} = .72$, $p = .49$). Follow-up analyses were conducted to interpret these effects. In general, participants' numerosity ratings indicated that they were less successful in segmenting the non-speech sequences. Collapsing over the level of statistical structure, numerosity evaluations were more accurate (closer to target, i.e., 4) for speech than non-speech sequences (mean numerosity score for speech = 4.03 ± .26 items, not significantly different from 4; mean numerosity score for non-speech = 3.87 ± .48 items, $p = .036$, indicating significant difference from 4). When examining the impact of objective statistical structure, we found that participants perceived more unique sounds for more random sequences, resulting in a main effect of statistical structure. A trend analysis revealed that statistical regularities affected perceived numerosity, similarly for speech and non-speech sequence, characterized by increased perception of numerosity with greater randomness (for speech, linear: $F_{(1,19)} = 3.672$, $p = .07$; quadratic: $F_{(1,19)} < 1$, $p = .89$; for non-Speech linear: $F_{(1,19)} = 6.78$, $p = .018$; quadratic: $F_{(1,19)} < 1$, $p = .88$). That is, participants tended to report more unique sounds when sequences were more random. These findings are illustrated in Fig. 1A.

For the explicit ratings of statistical structure, we found strong sensitivity to statistical structure (main effect of statistical structure $F_{(1,19)} = 60.8$, $p < .0001$), which was modulated by Category, exhibiting a significant statistical structure by Category interaction ($F_{(1,19)} =$

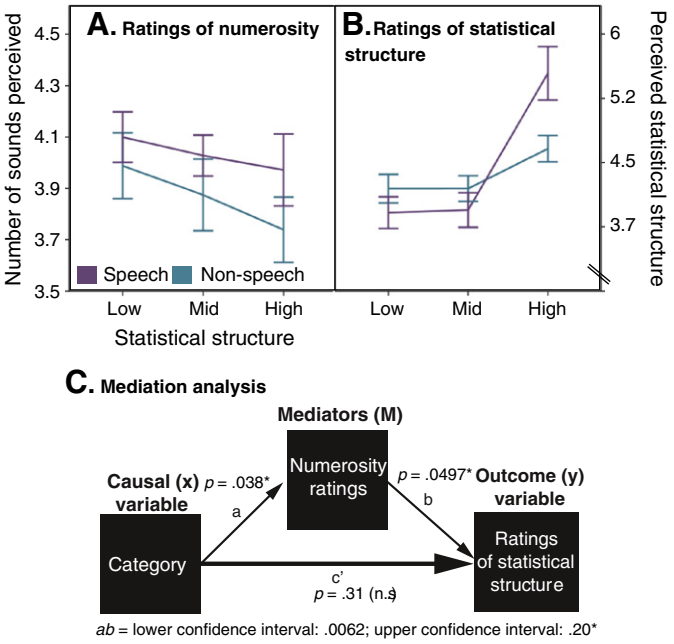

Fig. 1. Results of the post scan behavioral task. Panel A illustrates the perception of numerosity (number of unique sounds per sequence) as a function of Statistical Structure, separately for the speech and non-speech sounds. Panel B shows the perception of structure (ratings on a scale of 1–7) as a function of actual Statistical Structure, separately for the speech and non-speech sounds. The error bars represent the confidence intervals. Panel C illustrates the mediation analysis on the relationship of auditory category and perception of Statistical Structure. The direct effect of X (Category) on Y (perceived Statistical Structure), controlling for the effect of the mediator variable (perceived numerosity), is represented in this graph by the c′ path. The *a* path represents the direct effect of the causal variable on the mediator, while the *b* path represents the effect of the mediator on the outcome variable. The indirect effect of X though M is given by the *ab* path; the significance of the indirect effect is provided by bootstrapping confidence intervals. This model tests whether Category affects the perception of statistical structure, and whether that relationship is mediated by the perception of numerosity.

27.14, $p<.0001$). These results are illustrated in Fig. 1B. Importantly, the trend analysis revealed similar response profiles for speech and non-speech sequences characterized by an increase in perceived statistical structure as a function of objective Markov Entropy level (for the perception of statistical structure in speech: linear: $F_{(1,19)}=63.77$, $p<.0001$; quadratic: $F_{(1,19)}=39.69$, $p<.0001$; for the perception of statistical structure in non-speech: linear: $F_{(1,19)}=27.90$, $p<.0001$; quadratic: $F_{(1,19)}=7.67$, $p=.012$).

In sum, the behavioral findings indicate that participants could segment and evaluate the degree of within-sequence statistical regularity for both speech and non-speech sequences, with greater segmentation accuracy for speech. These results suggest a relationship between perceived numerosity and perceived statistical structure. To address this question, we conducted a mediation analysis to examine whether the relationship between Category and perceived statistical structure was mediated by perceived numerosity. The results, illustrated in Fig. 1C, reveal a significant relationship between category and perceived numerosity (the *a* path in the figure), a direct effect of perceived numerosity on perceived statistical structure (the *b* path). Importantly, there was a significant indirect (mediated) effect of Category on perceived statistical structure through perceived numerosity (the *ab* path), which shows that increased sensitivity to numerosity afforded by the speech signal lead to an apparent greater sensitivity to statistical structure. The importance of this result is in showing that the advantage of speech in terms of perceived structure likely originates from an easier segmentation process.

fMRI results

The fMRI analysis consisted of several steps. First, we conducted a whole-brain ANOVA that evaluated which brain regions were sensitive to Category, statistical structure, or to the Category by statistical structure interaction. This analysis also identified commonly and differently activated for speech and non-speech sequences. After documenting these findings on the whole-brain level we focus on the activity in subregions in the supratemporal plane.

Whole brain results

As shown in Fig. 2A, the 3-way ANOVA revealed that many bilateral supratemporal regions showed a main effect of Category, including the TTS, TTG and PT. The left ventral premotor cortex (PMv) in the precentral gyrus was also modulated by Category. Many regions showed a significant statistical structure by Time interaction (i.e., general differentiation between levels of statistical structure independent of time) consistent with our analysis (Section Construction of the sound sequences and formal validation of order manipulation) of how sensitivity to statistical structure develops over time (see Fig. 3A). Interestingly, these regions also included relatively low-level auditory regions, consisting of the TTG, TTS and PT bilaterally, as well as most of the left STG. As shown in Fig. 3B, the two-way Category × statistical structure interaction was significant in the left TTS and TTG. A list of all reliable activations for the ANOVA results is presented in Table 1.

In addition to conducting a 3-way ANOVA, we also examined the intersection of speech and non-speech sequence processing collapsing across statistical structure. Fig. 2B reveals the brain areas jointly activated during the presentation of both speech and non-speech sequences contrasted against rest (black outline). As can be seen in the figure, regions jointly active for both categories' sequences consisted of the TTG, TTS and PT. A list of all reliable joint activations is presented in Table 2. Also shown in Fig. 2B is the decomposition of the Category main effects revealed by the ANOVA. Regions showing stronger activation for speech compared to non-speech sequences are shown in purple – these included the left PMv and the lateral part of the supratemporal plane – those showing the opposite pattern are shown in turquoise. These were limited to the insula bilaterally, and most medial part of the supratemporal plane. The importance

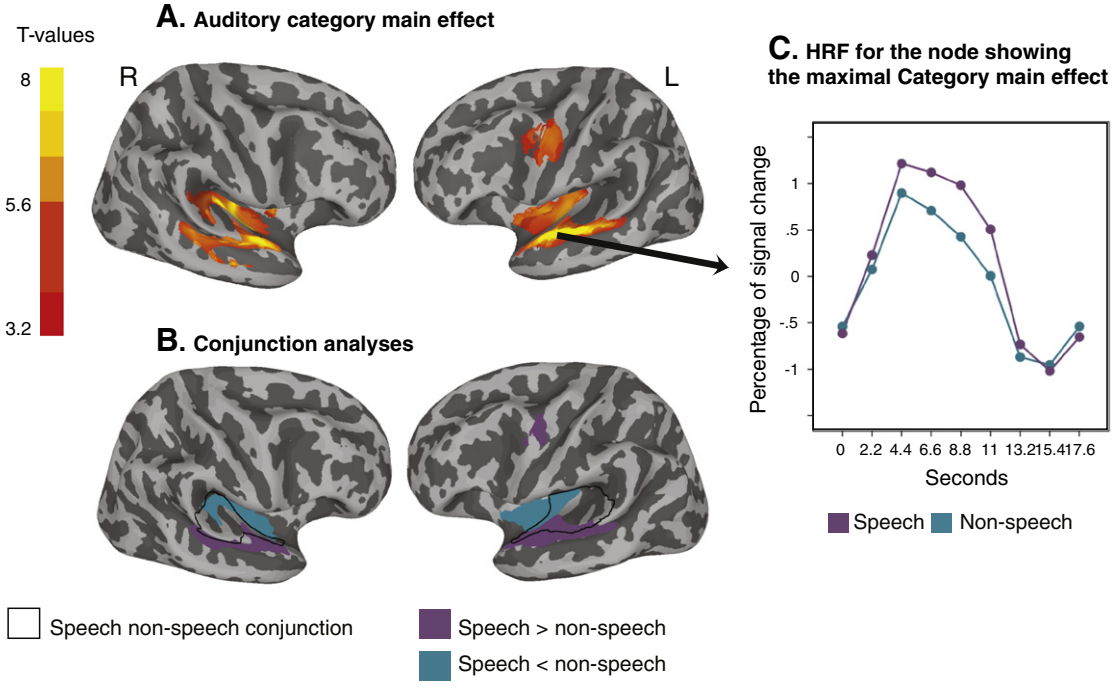

Fig. 2. Whole-brain analysis of auditory category effects. Panel A shows the regions that showed an overall main effect of auditory Category. Activation is shown on the group average smoothed flattened lateral surfaces. All analyses are controlled for family-wise error ($p<.05$) using cluster-level extent and a single voxel threshold of $p<.005$). Panel B illustrates the regions significantly active, at the group-level, for the contrasts of "speech greater than non-speech" (in purple), "non-speech greater than speech" (in turquoise) and for the conjunction of speech and non-speech (speech ∩ non-speech; black outline). Panel C illustrates the time-course of the HRF at the voxel of maximal Category effect in the left TTS, separately for speech (in purple) and non-speech (in turquoise). (For interpretation of the references to color in this figure legend, the reader is referred to the web of this article.)

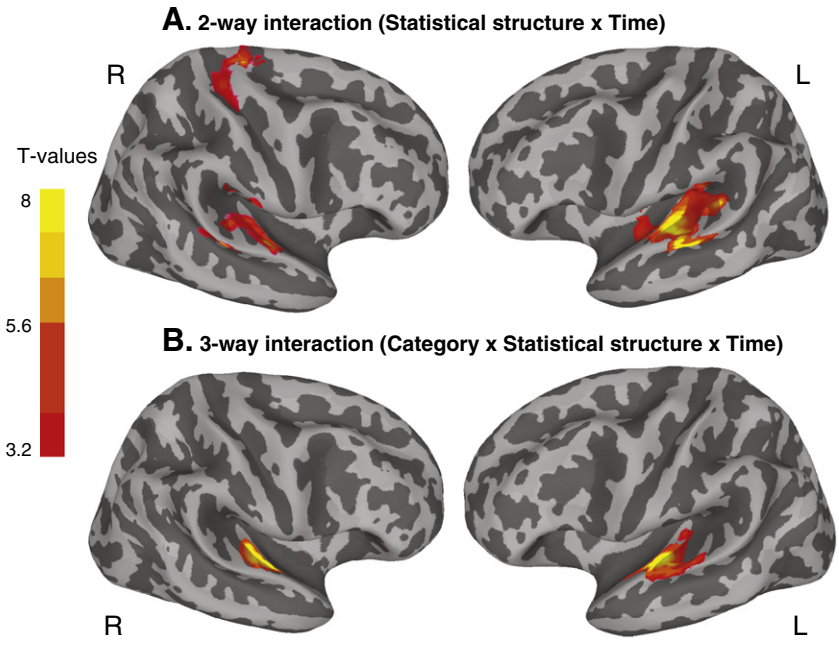

Fig. 3. Whole-brain analysis of sensitivity to Statistical Structure. The figure illustrates regions showing a statistically significant, Statistical Structure × Time interactions (panel A) and Statistical Structure × Category × Time interactions (panel B). Results are shown on the group average smoothed flattened lateral surfaces. Pale gray denotes a gyri and dark gray denotes sulci. R = right hemisphere. Analysis controlled for family-wise error ($p<.05$) using cluster-level extent and a single voxel threshold of $p<.005$).

of the whole-brain analysis is in validating our basic finding against the prior literature; the detailed ROI analysis presented below offers a more accurate evaluation of these findings in that the ROIs were determined on the basis of each participant's own anatomy hence conclusions do not rely on successful registration to common space across participants as do whole-brain analyses.

Supra-temporal ROI analyses

The ROI analysis consisted of a series of 2 (Category) × 3 (Statistical Structure) × 5 (Time in hemodynamic response function) ANOVA. These were conducted separately for each region, and their results are decomposed in the following paragraphs.

Sensitivity to auditory category. We began by identifying supratemporal regions where activity varied as a function of Category (speech vs. non-speech), as indicated by a statistically significant main effect of Category (C) in the ANOVA or as a reliable Category × Time interaction (CT) (FDR-corrected ($q=.05$, $i=13$)). As shown in Fig. 4, all but three supratemporal regions were sensitive to Category. Next, we determined the type of response driving the effect (as detailed in

Table 2
FWE-corrected group-level (N = 19), cortical surface results for: A. Input domain, B. input by time, C. input by structure and D. Structure by time, E. Input by structure by time. All coordinates are in Talairach space and represent the centroid surface node for each of the cluster (minimum cluster size: 168 contiguous surface nodes, each significant at p<.005).

Condition	Anatomical location	Hemi	x	Y	z	Max *p* value	Z	Number of nodes	Max F value
A. Input domain	TTG, extending caudally into TTS and PT, medially into the SF, insula and subcentral gyrus, and laterally covering most of the STG.	Left	−61	−15	2	<.0001	3.7	8015	115.5
	Ventral premotor cortex in the precentral gyrus.	Left	−51	−8	46	.0002	3.5	972	22.8
	Insula, extending laterally into SF and PT and anterior STG and STS	Right	40	−25	2	<.0001	3.7	9328	82.5
B. Input by Time	TTG, extending caudally into TTS and PT, medially into the SF, insula, and laterally covering most of the STG.	Left	−59	−9	−2	<.0001	3.7	11,347	47.2
	Subcentral gyrus and sulcus.	Left	−59	−12	13	<.0001	3.7	790	12.8
	Insula, extending laterally into SF and PT, TTS and TTG, and most of the STG and STS.	Right	44	−21	−1	<.0001	3.7	10,663	32.5
	Superior frontal gyrus.	Right	12	24	54	.0004	3.4	628	9.9
C. Input by Structure	Planum temporale and STS.	Left	−51	−30	5	.0002	3.5	478	1.9
	Ventral central sulcus.	Left	−40	−22	38	.0006	3.2	507	9.3
D. Structure by Time	TTS extending rostrally into TTG and caudally into PT. The cluster also extends medially into the SF, insula, and laterally into the posterior STG and supramarginal gyrus.	Left	−49	−20	3	.0026	2.8	4639	7.1
	Cingulate gyrus, extending into the SMA-proper.	Left	−11	−5	41	.0154	2.2	1275	4.8
	Dorsal precentral gyrus, extending into the dorsal central sulcus.	Right	21	−20	66	.0047	2.6	2055	6.3
	TTS extending rostrally into TTG and caudally into PT.	Right	56	−14	1	.0023	2.8	1324	7.2
	Cingulate gyrus, extending into the SMA-proper.	Right	5	−2	40	.0131	2.2	1115	4.9
	Posterior STG/STS	Right	59	−31	4	.0008	3.2	786	8.8
	Supramarginal gyrus.	Right	47	−27	24	.0192	2.1	851	4.4
E. Input by Structure by Time	TTS extending rostrally into TTG and caudally into PT, and laterally into the posterior STG.	Left	−52	−18	3	.0015	3.0	1997	7.9
	TTG, extending caudally into TTS.	Right	50	−16	5	.0051	2.6	993	6.1

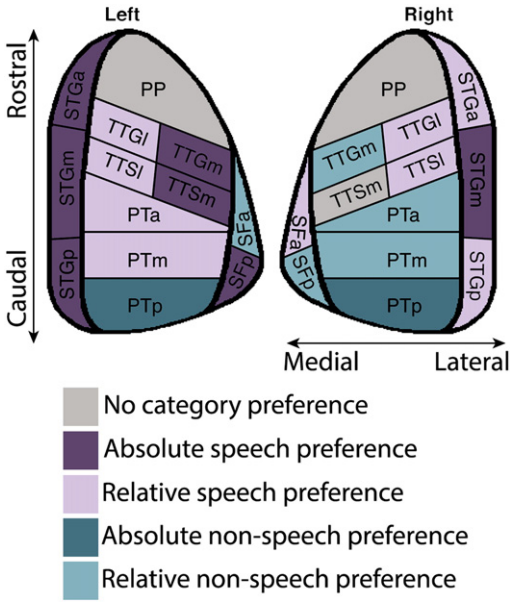

Fig. 4. Patterns of auditory category preference in bilateral supratemporal ROIs. This analysis identified auditory preference independent of sensitivity to statistical structure. Five patterns are color-coded and mapped onto a flattened schematic representation of the left and right supratemporal planeshowing the parcellation used in this study (different areas shown not to scale). ROI legend (see also text for more details): PP = planum polare; TTG = transverse temporal gyrus (m = medial, l = lateral); TTS = transverse temporal sulcus (m = medial, l = lateral); PT = planum temporale (a = anterior, m = middle, p = posterior); SF = caudal sylvian fissure (a = anterior, p = posterior); STG = superior temporal gyrus (a = anterior, m = middle, p = posterior). (For interpretation of the references to color in this figure legend, the reader is referred to the web of this article.)

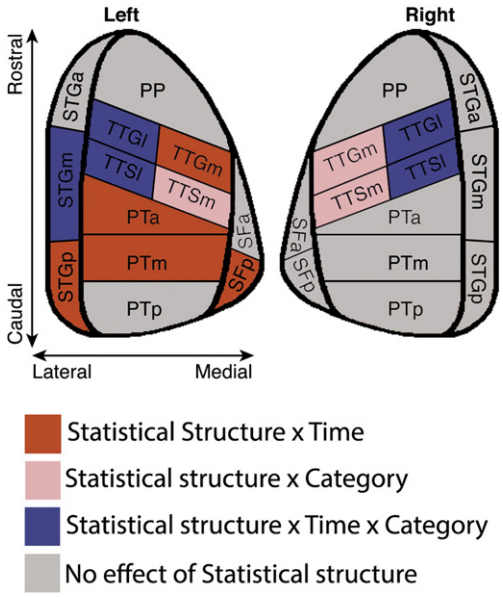

Fig. 5. Patterns of statistical structure effects in bilateral supratemporal ROIs. Four response patterns are color-coded and mapped onto a flattened schematic representation of the left and right supratemporal plane showing the parcellation used in this study (different areas shown not to scale).

Section ROI-level statistical analyses). Left supratemporal ROIs with a relative advantage for speech included the bilateral lateral TTG, the bilateral lateral TTS, the anterior and the mid PT, as well as the right anterior and posterior STG. ROIs with an absolute advantage for speech included the left medial TTG, the left medial TTS, the left anterior, middle and posterior STG, the right middle STG, and the posterior SF. Regions showing a relative advantage for the non-speech sounds included the left anterior SF, right medial TTG, right anterior and mid PT, and the right posterior SF. The bilateral posterior PT was the only ROI showing an absolute preference for non-speech. ROI findings are detailed in Table 3. The right medial TTS and the bilateral PP exhibited no category preference. To determine whether this lack of effect was driven by activation of similar magnitude for both categories, or by the absence of significant activity for either category, we conducted a set of additional one-sample t-tests against 0 (FDR corrected ($q = .05$, $i = 6$)). While activation in the right TTSm was significant for both (Speech: $t_{(18df)} = 7.15$, $p = .008$; Bird: $t_{(18df)} = 7.9$, $p = .016$), for bilateral PP it was not significant for either Category.

Category independent sensitivity to statistical structure. In total, 13 of 26 supratemporal ROIs were sensitive to statistical structure, either in a category-independent or a category-specific manner. The different sensitivity profiles are summarized in Fig. 5. As can be seen in the figure, the 5 ROIs that responded to statistical structure independently of category were located mainly in posterior supratemporal plane, extending into medial TTG. The other 8 ROIs showed differential sensitivity to statistical structure as function of Category (Table 4).

In the 5 regions that exhibited a reliable statistical structure by Time interaction (SST) – the simplest manifestation of sensitivity to statistical structure expected – different levels of statistical structure were associated with different shapes of hemodynamic responses. To characterize the impact of statistical structure, we collapsed the data across the two auditory categories and identified the time point at which the BOLD signal showed the maximally differentiated response for each level of statistical structure. To this aim, we first computed the difference between the three SS levels at each time point in the HRF and then we identified the time-point at which this difference was the greatest. After identifying this time point, a trend analysis was conducted to describe the relation between the three SS levels. As can be seen in Supplementary material S7, the response profiles in all ROIs were shaped as an upright quadratic ("V-shaped") function, showing minimum activation for the mid-structured sequences.

As reported in our analysis of category sensitivity (Section Sensitivity to auditory category above), all five ROIs exhibited a statistically significant preference for one category or the other. It is therefore interesting to note that despite such preference, these regions' sensitivity to statistical structure was independent of Category, suggesting that auditory category preference may not be a strong indicator of scope of processing.

Category-specific sensitivity to statistical structure. In approximately 62% of all supratemporal ROIs that were sensitive statistical structure, sensitivity to statistics was contingent upon Category. This was statistically indicated by a 2-way interaction (Category × statistical structure; CSS interaction below), or by a 3-way interaction (Category × Statistical Structure × Time; CSST interaction below). ROIs showing a reliable CSS interaction were limited to bilateral TTSm and the right TTGm (see Supplementary material S8). Both regions showed sensitivity to statistical structure in non-speech stimuli only (a quadratic trend). Interestingly, as described in our analysis of category preference (Section Sensitivity to auditory category), both regions showed

Table 3
Group-level (N = 19), cortical surface results for intersection of the speech and non-speech sequences. All coordinates are in Talairach space and represent the centroid surface node for each of the cluster (Family-wise corrected using cluster extent with minimum cluster size: 79 contiguous surface nodes, each significant at $p < .001$).

Anatomical location	Hemisphere	x	y	z	Number of nodes
TTG, extending caudally into the TTS and PT, and medially into SF.	Left	−41	−28	8	5014
Subcentral gyrus		−48	−17	16	322
TTG, extending caudally into the TTS, PT, and STG, medially into SF, and rostrally into PP.	Right	51	−25	6	3785

Table 4
Summary of the ROI corrected ANOVA results.

#	ROI	Hemisphere	Category effect (df = 1,18)	Category × Time (df = 4,72)	Specialization		SS	SS × Time (df = 8,144)	Category × SS (df = 2,36)		Category × SS × Time (df = 8,144)	
					Category	Degree of specialization			p	Applicability	p	Applicability
1	Lateral TTG	Left	.0003	n.s.	Speech	Relative	n.s.	n.s.	n.s.	N/A	.002	Both
2	Medial TTG	Left	n.s.	.002	Speech	Absolute	n.s.	.002	n.s.	N/A	n.s.	N/A
3	Lateral TTS	Left	.00001	n.s.	Speech	Relative	n.s.	n.s.	n.s.	N/A	.0004	Both
4	Medial TTS	Left	n.s.	.0005	Speech	Absolute	n.s.	n.s.	.001	Non-speech	n.s.	N/A
5	Anterior PT	Left	.003	n.s.	Speech	Relative	n.s.	.006	n.s.	N/A	n.s.	N/A
6	Mid PT	Left	n.s.	.0000000001	Speech	Relative	n.s.	.021	n.s.	N/A	n.s.	N/A
7	Posterior PT	Left	n.s.	.038	Non-speech	Absolute	n.s.	n.s.	n.s.	N/A	n.s.	N/A
8	Anterior SF	Left	.005	n.s.	Non-speech	Relative	n.s.	n.s.	n.s.	N/A	n.s.	N/A
9	Posterior SF	Left	n.s.	.0002	Speech	Absolute	n.s.	.034	n.s.	N/A	n.s.	N/A
10	Anterior STG	Left	.0001	n.s.	Speech	Absolute	n.s.	n.s.	n.s.	N/A	n.s.	N/A
11	Middle STG	Left	.0000000004	n.s.	Speech	Absolute	n.s.	n.s.	n.s.	N/A	.014	Both
12	Posterior STG	Left	.000002	n.s.	Speech	Absolute	n.s.	.007	n.s.	N/A	n.s.	N/A
13	PP	Left	n.s.	n.s.	N/A	N/A	n.s.	n.s.	n.s.	N/A	n.s.	N/A
14	Lateral TTG	Right	.007	n.s.	Speech	Relative	n.s.	n.s.	n.s.	N/A	.002	Both
15	Medial TTG	Right	n.s.	.002	Non-Speech	Relative	n.s.	n.s.	.005	Non-speech	n.s.	N/A
16	Lateral TTS	Right	.006	n.s.	Speech	Relative	n.s.	n.s.	n.s.	N/A	.0001	Both
17	Medial TTS	Right	n.s.	n.s.	N/A	N/A	n.s.	n.s.	.001	Non-speech	n.s.	N/A
18	Anterior PT	Right	n.s.	.002	Non-Speech	Relative	n.s.	n.s.	n.s.	N/A	n.s.	N/A
19	Mid PT	Right	n.s.	.0003	Non-Speech	Relative	n.s.	n.s.	n.s.	N/A	n.s.	N/A
20	Posterior PT	Right	.003	n.s.	Non-Speech	Absolute	n.s.	n.s.	n.s.	N/A	n.s.	N/A
21	Anterior SF	Right	.00003	n.s.	Speech	Relative	n.s.	n.s.	n.s.	N/A	n.s.	N/A
22	Posterior SF	Right	.0001	.00001	Non-Speech	Relative	n.s.	n.s.	n.s.	N/A	n.s.	N/A
23	Anterior STG	Right	.013	n.s.	Speech	Relative	n.s.	n.s.	n.s.	N/A	n.s.	N/A
24	Middle STG	Right	.00001	n.s.	Speech	Absolute	n.s.	n.s.	n.s.	N/A	n.s.	N/A
25	Posterior STG	Right	.003	n.s.	Speech	Relative	n.s.	n.s.	n.s.	N/A	n.s.	N/A
26	PP	Right	n.s.	n.s.	N/A	N/A	n.s.	n.s.	n.s.	N/A	n.s.	N/A

strong responses to both speech and non-speech stimuli, and thus the finding of sensitivity to statistics for non-speech stimuli only indicates a that statistical processing was implemented solely for non-speech inputs. As we mention in the Discussion, this process may be related to the enhanced segmentation requirements for less familiar stimuli.

For the other ROIs sensitive to statistical structure, sensitivity was marked by a more complex response patterns characterized by a reliable three-way CSST interaction (FDR-corrected). This interaction indicates that the BOLD responses to statistical structure in the speech and non-speech sequences were differently shaped across time-points. Specifically, for each category, we identified the HRF time-point of maximal difference between levels of statistical structure (as described in Section ROI-level statistical analyses). Fig. 6 shows response patterns at these time-points. As shown in the figure, these regions included the bilateral lateral TTG and lateral TTS, and the left middle STG. Interestingly, for speech, sensitivity for statistical structure occurred early on, within 4 of sequence onset (a timing that is consistent with our KL-divergence analysis described in Section Construction of the sound sequences and formal validation of order manipulation), while for non-speech maximal differentiation between SS levels was found at about 10 s post sequence onset, that is, few seconds after the end of the sequences. Moreover, while sensitivity to statistical structure in speech was primarily shaped as an inverted quadratic function with maximal activation for the mid-structured sequences, sensitivity to statistical structure in non-speech had the reversed shape (V-shaped), indicating higher activity for high and low levels of predictability.

Linear responses to statistical structure were found in only two ROIs: the right lateral TTS and the left lateral TTG. In the former, sensitivity to statistical structure was found for speech only, with increased activity for greater randomness. In the left lateral TTG, a similar linear pattern was found for both speech and non-speech sequences.

Discussion

Prior work examining sensitivity to statistical structure in speech inputs (McNealy et al., 2006) has shown that parts of the supratemporal plane are sensitive to random vs. deterministic (structured) speech inputs, with stronger BOLD response for the former. Overath et al. (2007) found that bilateral PT is sensitive to statistical structure in tone sequences. However, whether these statistical processes are employed in a similar manner for different kinds of auditory input had not been examined. The present study addressed this issue, identifying supratemporal regions sensitive to statistical structure as a function of, or independently from, auditory category. The present study also contributes to the understanding of the neurobiology of the supratemporal plane in showing that different subregions of this area exhibit markedly different profiles to statistical structure.

In contrast to prior work, we examined the neural mechanisms underlying the ability to process statistical structure, operationalized as the neural sensitivity to transition probability (TP) structure in sequences of speech and non-speech sounds that were presented in the context of an incidental visual task. Such dual-task contexts are known to be the most demanding in terms of sensitivity to statistical structure in auditory inputs; it has been claimed that in absence of explicit attention to auditory inputs such sensitivity is weak if present at all (Bekinschtein et al., 2009; Kimura et al., 2010). Our results challenge these previous results by revealing strong sensitivity to TP in the supratemporal plane even when participants were not focusing on auditory stimuli.

Importantly, posterior supratemporal regions were particularly sensitive to statistical structure. We expected that some regions would track statistics in a category-independent manner reflecting domain general mechanisms. This pattern was indeed identified, but in a relatively restricted set of posterior left supratemporal regions. We also expected that speech would offer a significant advantage in the processing of statistics resulting in a lower response magnitude or a more focalized activation pattern. And indeed, we identified a few supratemporal areas in which sensitivity to statistical structure was restricted to non-speech sounds even though these areas responded strongly to both sound categories. Given the behavioral findings, we interpret this pattern as reflecting reduced segmentation demands for more familiar speech sounds (Section Domain-general processing of statistical structure in lateral transverse temporal areas) rather than a

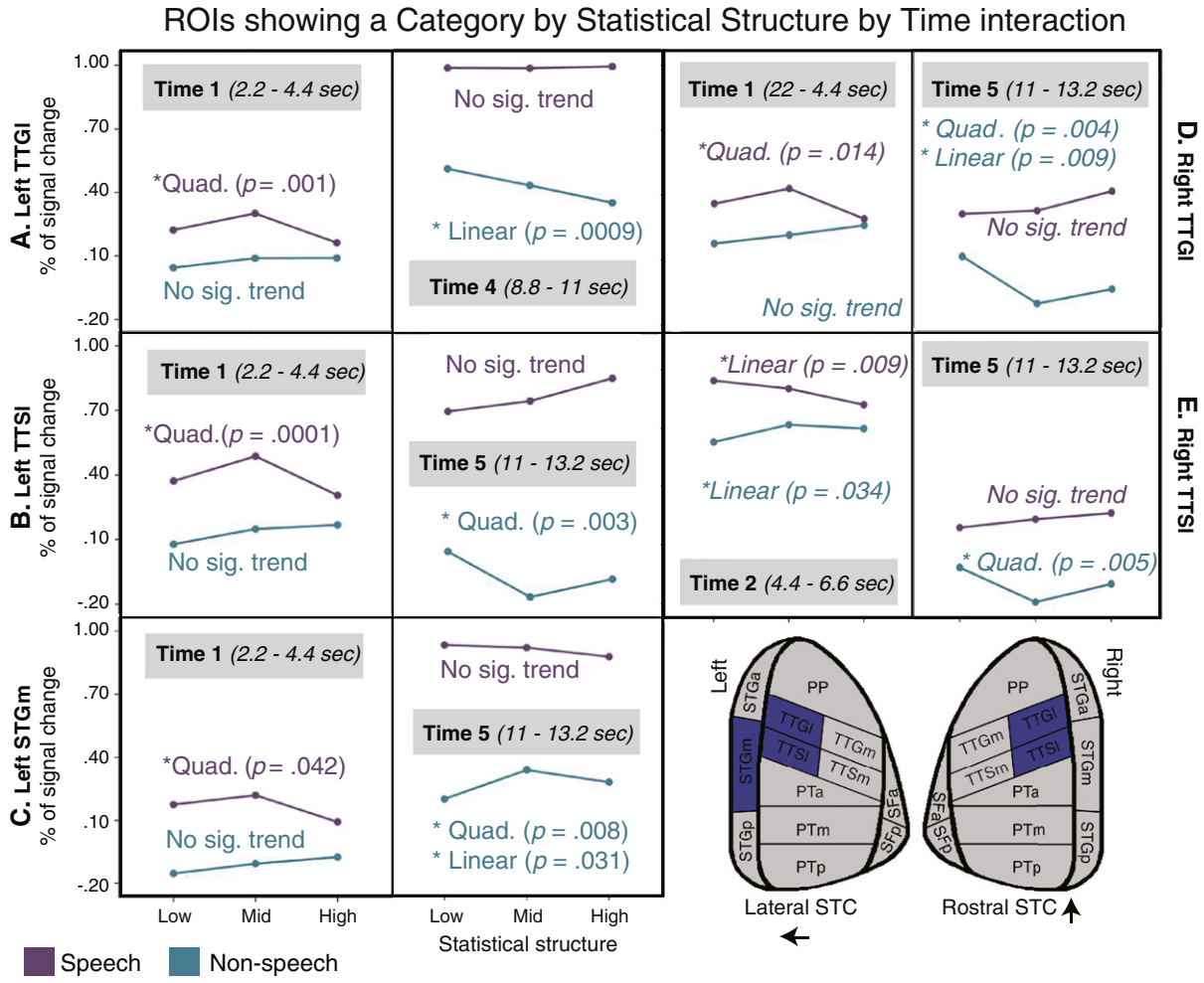

Fig. 6. Sensitivity to statistical structure modulated by auditory category in the supratemporal plane, at two different time-points. The five regions colored in blue in the schematic diagram of the supratemporal plane showed a reliable Statistical Structure × Category × Time (SSCT) interaction. For each region, a two-panel graph shows responses in the six conditions at the time point where the levels of statistical structure differed maximally for speech (left panel) and where the three levels statistical structure differed maximally for non-speech (right panel).

process mediating statistical coding per se. Finally, an unexpected but important finding was that category preference, defined in terms of increased activation magnitude for one stimulus category over another, was not a reliable predictor of whether a region tracks statistical structure in the given category (Section Segmentation in medial transverse temporal areas). On this issue, we identified (1) regions that responded strongly to both speech and non-speech categories but yet tracked statistical structure in only one category, and (2) regions who tracked statistics in a way that is independent of category yet exhibited a clear preference for one category over the other in terms of response magnitude. To summarize, our findings indicate that supratemporal cortex is highly sensitive to statistical structure in auditory inputs. More specifically, some regions are sensitive to statistics in a general manner, whereas others show category-based differences in the tracking of statistics.

Sensitivity to auditory category vs. acoustics

In the current work we contrasted speech to another kind of natural sounds (bird sounds). For this reason, the two categories were not completely matched acoustically, which imposes several potential interpretive limitations. First, when quantifying overall BOLD responses, it is possible that any or all response differences between speech and non-speech stimuli derive from differences in acoustic patterns rather than speech category per se. Thus, what we term 'speech preference' could be related to differential sensitivity to any of several acoustic features known to be encoded in the supratemporal plane, such as pitch, degree of variation of spectral centroid, median harmonicity, and the degree to which loudness changes over time (see for example Giordano et al., In press; Leaver and Rauschecker, 2010). Second, we are inclined to follow prior work in interpreting sensitivity to statistical structure in speech sequences in terms of sensitivity to constraints governing speech-level syllable or phoneme transitions (i.e., speech-code statistics; McNealy et al., 2006; Buiatti et al., 2009). However, it is, *in principle*, possible that the apparent sensitivity to statistics within speech sequences, as found here and in prior work, reflects the tracking of low-level sub-phonemic (acoustic) features, or a combination of phonemic and sub-phonemic tracking in anatomically distinct areas. For instance our results could indicate sensitivity to transition probabilities that exist between formant patterns rather than, or in addition to, transitions between speech-level units such as phonemes or syllables. Because it is impossible to determine which stimuli features were driving the effects for speech, the possibility of sensitivity to acoustical features (i.e. a 'low-level' account) cannot be completely ruled out. However, when considering the pattern of our behavioral results, fMRI results, and prior work as a whole, we would argue that it is more likely that differences in statistical sensitivity patterns that were found for speech and non-speech sequences are not reducible to acoustical differences. Our behavioral data revealed different

segmentation patterns for speech vs. non-speech stimuli, corroborating prior work (Marcus et al., 2007) that identified an advantage for extracting simple statistical patterns from sequences of speech vs. non-speech sounds. It is very likely that the neural cost of segmentation varies for speech and non-speech sounds, with greater resources being allocated for processing less familiar sounds. This account parsimoniously explains why certain regions in the current study showed different BOLD profiles to statistical structure as function of category. In contrast, a 'low-level' account holds that one is tracking acoustical patterns, and in the absence of several linking hypotheses going beyond acoustics *per se*, would find it difficult to explain why the relation between BOLD activity and statistical regularity showed one pattern for speech sounds but another pattern for non-speech sounds in regions strongly responsive to *both* auditory categories.

Nonetheless, the low-level account could explain one important pattern: the specific case where a certain region tracked statistical structure for the sound category it responded to more strongly, but did not track statistical structure for the sound category it responded to more weakly. This interaction pattern could indicate a floor-response for the latter category. In practice, however, this pattern never occurred in our data. Specifically, 7 of the 8 regions that showed different responses to statistical structure as a function of category responded strongly to both sound categories at the point of maximal differentiation (see Fig. 6 and Supplementary material S8). A minor exception was the middle part of left STG: that region tracked statistical structure for both categories, but its activity was low for the non-speech sounds. In sum, all ROIs showing category-dependent sensitivity to statistical structure were sensitive to statistical structure in both categories.

Anterior to posterior supratemporal plane: from speech decoding to the processing of statistical structure

Anterior supratemporal plane

Our findings suggest that anterior supratemporal regions are involved in decoding speech independent of larger-scale statistical structure, while posterior supratemporal areas are involved in processing speech and non-speech sounds over time as revealed by their sensitivity to statistical structure. The anterior supratemporal plane, including anterior STG and PP, exhibited the simplest response profiles. While the anterior STG showed preference for speech bilaterally, the bilateral PP, in contrast, was not significantly active in any condition. Importantly, these regions did not show any sensitivity to statistical structure. The preference for speech, combined with lack of sensitivity to statistics tentatively suggests these regions operate at the level of single speech units. This interpretation is consistent with prior findings linking the anterior STG/STS area to speech and voice processing in both human (Hickok and Poeppel, 2000; Rauschecker and Scott, 2009) and non-human primates (Petkov et al., 2008, 2009; Rauschecker and Tian, 2000; Tian et al., 2001), through the ventral auditory 'what' pathway. The lack of activation in PP, bilaterally, may be related to the fact that participants were presented with only one speaker, hence only one voice, with which they were familiarized before the beginning of the experiment. It follows from this that the contribution of any potential voice processing/recognition/normalization mechanisms was likely minimized, which may explain the lack of activation in PP.

Posterior supratemporal plane

In contrast to the anterior supratemporal areas, several posterior supratemporal areas exhibited sensitivity to statistical structure. To preface the discussion, we found an absolute preference for speech in the middle and posterior segments of the left STG as well as in the middle part of the right STG. However, responses to statistical structure in these regions did not parallel this preference pattern. Specifically, activation in the left middle and posterior STG tracked statistics in *both* speech and non-speech with some regions showing the same response pattern to statistical structure in both domains. Particularly noteworthy is the fact that the left posterior STG does not distinguish between auditory categories in its response to variations in statistical structure. These findings are consistent with previous fMRI, near-infrared spectroscopy (NIRS) and MEG studies showing modulation of activation in posterior temporal areas contingent upon the presence of statistical information in streams of non-speech sounds (Abla and Okanoya, 2008; Furl et al., 2011; Overath et al., 2007) and even visual sequences (Bischoff-Grethe et al., 2000).

The left PT was also sensitive to statistical structure, but anterior/medial responses differed from those found in posterior PT. These findings corroborate prior findings of multiple functional areas in the PT/SF complex (Fullerton and Pandya, 2007; Galaburda and Sanides, 1980; Rivier and Clarke, 1997; Scheich et al., 1998; Sweet et al., 2005; Tardif and Clarke, 2001; von Economo and Horn, 1930). This finding is consistent with a recent fMRI study that identified three non-overlapping regions in the bilateral human PT, each with a distinct response to speech (Tremblay et al., in press). We found sensitivity to statistical structure for both speech and non-speech in the left PT, but the right PT did not track statistics at all. Our findings are consistent with those of Overath et al. (2007) that found sensitivity to statistical structure of tone sequences in PT. However, they identified sensitivity in PT bilaterally (the difference in findings could be due to lack of control for multiple comparisons in the latter study).

On the basis of our findings and those of Overath et al. we propose that the left anterior and middle PT are involved in constructing predictions of what is likely to be presented next. In the present experiment, such predictive coding may have entailed building a representation of auditory sequence structure (e.g. sound pairs with high TP, etc.) and testing incoming sounds against the predicted ones. Such mechanisms can be used to detect inconsistencies, particularly in the highly predictable sequences, resulting in online adjustment of transient internal representations and error signals generation. This interpretation would be compatible with the computational hub hypothesis (Griffiths and Warren, 2002), according to which the left PT disambiguates complex sounds by matching their temporal and spectral characteristics to stored templates (Griffiths and Warren, 2002). The matching process may operate at multiple levels, comparing single sounds to templates but also complex sound streams (pairs, triplets, etc. of sounds) against internal representations. This hypothesis is consistent with prior findings, which demonstrated that PT responds to the presentation of unexpected events in auditory sequences, suggesting that PT is sensitive to the structure of sound sequences (Mustovic et al., 2003). In sum, our findings extend prior work that had identified sensitivity to statistical structure or surprise effects in PT, and shows that left PT mediates what are likely domain-general processing of statistical structure.

Domain-general processing of statistical structure in lateral transverse temporal areas

The most complex functional response patterns (3-way interactions) were found in lateral parts of the TTG and TTS bilaterally and left middle STG. While these areas showed relative preference for speech, they tracked statistics in both speech and non-speech sounds. However, sensitivity to statistical structure was manifested as distinct response profiles for speech and non-speech sequences. These results suggest that non-primary supratemporal areas houses powerful neural circuits capable of extracting and using complex statistical information. As shown by our analysis of the HRF profiles, these neural circuits appeared to process statistics more quickly (and perhaps more easily) from sequences of familiar sounds such as speech than from sequences of less familiar sounds, not because of preference

for speech per se, but possibly because of greater familiarity with speech than any other sounds. This may explain the recent finding that children are able to detect regularities such as ABB and ABA patterns in speech sequences, but not in auditory non-speech sequences consisting of tones, timbres, or animal call (Marcus et al., 2007).

Segmentation in medial transverse temporal areas

The bilateral medial TTS and the right medial TTG, which form the bulk of the bilateral primary auditory area (PAC), responded strongly to both speech and non-speech, yet tracked statistical structure *only* in the non-speech sequences. On the basis of this result, we suggest that these regions play an important role not in the processing of statistical structure per se or the representation of uncertainty, but in the segmentation of an input stream into units, which is a precursor to computation of statistics. In our study, both the speech and non-speech sequences had the same duration (225 ms) and were presented at the rate of 3.6 Hz, suggesting that both could be successfully encoded by the auditory system, which is thought to implement a 200-ms temporal integration window (Näätänen, 1992; Yabe et al., 1997, 1998). Still, the speech sounds enjoyed a special advantage: since the units forming the speech sequences were frequent Italian syllables, they were extremely well known to participants. In contrast, for the non-speech sequences, the units (the bird sounds) were unknown to the participants, and consequently, any functional process that makes uses of statistical information (e.g., for prediction) would first need to segment the units from the continuous input prior to establishing the statistical properties of the sequences. Processing less familiar sounds such as bird sounds, but also non-native languages, sequences of environmental sounds or other animal calls, may therefore be more demanding in terms of segmentation in the context of statistical information processing, which may result in increased activation magnitude or in the recruitment of additional brain areas. Because segmentation is tightly linked to statistical structure, statistics can only be computed on the basis of an a-priori internal representation of the unique elements present in a signal that has been established through segmentation. This idea is supported by our behavioral results: participants perceived fewer elements in the non-speech sequences and they also differed in their ability to rate SS across categories, being more accurate for speech than non-speech. Moreover, a mediation analysis showed that the relationship between Category and perceived regularity (ratings of sequence structure on a 7-point scale) is mediated by perceived number of elements. These findings converge to suggest that segmentation of the non-speech sounds was more laborious.

Our finding of complex response patterns in the human PAC (i.e. responses going beyond category representation) is consistent with recent findings of higher level processing in this region (Kilian-Hutten et al., 2011; Riecke et al., 2007; Staeren et al., 2009). For instance, Kilian-Hutten et al. (2011) recently showed that the posterior bank of TTG is involved in the disambiguation of speech sounds based on perceptual interpretations. In keeping with these previous results, the current results suggest that PAC, bilaterally, is involved in segmenting streams of sounds, a function that goes beyond category representation.

Category vs. process specialization in the supratemporal plane

Perhaps the most unexpected finding of the present study is that auditory preference was not a reliable indicator of the domain of applicability of statistical information processing. In the left medial TTG, left anterior and mid PT, and left posterior STG, for example, despite preference for speech, we found sensitivity to statistics in both categories. These findings have important implications for models of speech perception, which often focus on, or operationalize, "speech-related processes" by contrasting speech to non-speech stimuli and interpreting increased BOLD activity for speech in a particular region as reflecting the implementation of high-level speech-specific processes. Our results advise against this logic. Stronger responses to specific auditory categories may reflect familiarity, which may trigger 'expert-like' categorization processes not involved in processing unfamiliar sounds, as suggested by others (Mettler, 1932) but should not be taken as evidence of domain-specific neurocomputations. It is in this light that we interpret the findings that 1) some regions showed sensitivity to statistics in non-speech sounds but not for speech sounds, and 2) speech sequences were associated with more accurate segmentation. Both these indicate that there may be an advantage for processing statistical information for speech stimuli, which is not the product of dedicated neural resources or 'modules', but rather emerges from distinct neural computations within shared cortical substrates, a hypothesis previously suggested by others (Price et al., 2005).

Conclusions

The study of speech perception has been concerned to a large extent with the neurocomputations associated single syllables and phonemes, leaving processes specifically associated with the integration of information over large streams of sounds largely unexplored (but see McNealy et al., 2006 for an exception). The present finding of reliable, widespread sensitivity to statistical structure in several parts of the supratemporal plane, in a study where participants performed an unrelated visual task, suggests that statistical information processing is a spontaneously occurring and obligatory part of auditory sequence perception, even when not essential for overt behavior. By going beyond the processing of single stimuli, our findings demonstrate that posterior supratemporal were sensitive to the statistical structure of preferred (speech) as well as non-preferred auditory inputs (following Overath et al., 2007). As such, the results highlight the potential benefits of using measures beyond category preference to better understand the nature of the neural mechanisms implemented in the supratemporal cortex. Finally, the observed differences in regional activation patterns in the supratemporal plane, which presumably reflect activation in different populations of neurons, show the considerable heterogeneity of processing in this region and indicate the importance of examining supratemporal activation patterns at the highest possible anatomical resolution (for similar findings, see Obleser et al., 2010; Petkov et al., 2004). Further studies are needed to examine how the current findings generalize to other modalities (e.g. visual), other types of sounds, in particular, naturalistic complex sounds (computer sounds, sounds of cars, natural elements such as thunder, etc.) and more complex situations, such as processing speech sounds from multiple talkers. It will also be necessary to examine how other parts of the functional speech perception network such as the STS, inferior frontal gyrus and ventral premotor cortex, regions known to be involved in the processing of speech, responds to statistical structure and how they interact with parts of the supratemporal plane that are sensitive to this information.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2012.10.055.

Acknowledgments

We thank Margaret Moreno for her help in collecting the data, Monika Molnar for her PRATT, Francesco Cutugno for advice on Italian phonetic resources, the staff of the MRI lab of the Center for Brain and Mind Research, and all the participants. This study was supported by a research grant from the European Research Council under the 7th framework starting grant program (ERC-STG #263318) to U. Hasson and by a postdoctoral fellowship from the Canadian Institute for Health research (CIHR) to P. Tremblay. Their support is gratefully acknowledged. Thanks also the MRI staff of the Functional

Neuroimaging Lab at the Center for Mind & Brain Sciences (CIMeC) at the University of Trento.

References

Abla, D., Okanoya, K., 2008. Statistical segmentation of tone sequences activates the left inferior frontal cortex: a near-infrared spectroscopy study. Neuropsychologia 46 (11), 2787–2795.

Argall, B.D., Saad, Z.S., Beauchamp, M.S., 2006. Simplified intersubject averaging on the cortical surface using SUMA. Hum. Brain Mapp. 27 (1), 14–27.

Baroni, M.S., Bernardini, A., Ferraresi, A., Zanchetta, E., 2009. The WaCky Wide Web: a collection of very large linguistically processed Web-crawled corpora. J. Lang. Resour. Eval. 43 (3), 209–226.

Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., Naccache, L., 2009. Neural signature of the conscious processing of auditory regularities. Proc. Natl. Acad. Sci. U. S. A. 106 (5), 1672–1677.

Benson, R.R., Whalen, D.H., Richardson, M., Swainson, B., Clark, V.P., Lai, S., Liberman, A.M., 2001. Parametrically dissociating speech and nonspeech perception in the brain using fMRI. Brain Lang. 78 (3), 364–396.

Bischoff-Grethe, A., Proper, S.M., Mao, H., Daniels, K.A., Berns, G.S., 2000. Conscious and unconscious processing of nonverbal predictability in Wernicke's area. J. Neurosci. 20 (5), 1975–1981.

Boersma, P., Weenink, D., 2011. Praat: doing phonetics by computer (Version 5.2.10). Retrieved from http://www.praat.org/.

Buiatti, M., Pena, M., Dehaene-Lambertz, G., 2009. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. NeuroImage 44 (2), 509–519.

Cole, R.A., Jakimik, J., 1980. How are syllables used to recognize words? J. Acoust. Soc. Am. 67 (3), 965–970.

Cosi, P., Avesani, C., 2001. FESTIVAL speaks Italian. Paper presented at the Proceedings Eurospeech 2001, Aalborg, Denmark.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.

Dahl, A., Omdal, R., Waterloo, K., Joakimsen, O., Jacobsen, E.A., Koldingsnes, W., Mellgren, S.I., 2006. Detection of cerebral embolic signals in patients with systemic lupus erythematosus. J. Neurol. Neurosurg. Psychiatry 77 (6), 774–779.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. NeuroImage 9 (2), 179–194.

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31 (3), 968–980.

Duvernoy, H.M., 1991. The Human Brain: Structure, Three-dimensional Sectional Anatomy and MRI. Springer-Verlag, New York.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. NeuroImage 9 (2), 195–207.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D.H., Dale, A.M., 2004. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14 (1), 11–22.

Fiser, J., Aslin, R.N., 2001. Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychol. Sci. 12 (6), 499–504.

Fiser, J., Aslin, R.N., 2002. Statistical learning of new visual feature combinations by infants. Proc. Natl. Acad. Sci. U. S. A. 99 (24), 15822–15826.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. 33 (5), 636–647.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322 (5903), 970–973.

Friston, K., Kiebel, S., 2009. Predictive coding under the free-energy principle. Philos. Trans. R. Soc. Lond. B Biol. Sci. 364 (1521), 1211–1221.

Fullerton, B.C., Pandya, D.N., 2007. Architectonic analysis of the auditory-related areas of the superior temporal region in human brain. J. Comp. Neurol. 504 (5), 470–498.

Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., Griffiths, T.D., 2011. Neural prediction of higher-order auditory sequence statistics. NeuroImage 54 (3), 2267–2277.

Galaburda, A., Sanides, F., 1980. Cytoarchitectonic organization of the human auditory cortex. J. Comp. Neurol. 190 (3), 597–610.

Giordano, B.L., McAdams, S., Kriegeskorte, N., Zatorre, R., Belin, P., in press. Abstract encoding of auditory objects in cortical activity patterns. Cereb. Cortex (Electronic publication ahead of print).

Griffiths, T.D., Warren, J.D., 2002. The planum temporale as a computational hub. Trends Neurosci. 25 (7), 348–353.

Hasson, U., Nusbaum, H.C., Small, S.L., 2006. Repetition suppression for spoken sentences and the effect of task demands. J. Cogn. Neurosci. 18 (12), 2013–2029.

Hayes, A.F., 2008. SPSS macro for multiple mediation. Retrieved from http://www.comm.ohio-state.edu/ahayes/.

Hickok, G., Poeppel, D., 2000. Towards a functional neuroanatomy of speech perception. Trends Cogn. Sci. 4 (4), 131–138.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., Dupoux, E., 2003. Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. J. Neurosci. 23 (29), 9541–9546.

Johnstone, T., Ores Walsh, K.S., Greischar, L.L., Alexander, A.L., Fox, A.S., Davidson, R.J., Oakes, T.R., 2006. Motion correction and the use of motion covariates in multiple-subject fMRI analysis. Hum. Brain Mapp. 27, 779–788.

Kent, R.D., 2000. Research on speech motor control and its disorders: a review and prospective. J. Commun. Disord. 33 (5), 391–427 quiz 428.

Kilian-Hutten, N., Valente, G., Vroomen, J., Formisano, E., 2011. Auditory cortex encodes the perceptual interpretation of ambiguous sound. J. Neurosci. 31 (5), 1715–1720.

Kimura, M., Widmann, A., Schroger, E., 2010. Top-down attention affects sequential regularity representation in the human visual system. Int. J. Psychophysiol. 77 (2), 126–134.

Kirkham, N.Z., Slemmer, J.A., Johnson, S.P., 2002. Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition 83 (2), B35–B42.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. On Information and Sufficiency. Ann. Math. Stat. 22 (1), 79–86.

Leaver, A.M., Rauschecker, J.P., 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. 30 (22), 7604–7612.

Lehiste, I., Shockey, L., 1972. On the perception of coarticulation effects in English VCV syllables. J. Speech Hear. Res. 15 (3), 500–506.

Marcus, G.F., Fernandes, K.J., Johnson, S.P., 2007. Infant rule learning facilitated by speech. Psychol. Sci. 18 (5), 387–391.

McNealy, K., Mazziotta, J.C., Dapretto, M., 2006. Cracking the language code: neural mechanisms underlying speech parsing. J. Neurosci. 26 (29), 7629–7639.

McNealy, K., Mazziotta, J.C., Dapretto, M., 2010. The neural basis of speech parsing in children and adults. Dev. Sci. 13 (2), 385–406.

Meltzer, J.A., Negishi, M., Constable, R.T., 2008. Biphasic hemodynamic responses influence deactivation and may mask activation in block-design fMRI paradigms. Hum. Brain Mapp. 29 (4), 385–399.

Mettler, F.A., 1932. The Marchi method for demonstrating degenerated fiber connections within the central nervous system. Stain Technol. 7 (3), 95–106.

Mustovic, H., Scheffler, K., Di Salle, F., Esposito, F., Neuhoff, J.G., Hennig, J., Seifritz, E., 2003. Temporal integration of sequential auditory events: silent period in sound pattern activates human planum temporale. NeuroImage 20 (1), 429–434.

Näätänen, R., 1992. Attention and Brain Function. Lawrence Erlbaum Associates, Hillsdale, NJ.

Naatanen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Alho, K., 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385 (6615), 432–434.

Newport, E.L., Aslin, R.N., 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. Cogn. Psychol. 48 (2), 127–162.

Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J.B., 2005. Valid conjunction inference with the minimum statistic. NeuroImage 25 (3), 653–660.

Obleser, J., Elbert, T., Lahiri, A., Eulitz, C., 2003. Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. Brain Res. Cogn. Brain Res. 15 (3), 207–213.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Rauschecker, J.P., 2006. Vowel sound extraction in anterior superior temporal cortex. Hum. Brain Mapp. 27 (7), 562–571.

Obleser, J., Zimmermann, J., Van Meter, J., Rauschecker, J.P., 2007. Multiple stages of auditory speech perception reflected in event-related FMRI. Cereb. Cortex 17 (10), 2251–2257.

Obleser, J., Leaver, A.M., Vanmeter, J., Rauschecker, J.P., 2010. Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. Front. Psychol. 1, 232.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9 (1), 97–113.

Overath, T., Cusack, R., Kumar, S., von Kriegstein, K., Warren, J.D., Grube, M., Griffiths, T.D., 2007. An information theoretic characterisation of auditory encoding. PLoS Biol. 5 (11), e288.

Pelucchi, B., Hay, J.F., Saffran, J.R., 2009a. Learning in reverse: eight-month-old infants track backward transitional probabilities. Cognition 113 (2), 244–247.

Pelucchi, B., Hay, J.F., Saffran, J.R., 2009b. Statistical learning in a natural language by 8-month-old infants. Child Dev. 80 (3), 674–685.

Pena, M., Bonatti, L.L., Nespor, M., Mehler, J., 2002. Signal-driven computations in speech processing. Science 298 (5593), 604–607.

Petkov, C.I., Kang, X., Alho, K., Bertrand, O., Yund, E.W., Woods, D.L., 2004. Attentional modulation of human auditory cortex. Nat. Neurosci. 7 (6), 658–663.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat. Neurosci. 11 (3), 367–374.

Petkov, C.I., Logothetis, N.K., Obleser, J., 2009. Where are the human speech and voice regions, and do other animals have anything like them? Neuroscientist 15 (5), 419–429.

Poeppel, D., Phillips, C., Yellin, E., Rowley, H.A., Roberts, T.P., Marantz, A., 1997. Processing of vowels in supratemporal auditory cortex. Neurosci. Lett. 221 (2–3), 145–148.

Preacher, K.J., Hayes, A.F., 2004. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav. Res. Methods Instrum. Comput. 36 (4), 717–731.

Price, C., Thierry, G., Griffiths, T., 2005. Speech-specific auditory processing: where is it? Trends Cogn. Sci. 9 (6), 271–276.

Raizada, R.D., Poldrack, R.A., 2007. Selective amplification of stimulus differences during categorical processing of speech. Neuron 56 (4), 726–740.

Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2 (1), 79–87.

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. 12 (6), 718–724.

Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc. Natl. Acad. Sci. U. S. A. 97 (22), 11800–11806.

Reil, J.C., 1809. Das balken-system oder die balken-organisation im großen gehirn. Arch. Physiol. 9, 172–195.
Riecke, L., van Opstal, A.J., Goebel, R., Formisano, E., 2007. Hearing illusory sounds in noise: sensory-perceptual transformations in primary auditory cortex. J. Neurosci. 27 (46), 12684–12689.
Rivier, F., Clarke, S., 1997. Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. NeuroImage 6 (4), 288–304.
Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B Biol. Sci. 336 (1278), 367–373.
Rosenthal, R., Rosnow, R.L., Rubin, D.B., 2000. Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach. Cambridge University Press, Cambridge.
Saad, Z.S., Glen, D.R., Chen, G., Beauchamp, M.S., Desai, R., Cox, R.W., 2009. A new method for improving functional-to-structural MRI alignment using local Pearson correlation. NeuroImage 44 (3), 839–848.
Saffran, J.R., Aslin, R.N., Newport, E.L., 1996. Statistical learning by 8-month-old infants. Science 274 (5294), 1926–1928.
Saffran, J.R., Johnson, E.K., Aslin, R.N., Newport, E.L., 1999. Statistical learning of tone sequences by human infants and adults. Cognition 70 (1), 27–52.
Saffran, J.R., Pollak, S.D., Seibel, R.L., Shkolnik, A., 2007. Dog is a dog is a dog: infant rule learning is not specific to language. Cognition 105 (3), 669–680.
Scheich, H., Baumgart, F., Gaschler-Markefski, B., Tegeler, C., Tempelmann, C., Heinze, H.J., Stiller, D., 1998. Functional magnetic resonance imaging of a human auditory cortex area involved in foreground-background decomposition. Eur. J. Neurosci. 10 (2), 803–809.
Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E., 2009. Sound categories are represented as distributed patterns in the human auditory cortex. Curr. Biol. 19 (6), 498–502.
Sueur, J., Aubin, T., Simonis, C., 2008. Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics 18, 213–226.
Sweet, R.A., Dorph-Petersen, K.A., Lewis, D.A., 2005. Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. J. Comp. Neurol. 491 (3), 270–289.
Tardif, E., Clarke, S., 2001. Intrinsic connectivity of human auditory areas: a tracing study with DiI. Eur. J. Neurosci. 13 (5), 1045–1050.
Tian, B., Reser, D., Durham, A., Kustov, A., Rauschecker, J.P., 2001. Functional specialization in rhesus monkey auditory cortex. Science 292 (5515), 290–293.
Tobia, M.J., Iacovella, V., Hasson, U., 2012a. Multiple sensitivity profiles to diversity and transition structure in non-stationary input. NeuroImage 60 (2), 991–1005.
Tobia, M.J., Iacovella, V., Davis, B., Hasson, U., 2012b. Neural systems mediating recognition of changes in statistical regularities. NeuroImage 63 (3), 1730–1742.
Tremblay, P., Deschamps, I., Gracco, V.L., in press. Regional heterogeneity in the processing and the production of speech in the human planum temporale Cortex (Electronic publication ahead of print).
von Economo, C., Horn, L., 1930. ber Windungsrelief, Masse und Rindenarchitektonik der Supratemporalfläche, ihre individuellen und ihre Seitenunterschiede. Z. Neurol. Psychiatr. 130, 678–757.
Yabe, H., Tervaniemi, M., Reinikainen, K., Naatanen, R., 1997. Temporal window of integration revealed by MMN to sound omission. Neuroreport 8 (8), 1971–1974.
Yabe, H., Tervaniemi, M., Sinkkonen, J., Huotilainen, M., Ilmoniemi, R.J., Naatanen, R., 1998. Temporal window of integration of auditory information in the human brain. Psychophysiology 35 (5), 615–619.
Zaitsev, M., Hennig, J., Speck, O., 2004. Point spread function mapping with parallel imaging techniques and high acceleration factors: fast, robust, and flexible method for echo-planar imaging distortion correction. Magn. Reson. Med. 52 (5), 1156–1166.