Article

Cortical encoding of speech enhances task-relevant acoustic information

RUTTEN, Sanne, *et al*.

Abstract

Speech is the most important signal in our auditory environment, and the processing of speech is highly dependent on context. However, it is unknown how contextual demands influence the neural encoding of speech. Here, we examine the context dependence of auditory cortical mechanisms for speech encoding at the level of the representation of fundamental acoustic features (spectrotemporal modulations) using model-based functional magnetic resonance imaging. We found that the performance of different tasks on identical speech sounds leads to neural enhancement of the acoustic features in the stimuli that are critically relevant to task performance. These task effects were observed at the earliest stages of auditory cortical processing, in line with interactive accounts of speech processing. Our work provides important insights into the mechanisms that underlie the processing of contextually relevant acoustic information within our rich and dynamic auditory environment.

Reference

RUTTEN, Sanne, *et al*. Cortical encoding of speech enhances task-relevant acoustic information. *Nature Human Behaviour*, 2019, vol. 3, no. 9, p. 974-987

DOI : 10.1038/s41562-019-0648-9
PMID : 31285622

ARTICLES
https://doi.org/10.1038/s41562-019-0648-9

nature human behaviour

Corrected: Publisher Correction

Cortical encoding of speech enhances task-relevant acoustic information

Sanne Rutten[1]*, Roberta Santoro[1], Alexis Hervais-Adelman[1,2], Elia Formisano[3,4,5,6] and Narly Golestani[1,6]

Speech is the most important signal in our auditory environment, and the processing of speech is highly dependent on context. However, it is unknown how contextual demands influence the neural encoding of speech. Here, we examine the context dependence of auditory cortical mechanisms for speech encoding at the level of the representation of fundamental acoustic features (spectrotemporal modulations) using model-based functional magnetic resonance imaging. We found that the performance of different tasks on identical speech sounds leads to neural enhancement of the acoustic features in the stimuli that are critically relevant to task performance. These task effects were observed at the earliest stages of auditory cortical processing, in line with interactive accounts of speech processing. Our work provides important insights into the mechanisms that underlie the processing of contextually relevant acoustic information within our rich and dynamic auditory environment.

Speech is the most crucial acoustic signal in our daily life. It conveys information for interpersonal communication as well as for the recognition of the identity and emotional state of an individual[1]. The type of information that is most relevant depends on the specific context and the momentary behavioural goal[2]. Consequently, speech processing needs to be highly adaptive and efficient[2]. Such efficiency and adaptiveness is achieved in the human brain through the active interplay between bottom-up processing of the physical input in early sensory areas, and top-down modulatory mechanisms driven by upstream auditory and non-auditory (such as frontal) areas[3,4]. Interactive speech models therefore propose that initial bottom-up processing is implemented on the incoming input, activating multiple possible linguistic representations of the sound[5,6]. Simultaneously, higher-level speech-recognition mechanisms exert inhibitory influences on these competing interpretations, ultimately resulting in the activation of the correct interpretation[6]. Top-down effects are therefore thought to alter the bottom-up processing of speech sounds. However, it is unclear whether—and in what manner—these top-down modulations alter the neural representations of the acoustic content of speech (referred to as speech sound encoding hereafter). It is also unknown where these changes occur in the cortical processing pathway.

On the basis of previous studies in animals and humans, it has been suggested that cortical sound encoding can be characterized by a set of modulation filters[7–10]. After initial frequency decomposition in the cochlea, sounds are decomposed during subcortical and cortical processing with respect to their joint spectral and temporal modulation content[8]. This decomposition provides a multiresolution representation of sounds. Evidence suggests that phonetic information is encoded in this multidimensional space in the human superior temporal gyrus (STG)[11]. Multiple studies in animals have shown that neural spectrotemporal sensitivities in primary auditory areas are flexible and can dynamically adapt to task demands, task difficulty or learned associations[12–15]. The extent to which these top-down influences on early auditory processing occur in the human brain remains to be explored.

In humans, top-down effects on the neural processing of speech have been found, but in higher-level areas, such as the inferior parietal or frontal cortices or in (auditory) association areas located in the posterior STG (postSTG) and sulcus[2,16,17]. Top-down influences have been shown to modulate specific aspects of neural speech representations[18]. For example, one study has shown that critical spectrotemporal features of incoming speech signals can be more accurately retrieved from neuronal responses within the posterior superior temporal lobe if the speech is attended to, compared to the features that can be retrieved from simultaneously presented unattended speech[18]. Other studies have shown that previous exposure to similar or identical stimuli directly influences the bottom-up representations of speech sounds within the posterior temporal lobe. For example, in the context of perceptual restoration of missing phonemes, linguistic predictions coming from top-down lexical/linguistic knowledge modulate neuronal responses in the postSTG before the onset of the critical phoneme. These modulations have been shown to enhance the representation of acoustic energy specifically related to the restored phoneme[4]. Another study has shown that an increase in the intelligibility of degraded speech—arising after exposure to the corresponding intact speech signal—evokes an amplification of the representation of spectrotemporal features that are characteristic of speech[19].

The above studies show that the neural mechanisms that underlie speech processing actively adapt to task demands, attention and previous semantic knowledge. Moreover, the two previous studies[4,19] show that top-down effects—driven by previous exposure—modulate specific aspects of how acoustic information is encoded within the temporal lobe. These dynamic changes in sound encoding have mostly been found in the postSTG and sulcus and/or in the inferior frontal gyrus—regions upstream of the primary auditory cortex. However, interactive models of speech processing predict

[1]Brain and Language Lab, Department of Psychology, Faculty of Psychology and Education Sciences, University of Geneva, Geneva, Switzerland. [2]Neurolinguistics, Department of Psychology, University of Zurich, Zurich, Switzerland. [3]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands. [4]Maastricht Brain Imaging Center, Maastricht, the Netherlands. [5]Maastricht Centre for Systems Biology, Maastricht University, Maastricht, the Netherlands. [6]These authors jointly supervised this work: Elia Formisano, Narly Golestani.
*e-mail: Sanne.Rutten@unige.ch

that top-down mechanisms affect even early auditory pre-lexical levels of processing, enabling early processes to tune to or amplify the acoustic features that are critically relevant to the processing of speech[5,6]. According to these models, it could be expected that task requirements modulate the encoding of speech sounds already in early auditory cortical areas (that is, in Heschl's gyrus (HG) and Heschl's sulcus (HS)). However, there is inconsistent evidence for task-dependent modulation of activation in primary auditory areas. For example, performing semantic categorization tasks on speech stimuli minimally modulates activity in HG[20–22]. By contrast, other studies using pattern-classification techniques have shown that response patterns evoked by speech sounds in early auditory areas are modulated by task demands, perception and learning[23–26]. However, none of these latter studies related these observations to the processing of specific features of speech input. Recent developments in model-based functional magnetic resonance imaging (fMRI) analysis[27,28] now enable us to relate spatially distributed neural sound representations to the specific acoustic features that underlie task requirements[8,29,30].

In the present study, we used model-based fMRI to examine how context—which we operationalized as the execution of different tasks—modulates the encoding of speech throughout the human auditory cortex. Specifically, we investigated how the neural encoding of the same speech sounds changes as a function of preferential processing of different acoustic features within the sounds. During high-resolution fMRI measurements, participants performed either a linguistic (identification of stop consonants) or a paralinguistic (speaker identification) task on identical speech stimuli (pseudo-words with similarities to French phonology but containing no meaning). Then, we used model-based decoding[30] to examine the acoustic energy that is encoded by the brain along three acoustic dimensions: frequency, spectral modulation and temporal modulation. These acoustic dimensions are differentially important for characterizing specific aspects of linguistic versus paralinguistic information in the speech signal. For example, plosive consonants (such as /p/, /t/ and /k/) have sudden and spectrally broad bursts, whereas voice processing—or speaker identification—relies more heavily on fine spectral detail and pitch[31].

Consequently, to accurately perform these respective tasks, participants needed to focus on different types of acoustic information in the sounds. We therefore expected that this would be reflected in dynamic changes to the encoding of identical auditory input. Specifically, on the basis of previous findings on the neural processing of different aspects of speech sounds[11,31,32], we hypothesized that performance of the speaker task would result in the preferential encoding of higher spectral modulations, and that performance of the phoneme task would result in the preferential encoding of lower spectral modulations and faster temporal modulations. Moreover, using regionally specific analyses, we examined task-driven modulation of neural encoding across different auditory sub-regions, and assessed whether such modulation occurs even in early auditory areas.

Results

Identification tasks and behavioural performance. In the fMRI scanner (7 T), participants performed a phoneme- and a speaker-identification task on the same pseudo-words (see the 'Task and stimuli' section in the Methods). During the speaker-identification task, participants were asked to indicate whether a stimulus was spoken by speaker 1, speaker 2 or speaker 3, whereas during the phoneme-identification task, the participants heard the same pseudo-words but were asked to indicate whether they contained a /p/, /t/ or /k/ sound. Pseudo-words were used to diminish the use of lexical information for anticipating the presence of a target sound and to promote reliance on the auditory input during the phoneme task. To specifically guide the acoustic focus towards spectral information during the speaker task, we did not use actual different speakers within this study. Instead we created the percept of three different speakers by manipulating the fundamental frequency of the pseudo-words recorded from one female speaker (see the 'Task and stimuli' section in the Methods). Performance in the scanner was well above chance for both tasks (mean ± s.e.m. for the speaker task = 88.8% ± 2%, and for the phoneme task = 96.5% ± 0.9%); however, participants had more difficulty in identifying the different speakers compared with the different phonemes ($t_{12} = -4.193$, $P = 0.001$ (two-tailed), difference (mean ± s.e.m.) = −7.7% ± 1.8%, 95% confidence interval (CI) = −11.7% to −3.71%; Supplementary Fig. 1). There were no significant differences in performance on the different targets within each task (speaker task: $F_{2,24} = 0.852$, $P = 0.439$; phoneme task: $F_{2,24} = 0.320$, $P = 0.729$; Supplementary Fig. 1).

Speaker and phoneme modulation profiles based on the stimuli. The aim of this study was to examine whether identical speech sounds are encoded differently in the auditory cortex when different tasks are performed on the sounds. We therefore determined which acoustic aspects of the sounds themselves were the most informative for performance of the respective tasks. To do this, we modelled the acoustic energy of our stimuli using a model that mimics cortical sound representations[30,33]. This sound representation model represents the energy of the sounds along the following acoustic dimensions: frequency (f), spectral modulation (Ω) and temporal modulation (ω; see the 'Sound representation model for processing of speech sounds' section in the Methods; Supplementary Fig. 2a). Consequently, we evaluated the acoustic energy of the sounds in a three-dimensional modulation space that encompasses different sound features, with each sound feature representing the time-averaged acoustic energy at a specific frequency and at specific spectral and temporal modulations.

Next, we analysed the relative importance of different sound features for the identification of the different targets in the respective tasks by assessing which sound features contained significant acoustic variation within each target (see the 'Estimation of speaker and phoneme modulation profiles based on the stimuli' section in the Methods). This provided target-specific modulation profiles that represent the sound features that are the most informative for identifying each target. After statistical validation (see the 'Estimation of speaker and phoneme modulation profiles based on the stimuli' section in the Methods), marginal modulation profiles were obtained by calculating an average of the acoustic variation across the irrelevant acoustic dimension.

Generally, the modulation profiles of the different targets were characterized by multiple frequency-specific spectrotemporal modulations (Fig. 1). However, the acoustic variation in the modulation profiles of the different speakers were most pronounced at higher spectral modulations (>1.1 cycles per octave) for centre frequencies of up to 0.8 kHz and above 2.9 kHz. These profiles did not show high acoustic variation for specific temporal modulations (see Fig. 1a and Supplementary Fig. 3 for three-dimensional representations of the speaker profiles). By contrast, the modulation profiles of the different phonemes could mostly be characterized by acoustic variation at fast rates of temporal modulation (>7.8 Hz up and down) and at broader spectral modulations for a wide range of centre frequencies, but mostly for frequencies above 0.6 kHz (see Fig. 1a and Supplementary Fig. 3 for three-dimensional representations of the phoneme profiles). The pattern of greater acoustic variation at faster temporal modulations and broader spectral modulations became more pronounced when the profiles were calculated on the basis of the parts of the speech samples that corresponded to the target phonemes only (Supplementary Fig. 4).

We further examined which acoustic variations were characteristic for each target class (that is, for all speakers or all phonemes).

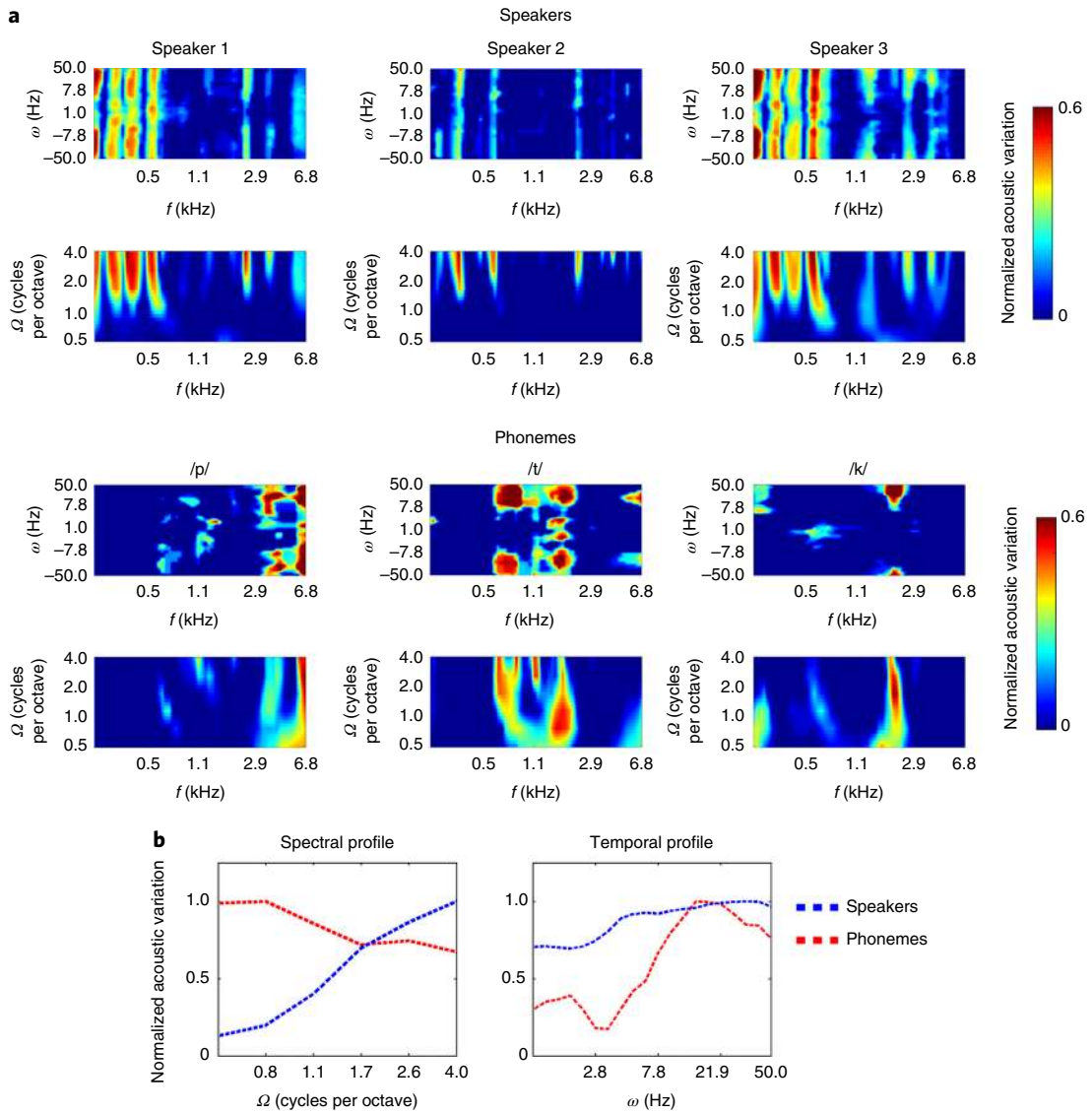

Fig. 1 | Target-specific modulation profiles for the six individual target sounds. a, Two-dimensional modulation profiles for the individual target sounds, showing the normalized acoustic variation of the individual speakers (top two rows) and phonemes (bottom two rows). The profiles show the sound features that had significant acoustic variation (*t*-test with $P < 0.05$, uncorrected), and the colour code indicates the mean acoustic variation (normalized) across the sounds belonging to one target. Non-significant sound features are shown in blue. The speakers showed the highest acoustic variation for higher spectral modulations (>1.1 cycles per octave) at centre frequencies of up to 0.8 kHz. The energy variation was not selective for specific rates of temporal modulation. By contrast, the phonemes showed the highest variation at fast rates of temporal modulation (>7.8 Hz up and down) and at broad spectral modulations for centre frequencies above 0.6 kHz. Plots are interpolated for display purposes. **b**, One-dimensional modulation profiles showing the mean acoustic variation (normalized) across the three speakers (blue) and the three phonemes (red). The spectral profiles indicate that the speakers showed increasingly higher acoustic variation towards higher spectral modulations. The phonemes showed the greatest acoustic variation at the lowest scale (0.5 cycles per octave), and the variation decreased for higher spectral modulations. The temporal modulation profiles for the different speakers showed high variation across all rates of temporal modulation, and did not show specificity for particular rates of temporal modulation. The temporal profiles for the different phonemes showed clear increases in variation at faster rates of temporal modulation, with the peak at 18 Hz. Modulation profiles are normalized for each target class for display purposes. Temporal modulation profiles are averaged across upward and downward temporal modulations (see the 'Estimation of speaker and phoneme modulation profiles based on the stimuli' section in the Methods).

To do this, we initially calculated the average across the profiles of the individual speakers or the individual phonemes, and then created one-dimensional (frequency-unspecific) modulation profiles by calculating the average along the two irrelevant acoustic dimensions (for the temporal profiles we also calculated the average across upward and downward modulations; see the 'Estimation of speaker and phoneme modulation profiles based on the stimuli' section in the Methods).

The spectral modulation profiles showed that the speakers have increasingly higher acoustic variation towards higher spectral modulations, whereas the phonemes have highest variation at the lowest scale (Fig. 1b). The temporal modulation profiles revealed that the speakers do not show a clear increase in acoustic variation at specific modulation rates. By contrast, the phonemes show an increase in acoustic variation towards faster rates of temporal modulation (peak at 18 Hz; Fig. 1b).

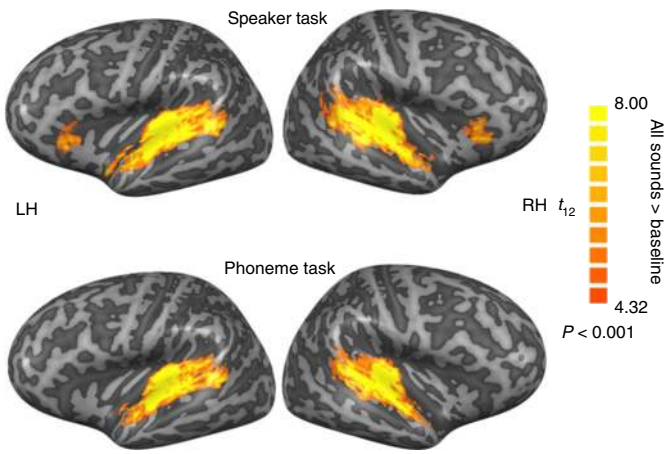

Fig. 2 | Activations evoked by speech sounds during the speaker and phoneme tasks. The maps show the functional contrast during the performance of the speaker (top; speaker task > baseline) and of the phoneme tasks (bottom; phoneme task > baseline). The maps are shown on inflated group surface reconstructions of the LH and RH after alignment of the cortices of all participants ($n = 13$). The activation maps were corrected for multiple comparisons ($P_{corr} < 0.05$) by applying a cluster size thresholding procedure on the basis of permutation and an initial uncorrected $P = 0.001$ (see the 'fMRI analysis for univariate group contrasts' section in the Methods).

Given the importance of these different sound features in the stimuli for differentiating between speakers and between phonemes, we expected that neural encoding of higher spectral modulations would be amplified during the speaker task, and that neural encoding of lower spectral and faster temporal modulations would be amplified during the phoneme task.

Auditory cortical responses during the speaker and phoneme tasks. The participants ($n = 13$) performed the two tasks while their sound-evoked neural responses were measured using fMRI. The speech sounds evoked significant blood-oxygen-level-dependent (BOLD) responses in a wide expanse of the bilateral superior temporal cortex, including HG, HS, planum temporale (PT), planum polare (PP), STG and superior temporal sulcus (STS; Fig. 2). Beyond the auditory cortex, activation was also found in the insula and orbitofrontal cortex. Despite the significant difference in behavioural performance on the two tasks (see 'Identification tasks and behavioural performance' section above), a general linear model (GLM) contrast analysis showed that the speaker task yielded higher responses in only a small cluster in the right PT (cluster-size: 39 vertices, t-statistic (mean cluster) = 5.1; Supplementary Fig. 5). The phoneme task did not evoke enhanced responses compared with the speaker task (no significant voxels after correction for multiple comparisons (corrected P value (P_{corr}) < 0.05) using a cluster size thresholding procedure on the basis of permutation and an initial uncorrected $P = 0.001$).

Neural encoding of speech sounds in auditory regions of interest. We examined the neural encoding of the sounds in six different auditory areas by reconstructing the sound features of the stimuli from the fMRI responses that we obtained during task performance. On the cortical surface reconstruction of each participant, we manually labelled the six following regions of interest (ROIs): HG, PT, PP, anterior STG (antSTG), middle STG (midSTG) and postSTG (Fig. 3; see the 'Delineation of anatomical ROIs' section in the Methods for delineation criteria). For each ROI and each hemisphere separately, we trained a linear decoder to reconstruct the acoustic energy of every sound feature, as defined by the sound representation model[30] (see the 'Estimation of the linear decoders' section in the Methods and Supplementary Fig. 2b–d for an overview of the modelling steps and Supplementary Fig. 6 for the number of voxels used per ROI). Specifically, for each sound feature, we trained a linear decoder on the fMRI responses within each ROI to predict the acoustic energy for that given feature. Reconstruction accuracies (that is, prediction accuracies) were assessed by computing the Pearson's correlation coefficient (r) between the actual acoustic energy for each sound feature in the test sounds and the energy of that feature as predicted by the decoder (see the 'Estimation of ROI-specific MTFs for each participant' section in the Methods; Supplementary Fig. 2). Group reconstruction accuracies for all of the features of the sound representation model resulted in a task-specific modulation transfer function (MTF). The MTF describes reconstruction accuracies as a function of frequency and of spectral and temporal modulations, and thus provides an overview by which sound features could be accurately reconstructed from the fMRI responses. All group MTFs were statistically validated and thresholded (see the 'Estimation of ROI-specific MTFs at the group-level' section in the Methods and Supplementary Fig. 7a for the distributions of reconstruction accuracies for all sound features). After statistical validation of the MTFs, we obtained marginal modulation profiles by calculating the average of the reconstruction accuracies along the irrelevant acoustic dimension.

During both tasks, the sound features of the stimuli could be accurately reconstructed for a broad range of spectral and temporal modulations; however, reconstructions for faster rates of temporal modulation (>10 Hz up and down) and for centre frequencies between 0.5 kHz and 1.7 kHz were generally better (see Fig. 4 and Supplementary Fig. 8 for three-dimensional representations of the MTFs). For the speaker task, a broader range of centre frequencies could be accurately reconstructed compared with the phoneme task (Fig. 4b). By contrast, the marginal profiles for the phoneme task show a clearer pattern of peak reconstruction accuracies for faster rates of temporal modulation (>10 Hz). For example, the left PT and the PP show clearer segregation in reconstruction accuracies between slower and faster temporal modulations compared with those seen during the speaker task for corresponding ROIs (Fig. 4c).

Analysis of task differences using single sound features. To determine which brain regions encoded the sounds differently as a function of task, we tested for task differences in the MTFs of the different ROIs. For this, we examined whether reconstruction accuracies for specific sound features were higher for one task than for the other (see the 'Estimation of task differences between group MTFs' section in the Methods). For ROIs that did not show hemispheric differences, the task effects were examined using the MTFs that were averaged across hemispheres, whereas for ROIs that did show hemispheric differences (PT and midSTG; Supplementary Results 1), we examined the task effects for the two hemispheres separately (see the 'Estimation of task differences between group MTFs' section in the Methods).

Results showed task differences in the MTFs for five out of the six ROIs. During the speaker task, reconstruction accuracies for specific sound features were higher within the MTFs of the following ROIs: bilateral HG, PP and postSTG, right PT and left midSTG. During the phoneme task, we found higher reconstruction accuracies within the MTFs of bilateral postSTG and right midSTG (see Fig. 5 and Supplementary Fig. 9 for unthresholded differences, Supplementary Fig. 10 for three-dimensional representations and Supplementary Fig. 11 for task effects in voice-selective areas).

We further examined which specific spectral and temporal modulations were encoded differently across multiple ROIs by computing task-specific spectral and temporal modulation profiles as follows. For each participant, we averaged the reconstruction accuracies of the sound features across the ROIs that showed significant

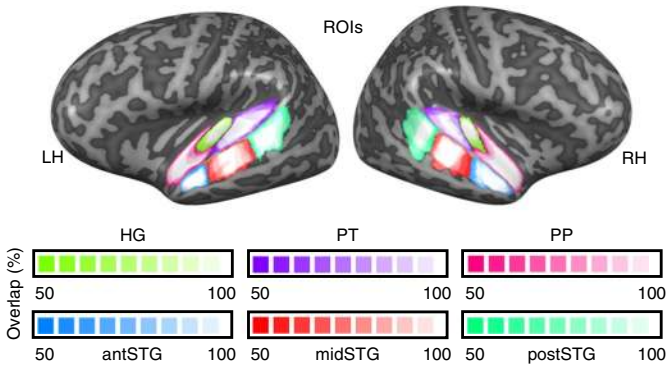

Fig. 3 | Probabilistic maps of the ROIs. Probabilistic maps, colour-coded for each ROI, showing the manually labelled ROIs across participants overlaid on inflated group surface reconstructions of the LH and RH. The ROI-specific colour scales indicate the percentage of overlap of the respective ROIs across all the participants ($n=13$).

task differences (that is, across HG, PT (right hemisphere (RH)), midSTG (left hemisphere (LH)) and postSTG for the speaker task, and across midSTG (RH) and postSTG for the phoneme task; see the 'Comparison of task-specific spectral and temporal modulation profiles' section in the Methods). One-dimensional profiles were then obtained by calculating the average of the reconstruction accuracies across the irrelevant acoustic dimensions.

We tested for differences in reconstruction accuracies between the spectral profiles of the speaker and of the phoneme tasks at each scale and, similarly, for the temporal profiles at each modulation rate (see the 'Comparison of task-specific spectral and temporal modulation profiles' section in the Methods). Results showed that reconstruction accuracies during the speaker task were significantly higher than those during the phoneme task for spectral modulations above 1.1 cycles per octave ($\Omega=1.7$: $t_{12}=3.548$, $P=0.004$; $\Omega=2.6$: $t_{12}=4.225$, $P=0.001$; $\Omega=4.0$: $t_{12}=5.092$, $P<0.001$; all tests were Bonferroni-corrected for the number of tests; Fig. 6). By contrast, reconstruction accuracies during the phoneme task were significantly higher than those during the speaker task at the lowest spectral scale ($\Omega=0.5$: $t_{12}=-6.387$, $P<0.001$; Fig. 6). Results for the temporal modulations showed that accuracies overlapped between the two tasks at faster rates of temporal modulation reconstruction, whereas at slower rates of temporal modulation reconstruction accuracies were significantly higher for the speaker task than the phoneme task ($\omega=1.0$: $t_{12}=4.932$, $P<0.001$; $\omega=1.2$: $t_{12}=4.960$, $P<0.001$; $\omega=2.8$: $t_{12}=5.361$, $P<0.001$; $\omega=3.4$: $t_{12}=4.725$, $P<0.001$; Fig. 6). These differences indicate that the encoding of fast temporal modulations is more specific to the phoneme task compared with the encoding of a broader range of rates of temporal modulation during the speaker task.

Finally, we examined whether the differences in the neural encoding of the sounds observed in the tasks showed amplification of sound features that were relevant to task performance. On the basis of the profiles that we obtained from the stimuli (Fig. 1b), we expected that distinct spectral and temporal information would be informative for the two tasks. We therefore first correlated the spectral profiles of the two tasks obtained from the neural data with the spectral profiles that we obtained from the stimuli (see the 'Comparison of task-specific spectral and temporal modulation profiles' section in the Methods). We found a greater correlation between the spectral profile of the speakers in the stimuli and the neural spectral profiles obtained during the speaker task compared with the correlation between those obtained during the phoneme task ($t_{12}=9.298$, $P<0.001$ (two-tailed), mean difference (s.e.)=1.35 (0.15), 95% CI=1.03–1.67). Similarly, we found a greater correlation between the spectral profile of the phonemes in the stimuli and the neural spectral profiles obtained during the phoneme task compared with the correlation of those obtained during the speaker task ($t_{12}=9.818$, $P<0.001$ (two-tailed), mean difference (s.e.)=1.39 (0.14), 95% CI=1.08–1.70). We also expected to observe a greater reliance on faster temporal information during the phoneme task. Consistent with this prediction, we found a greater correlation between the temporal profile of the phonemes in the stimuli and the temporal profiles that we obtained from neural data obtained during the phoneme task compared with the correlation of those obtained during the speaker task ($t_{12}=6.510$, $P<0.001$ (two-tailed), mean difference (s.e.)=0.82 (0.13), 95% CI=0.55–1.10).

Analysis of target separability using multiple sound features. In addition to analysing task-related differences in neural encoding at the level of single-sound features, we also tested for amplification of task-relevant information at the whole-sound level. More specifically, we expected that target identification would enhance the neural separability between targets. Consequently, sounds belonging to different speakers would be most separable during the speaker task, and sounds belonging to different phonemes would be most separable during the phoneme task. We tested this prediction using linear classification (support vector machine (SVM)) on the reconstructed sounds, which we derived from the fMRI responses during performance of each task. For each individual, we trained SVMs to discriminate between the reconstructed sounds spoken by the different speakers regardless of the phoneme, and to discriminate between reconstructed sounds containing the different target phonemes regardless of the speaker (see the 'Analysis of target separability with the use of linear classification' section in the Methods). For each task, classifications were performed on all positively reconstructed sound features and separately for the two hemispheres (mean classification accuracies of speaker identity for the speaker task: $t_{12}=6.385$, $P<0.001$ (two-tailed), mean (s.e.)=0.54 (0.006), 95% CI=0.52–0.55; and for the phoneme task: $t_{12}=0.748$, $P=0.469$ (two-tailed), mean (s.e.)=0.50 (0.005), 95% CI=0.49–0.51. Mean classification accuracies of phoneme identity for the speaker task: $t_{12}=3.182$, $P<0.008$ (two-tailed), mean (s.e.)=0.51 (0.005), 95% CI=0.50–0.52; and for the phoneme task: $t_{12}=6.972$, $P<0.001$ (two-tailed), mean (s.e.)=0.53 (0.005), 95% CI=0.52–0.54; Fig. 7a).

We expected that the reconstructed sounds belonging to different targets would be more separable when they were relevant to task performance, which would be reflected by higher classification accuracies for task-relevant targets compared with classifications for task-irrelevant targets. We used repeated-measures analysis of variance (ANOVA) to test whether the classification accuracies differed between the target class (speakers or phonemes), task (speaker or phoneme task), ROI and hemisphere. We found an interaction between target class and task ($F_{1,12}=17.714$; $P=0.001$), as well as a three-way interaction between target class, task and ROI ($F_{5,8}=6.516$; $P=0.011$). When we tested each ROI separately, we found a main effect of the task only in the antSTG, with higher classification accuracies for the speaker task for both speakers and phonemes ($F_{1,12}=5.869$, $P=0.032$; Fig. 7b). We did not find such an effect of task in any other ROI (HG: $F_{1,12}=0.690$, $P=0.422$; PT: $F_{1,12}=0.915$, $P=0.358$; PP: $F_{1,12}=0.015$, $P=0.904$; midSTG: $F_{1,12}=0.116$, $P=0.739$; postSTG: $F_{1,12}=2.162$, $P=0.167$; Fig. 7b). Moreover, in accordance with our predictions, we found significant interactions between target class and task in the following ROIs: HG, PT, midSTG and postSTG (HG: $F_{1,12}=26.078$, $P<0.001$; PT: $F_{1,12}=7.638$, $P=0.017$; midSTG: $F_{1,12}=5.077$, $P=0.044$; postSTG: $F_{1,12}=15.161$, $P=0.002$). Within these ROIs, speaker classification accuracies were higher for the sounds that were reconstructed from the fMRI responses during the speaker task than for those of the phoneme task. The opposite was found for phoneme classification (Fig. 6c). These results show that the acoustic representations of the sounds change as a function of task requirements. For the

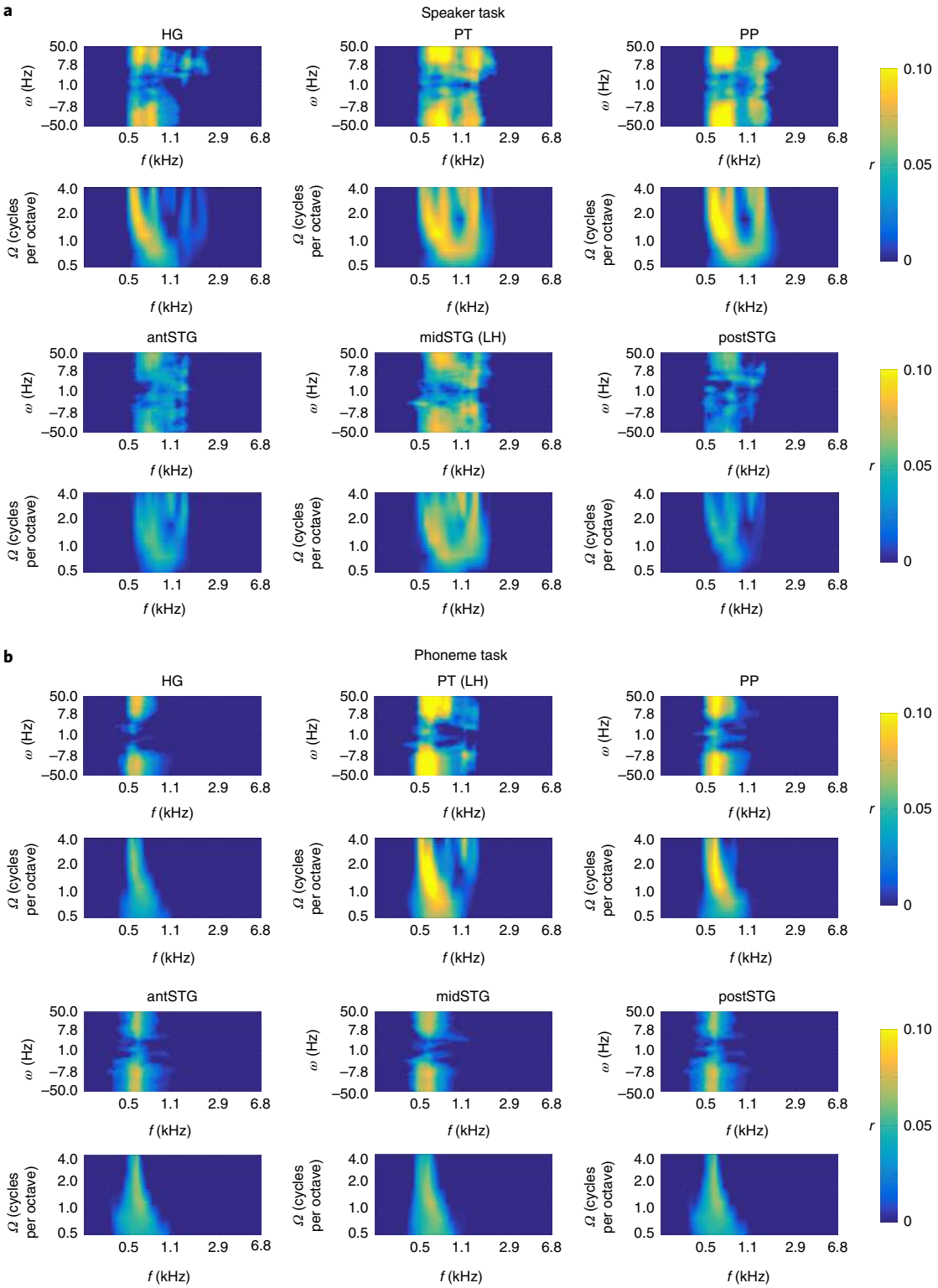

Fig. 4 | Marginal modulation profiles of the MTFs during the speaker and phoneme tasks. a,**b**, Two-dimensional marginal modulation profiles showing the reconstruction accuracies of the sound features during the speaker (**a**) and the phoneme tasks (**b**), for each ROI. The colour code indicates the group average *r* between the predicted acoustic energy and the actual energy of the sound features. **a**, During the speaker task, peak reconstruction accuracies were found for faster temporal modulations (>10 Hz up and down) and for higher spectral modulations (4 cycles per octave) at centre frequencies between 0.5 kHz and 1.7 kHz. **b**, During the phoneme task, peak reconstruction accuracies were found for fast temporal modulations (>10 Hz up and down) at centre frequencies between 0.4 kHz and 1.0 kHz. All correlations were statistically validated and thresholded (cluster-size threshold yielding $\alpha_{clu} < 0.05$; see the 'Estimation of ROI-specific MTFs at the group level' section in the Methods). Non-significant correlations are shown in blue. The plots have been interpolated for display purposes. Negative ω indicates upward temporal modulations and positive ω indicates downward temporal modulations. Before statistical assessment, all correlation values were normalized by applying the Fisher *z*-transformation.

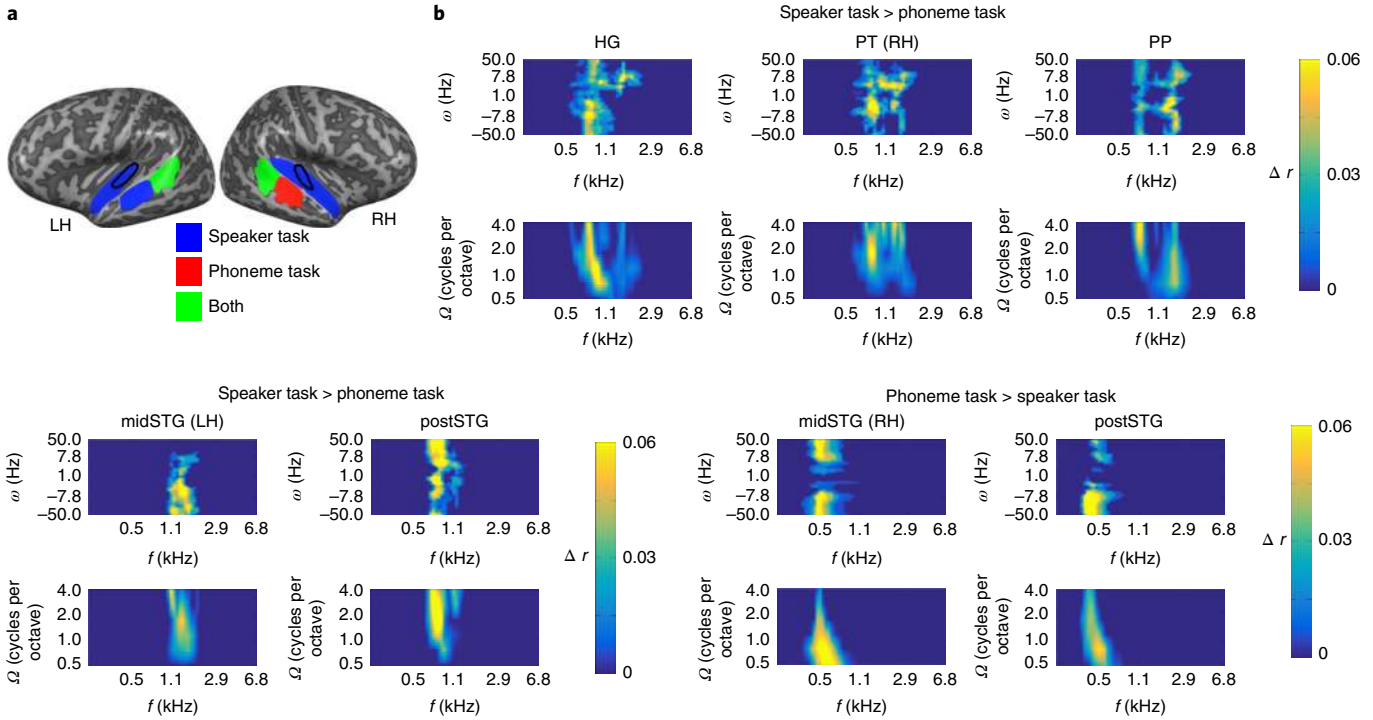

Fig. 5 | Marginal modulation profiles of the task-difference MTFs. a, ROIs that showed task modulation effects were overlaid on inflated group surface reconstructions of the LH and RH. ROIs with increased reconstruction accuracies for only the speaker task are shown in blue, ROIs with higher reconstruction accuracies for only the phoneme task are shown in red and ROIs with higher reconstruction accuracies for both tasks are shown in green. HG is outlined with a black line. **b**, Two-dimensional marginal modulation profiles of the task-difference MTFs for HG, PT and PP (top row), and for midSTG and postSTG (bottom row). The colour scales indicate the group-average differences in reconstruction accuracy (difference in *r*) between the speaker and phoneme tasks (ROIs that showed no task differences are not shown). The spectral profiles during the speaker task (speaker task > phoneme task) showed increased reconstruction accuracies for higher spectral modulations (>1.1 cycles per octave), whereas the spectral profiles for the phoneme task (phoneme task > speaker task) showed increased reconstruction accuracies for lower spectral modulations (<0.8 cycles per octave). The temporal profiles showed a reversed pattern, whereby the phoneme task had the highest reconstruction accuracies for fast temporal modulations (>±7.8), whereas the speaker task did not show selectivity towards specific rates of temporal modulation. All correlations of differences were statistically validated and thresholded (cluster-size threshold yielding $\alpha_{clu} < 0.05$; see the 'Estimation of task differences between group MTFs' section in the Methods). Non-significant correlations are shown in blue. The plots have been interpolated for display purposes. Negative ω indicates upward temporal modulations and positive ω indicates downward temporal modulations. Before statistical assessment, all correlation values were normalized by applying the Fisher *z*-transformation.

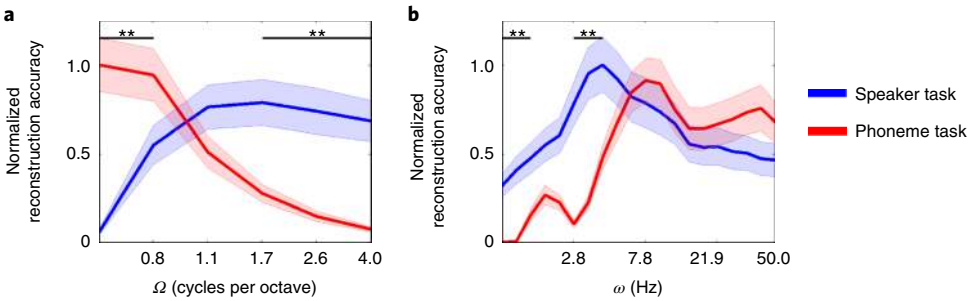

Fig. 6 | Dissociated spectral and temporal modulation profiles for the two tasks. a,**b**, One-dimensional modulation profiles for the speaker (blue) and phoneme (red) tasks. Each plot represents the group-average profile for each of the two tasks. Shaded areas represent the s.e. **a**, The spectral profiles showed dissociated modulation effects. Reconstruction accuracies for spectral modulations of 1.7, 2.6 and 4 cycles per octave were significantly higher during the speaker task than the phoneme task, whereas reconstruction accuracies for spectral modulations of 0.5 cycles per octave were higher during the phoneme task. **b**, The temporal profiles showed that reconstruction accuracies during the phoneme task were highest at faster rates of temporal modulation (non-significant), whereas reconstruction accuracies during the speaker task were higher for slower and intermediate rates of temporal modulation of 1.0 Hz, 1.2 Hz, 2.8 Hz and 3.4 Hz. Accuracies were normalized jointly for the two tasks for display purposes. $^{**}P < 0.001$ (Bonferroni-corrected), indicating modulations that different significantly between the two tasks. Temporal modulation profiles were averaged across upward and downward temporal modulations. Before statistical assessment, all correlation values were normalized by applying the Fisher *z*-transformation.

postSTG, we also found an interaction between hemisphere and task; classification accuracies were higher in the RH during the speaker task ($F_{1,12} = 5.050$, $P = 0.044$).

Discussion

Our results show that the performance of different tasks on identical speech stimuli modulates fine-grained task-relevant aspects of

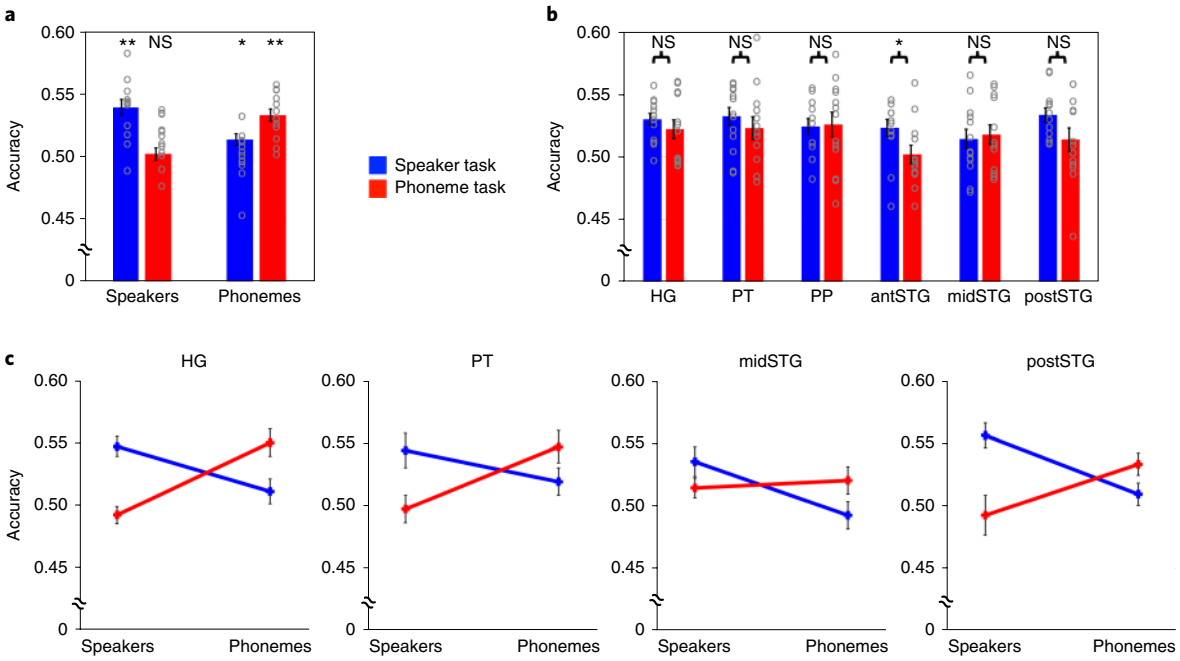

Fig. 7 | Target classification accuracies obtained during task performance. a–c, Task-relevant and task-irrelevant targets were classified using the reconstructed sounds derived from the fMRI data obtained during the speaker (blue) and the phoneme (red) tasks. **a**, Group-averaged classification accuracies. Classification according to speaker identity was most accurate for the reconstructed sounds obtained during the speaker task (mean accuracy (s.e.) = 0.54 (0.006)), and classification according to phoneme identity was most accurate for the reconstructed sounds obtained during the phoneme task (mean accuracy (s.e.) = 0.53 (0.005)). **b**, A main effect of task was not observed in five out of six ROIs. In the case of the antSTG, classification accuracies for both target classes were higher for the speaker task (mean difference (s.e.) = 0.021 (0.009)). **c**, HG, PT, midSTG and postSTG showed significant interactions between target class and task. Classification accuracies were higher for task-relevant targets than for task-irrelevant targets. $^{*}P < 0.05$; $^{**}P < 0.001$; NS, not significant. Circles represent individual data points and error bars represent s.e.m. Before statistical assessment, all correlation values were normalized by applying the Fisher *z*-transformation.

the neural encoding of those sounds. Furthermore, we found that this top-down modulation occurs even in HG, which is the earliest cortical stage of auditory processing[34]. The difference in encoding during the respective tasks was modulated in a manner that is consistent with the characteristic sound features of the targets that were attended to during these tasks. Acoustic analysis of the speech sounds showed that speaker information in the speech signal is characterized by modulation at higher spectral modulations, and—accordingly—the encoding of these sound features was amplified during the speaker task. By contrast, phonetic information in the stimuli was characterized by lower spectral modulations and faster temporal modulations, and this was paralleled by amplification in the neural encoding of corresponding sound features during the phoneme task.

Our study shows alterations in the bottom-up processing of speech sounds—which are probably driven by top-down influences—through selective enhancement or amplification of the encoding of spectrotemporal information that is relevant to the current behavioural goal. Our findings complement those of other studies with human participants that found contextual modulation of the neural representations of speech stimuli in early auditory areas[23,25,26], but we go further in terms of uncovering the acoustic specificity of these modulations with respect to behaviourally relevant goals. Our findings of similar effects of task modulation across different, lower- to higher-level auditory regions suggest the presence of multiple spatially distributed neural computations that work in parallel to facilitate flexible task-relevant behaviour. Moreover, our findings are in line with interactive speech models, which assume that the bottom-up processing of speech sounds is affected by interactions between multiple lexical and pre-lexical processes[3,6].

Our study provides a bridge between human sound processing and the mechanisms found in non-human species[12,13,35]. Multiple electrophysiology studies in animals have shown that neuronal-response profiles adapt to task requirements, and the matched-filter hypothesis[12,35,36] proposes a mechanism that could underlie this rapid task-driven flexibility. This hypothesis states that behaviourally induced changes in spectrotemporal sensitivities match the task-relevant spectrotemporal features within the sounds. The same principle has been proposed in the visual domain[37], and our results suggest that this mechanism also generalizes to the human auditory cortex during the processing of complex naturalistic sounds.

For both tasks, we found high reconstruction accuracies for faster temporal modulations, albeit this was more specific for the phoneme task. These results are in line with previous studies that showed the importance of temporal information for linguistic processing[38], and suggest that the human auditory cortex automatically processes faster temporal modulations in speech stimuli, regardless of whether linguistic or paralinguistic information is attended to. Furthermore, for the phoneme task, the reconstruction accuracies were highest at the lower spectral modulations, and they dropped significantly at higher scales. This parallels the acoustic variation that was found to distinguish the target phonemes (in our case, stop consonants) from one another. Together with the higher accuracies at faster temporal modulations, these findings suggest that there may be a trade-off between the relative importance of temporal and spectral detail during the perception of unvoiced stop consonants in that not only are faster temporal modulations enhanced during the processing of these sounds, but also in that higher spectral modulations are suppressed—or ignored—by the brain. Taken together, our results suggest that higher spectral density is more relevant

for voice processing only, in contrast to rapidly changing temporal information—which seems to be important for the processing of both speech phonetics and speaker identity—highlighting the importance of temporal information during the processing of different types of information in the speech signal.

Previous studies in humans that examined task effects in the context of speech processing using electrocorticography have found few modulation effects in HG[20–22]. A possible explanation for this discrepancy could be differences in the nature of the tasks. In the previous studies, participants mainly performed semantic categorization tasks on speech sounds, whereas in this study, participants had to retrieve detailed acoustic information embedded in the speech sounds, especially during the phoneme task. Therefore, the participants in the former studies probably relied on semantic processing, which takes place in more-upstream cortical areas such as the STS. By contrast, in our study there was heavier reliance on spectrotemporal parsing of the actual acoustic signal, which takes place in the earliest auditory areas[39–41]. Another explanation could be that task modulations of early auditory processing may be subtle[24], and thus sensitive analysis methods such as multivariate analyses—as used in this study—may be required to detect them.

The MTFs that we modelled with the data obtained during the speaker task were generally richer than the MTFs obtained with the data from the phoneme task. Together with the difference in task difficulty (Supplementary Fig. 1), it could be questioned whether the participants listened more attentively to the acoustic input during the speaker than during the phoneme task, which could have resulted in better encoding of the sounds during the former task. However, we argue that a generic attentional effect cannot explain our results. First, the difference in evoked fMRI responses between the two tasks was small (Supplementary Fig. 5) and we used an equal number of voxels across the tasks for the modelling of the encoding (see the 'Estimation of the linear decoders' section in the Methods). This ensured that the amount of input for the training of the linear decoders was the same for each task. Second, single-feature-based analyses showed encoding advantages for both the speaker and the phoneme tasks in different ROIs. Moreover, multi-feature analyses did not show generally higher classification accuracies for the speaker task compared to the phoneme task, with the exception of the antSTG. However, we did not observe task-related modulation in the neural encoding of the sounds in the antSTG. Finally, we found dissociations in the task modulation effects that were consistent with the amplification of task-relevant acoustic information in four out of six ROIs. If the modelling of sound encoding had been generally less successful due to decreased attention during the phoneme task, we should not have found any advantage of this task compared with the speaker task, nor relationships with the acoustic information that was critical for the respective tasks.

The difference in richness of the MTFs between the two tasks can, more probably, be explained by the fact that the sound representation model that we used may be less well tailored for the phonemes than for the speakers. The frequency shifts that we applied to create the percept of different speakers were clearly detectable in the modulation space, whereas the acoustic signatures of the different phonemes were less well pronounced. This difference was confirmed by significantly better classification of the speakers compared with the phonemes when we classified the different targets on the basis of their actual acoustic energy (mean difference in classification accuracies = 0.25; see Supplementary Results 2). The better fit of the model for the speaker task probably resulted in successful reconstruction of broader sound features during this task.

The computational modelling paradigm that we adopted in this study can be implemented in different ways, namely using single-voxel encoding and multivariate decoding (used here as described previously[30]). When using single-voxel encoding, a voxels response is modelled as a function of combined spectrotemporal information. However, single-voxel encoding relies on the responses of individual voxels and does not account for possible relationships between the responses of multiple voxels together. Previous studies have shown that task modulation effects are subtle and are distributed over multiple voxels[23,24]. To increase the sensitivity of our analysis, we applied multivariate decoding, which models the encoding of acoustic information across multi-voxel brain responses[30]. Neural encoding is established for each sound feature separately, which allows for the implementation of a rich acoustic model that accounts for the finer interdependent spectrotemporal acoustic detail that natural speech contains. As a consequence, this method does not account for correlations across sound features, whereas aspects of speech—especially phonemes—can be differentiated on the basis of a variety of combined spectral and temporal cues[42,43]. For this reason, we examined task-related differences in neural encoding not only with respect to single-sound features using model-based decoding, but also across multiple sound features within the sounds using a classification approach.

We found regional differences in the encoding of the sounds (Supplementary Figs. 12 and 13). For example, we found the highest reconstruction accuracies in the PT, which replicates its important role in the processing of spectrotemporally complex sounds[39,44]. Furthermore, we found a higher involvement of HG for the speaker task only, which can be explained by the spectral nature of this task and by the tonotopic representations that have been found in this region[45]. Furthermore, due to manipulation of the fundamental frequencies in our stimuli, the speaker task could have entailed aspects of pitch processing. Our finding of involvement of HG during the speaker task could—in part—reflect this, given that previous studies found pitch-sensitive regions within this area[46,47]. Furthermore, the few lateralized effects that we found are not consistent with previous observations regarding the lateralization of phonemic versus pitch processing[2,48], suggesting that such lateralization may depend on the use of specific tasks and stimuli.

Prediction accuracies of the MTFs in our study were generally lower compared with the ones found by Santoro et al.[30], who used a similar computational modelling approach to ours, or to those found by other studies that reconstructed speech from brain responses[11,49,50] (see Supplementary Figs. 14 and 15 and Supplementary Table 1 for information about the reliability of the reconstructed MTFs that we observed, and Supplementary Fig. 16 and Supplementary Table 2 for information about the variability in the estimations of the task effects). One explanation for the difference in prediction accuracies in this study compared with other studies could be our choice of stimuli, which were single pseudo-words with similar durations and syllabic structures. The use of these stimuli entailed the modelling of a restricted modulation space, compared with the modulation space that is sampled when using a wider variety of linguistic sounds or when using environmental sounds (linguistic and non-linguistic). A smaller modulation space makes successful feature reconstruction more challenging because decoders are trained on acoustic variation within the modulation space. Alternatively, lower prediction accuracies in our study could have been caused by the repeated presentations of speech stimuli, which might have attenuated BOLD responses that were evoked. This is another challenge for successful decoding, because the latter also depends on fluctuations in BOLD responses.

In conclusion, our results provide meaningful insights into the flexibility of auditory encoding in the human brain, non-invasively. Our data elucidate the neuro-computational mechanisms that enable the dynamic processing of task-relevant information within our rich and dynamic auditory environment—mechanisms which may generalize to other sensory modalities. The model-based decoding approach that we used enables a broad range of applications, for example, to address questions such as how expertise or pathology alter the encoding of sounds. However, open questions

remain regarding the neural mechanisms that underlie the encoding of higher-level, more-abstract aspects of speech. Models of encoding can therefore be refined by adding higher-level phonetic features, such as pitch, voicing, voice-onset-time and formant ratios[6,11,42]. Such extensions of the computational modelling of speech will enable a more-complete understanding of how speech perception comes about; however, the major challenge for future studies will be to establish the exact mechanisms that underlie the transformation of the acoustic signal into abstract representations.

Methods

Participants. Thirteen right-handed native French-speaking adults (6 women; mean age (s.d.) = 23 yr (4 yr)) participated in the study. The participants had self-reported normal hearing and reading abilities and none of them were musicians. The approval for the study was granted by the Cantonal Ethics Committee of the Vaud Canton (Switzerland). All participants gave written informed consent before the study, and they received monetary compensation for their participation. No statistical methods were used to predetermine the sample size, but our sample size is larger than those reported in previous publications[30,51,52]. Four other participants were originally included in this study but were later excluded because two did not complete the full experimental sessions and two displayed excessive head movement during the fMRI acquisition.

Task and stimuli. The speech stimuli included 120 pseudo-words that respected the rules of French phonology but had no meaning. The pseudo-words were created from a preselected French word list that was retrieved from the Lexique word database (http://www.lexique.org) using the Lexique Psycholinguistic Toolbox, which selectively scrambles letters within words while respecting French phonetic rules. The duration of the stimuli ranged from 1,000 ms to 1,200 ms, with a sampling frequency of 16 kHz. The stimuli ranged from three to five syllables in length, with a mean of four syllables. For each stimulus, sound onset and offset were ramped with a 10 ms linear slope, and the energy level (root mean square) was set to a constant value.

The participants performed a speaker identification (three target speakers) and a phoneme identification (three target stop consonants; /p/, /t/ and /k/) task. For example, the participants were presented with the item: /gabʁatadə/, and during the speaker-identification task, they identified the corresponding speaker (speaker 2 in this example), whereas during the phoneme-identification task, they indicated which target phoneme was in the stimulus (/t/ in this example). All targets (specific speaker or specific phoneme) were equally distributed across the stimuli (n = 40 per target), and task irrelevant targets were balanced across the task-relevant targets. For example, 13 of all the pseudo-words containing a /t/ target were spoken by speaker 1, 14 were spoken by speaker 2 and 13 were spoken by speaker 3. All of the pseudo-words contained only one out of the three target phonemes, and these targets occurred once or twice within each stimulus (with an equal probability of either). Furthermore, 25% of the words started with the target phoneme. The stimuli were spoken by a female professional phonetician and—to create the percept of different speakers—the fundamental frequency of the recorded pseudo-words was manipulated using an overlap-add technique based on waveform similarity (WSOLA) implemented in Audacity (www.audacityteam.org). The pitch-shifting algorithm that we used minimally alters the speed of the signal due to the combination of time stretching and resampling. For the creation of the stimuli corresponding to speaker 1, the fundamental frequency of a random subset (one third) of the stimuli was down-shifted by 7.5% with respect to the original value. For the creation of stimuli corresponding to speaker 2, the fundamental frequency of another third of the items was down-shifted by 0.01%, and for the creation of stimuli corresponding to speaker 3, the fundamental frequency of the remaining items was up-shifted by 10%. The shift in fundamental frequency for speaker 2 was not audible compared with the original stimuli. The naturalness of these pitch-shifted stimuli and the degree of success in creating the percept of different speakers were validated in a previous experiment, with different participants.

fMRI measurements. The fMRI experiment consisted of two sessions that were performed up to one week apart. Before the first session, the participants were familiarized with the two identification tasks to ensure that they understood and performed the task correctly. In both fMRI sessions, the participants performed the two identification tasks, which were administered in a counter-balanced order within and between sessions and across participants. The participants reported their responses through button presses using both hands. They were only required to respond on cued trials (13% of the trials), which were signalled by a visual cue presented after the stimulus (cued trials were excluded from the brain imaging data analyses). To motivate task engagement, the participants received an extra monetary bonus at the end of the two sessions if their task performance during scanning was at or above a desired performance level (75% correct); all participants achieved this. Sounds were presented binaurally at a comfortable listening level using the S14 model fMRI-compatible earphones by Sensimetrics (www.sens.com). The intensity of the sounds was adjusted to equalize their perceived loudness.

The stimuli were divided into four non-overlapping sets (n = 30 per set), each set containing a balanced subset of the different targets. Each of the four sets was presented within one fMRI session and repeated during the following session. In one session, the participants performed the speaker task on a particular stimulus set, and in the other session they performed the phoneme task on that set. Every stimulus set was presented within two consecutive functional runs, and the task changed after each stimulus set.

Every stimulus was presented three times, and the order of the stimuli was pseudo-randomized such that none of the targets relevant to the task were repeated during consecutive trials, and such that the irrelevant targets were not repeated more than two times in a row. One scan session consisted of a total of eight functional runs. One run lasted approximately 8 min and included 60 trials. The participants were cued to give a response during 13% of the trials, and during 12% of the trials no sound was presented (null trials) to increase inter-stimulus intervals. Data collection and analysis were not performed blind to the conditions of the experiment.

MRI parameters. Brain imaging was performed using a 7 T head-only scanner (Siemens Medical Solutions) with a 32-channel RF head array coil housed within a birdcage transmit coil (Nova Medical). A fast event-related scheme was used to collect the functional T2*-weighted images. Each volume consisted of 35 slices covering the superior temporal plane, and was acquired using a clustered echo planar imaging sequence (repetition time (TR) = 2,600 ms; time of acquisition (TA) = 1,250 ms, echo time (TE) = 20 ms, silent gap = 1,350 ms, voxel size = 1.5 × 1.5 × 1.5 mm^3, GRAPPA acceleration X3). Stimuli were presented within the silent gap between acquisitions, with a randomized inter-stimulus interval of two, three or four TRs. The minimal inter-stimulus interval was 5,100 ms.

Anatomical T1-weigted images were acquired using the MP2RAGE sequence, a modified magnetization-prepared rapid gradient-echo (MPRAGE) sequence that generates two image sets at different inversion times for bias field compensation[53], with the following parameters: resolution = 0.6 × 0.6 × 0.6 mm^3, TRmp2rage/TE/TI1/TI2 = 6,000 ms/2.05 ms/800 ms/2,700 ms. The MP2RAGE sequence included the sampling of fat excitation data (using a binomial pulse with a 7° flip angle), enabling retrospective motion correction of the T1-weighted images[54].

Data preprocessing. Functional and anatomical images were analysed using BrainVoyager and BrainVoyager QX (Brain Innovation). Preprocessing steps for the functional images consisted of slice scan-time correction (cubic spline interpolation), three-dimensional motion correction (trilinear/sinc) and temporal high-pass filtering to remove nonlinear drifts of maximum seven cycles per time course. Owing to head movement of two participants during acquisition of the anatomical scan, the anatomical images of these participants were motion corrected retrospectively by using the fat-selective excitation data as a three-dimensional motion navigator[54]. The functional images were coregistered to the anatomical images, and both were transformed into Talairach space. The volume time courses were moderately spatially smoothed (kernel width 2 mm). Cortical surface reconstructions were generated for all participants by segmenting the grey–white matter border in the anatomical images. Cortex-based alignment was performed on all participants using a moving-target-group-average approach on the basis of curvature information[55]. Alignment information was used for the random-effects general linear modelling of task effects, to obtain a group surface reconstruction and to compute the overlap of ROIs across participants.

fMRI analysis for univariate group contrasts. Univariate analyses were based on the functional time series that were resampled on the cortical surface reconstructions of the participants. Volume time courses were therefore mapped from the volume space onto the surface space using a customized MATLAB code (www.mathworks.com). Random-effects GLM analyses were performed on time-course data sampled on individual cortical reconstructions and aligned to the cortical group map using cortex-based alignment (BrainVoyager, Brain Innovations; Fig. 2). For the examination of task differences, we used one predictor per task (convolved with a double-gamma haemodynamic response function (HRF)). Note that the stimuli were identical for the two tasks, and that only task instructions differed. All GLMs included confounding predictors for each participant's motion-corrected parameters and for the excluded cued trials. Functional contrast maps (t-statistics) were calculated to assess task-specific sound-evoked responses (Supplementary Fig. 5). Functional contrasts were corrected for multiple comparisons ($P_{corr} < 0.05$) by applying a cluster-size threshold obtained using Monte Carlo simulations (initial threshold of $P = 0.001$ and 3,000 permutations), implemented in BrainVoyager.

Sound representation model for processing of speech sounds. Cortical representations of the stimuli were estimated using a biologically inspired model of auditory processing[7,30]. This is a two-stage model with an early processing stage that mimics the auditory transformations performed from the cochlea to the midbrain and a cortical processing stage that consists of more-complex spectrotemporal transformations that are presumed to take place in the auditory

cortex. Ultimately, the output of this auditory model provides a multidimensional representation of the initial sound waveform that encompasses the following acoustic dimensions: time, frequency, spectral and temporal modulations, and directionality.

In the early processing stage, the one-dimensional sound waveform was converted into a two-dimensional spectrotemporal representation describing the logarithmic (tonotopic) frequency content of a sound and how it evolved over time. This spectral analysis by the cochlea was simulated by a bank of 128 overlapping band-pass filters with a constant Q ($Q_{10dB}=3$) and with centre frequencies that were uniformly distributed along a logarithmic frequency (f) axis spanning 5.3 octaves ($f_{min}=180$ Hz and $f_{max}=7{,}040$ Hz). The output of these filters then entered a hair-cell stage that consists of a high-pass filter, a nonlinear compression and a low-pass filter. This was followed by a midbrain stage that was modelled by a first-order derivative with respect to the logarithmic frequency axis, a half-wave rectifier and a short time window integration (time constant = 8 ms).

The resulting auditory spectrogram then entered the cortical processing stage. Within this stage, the spectral and temporal modulation content of the spectrogram was estimated by a set of cortical filters. These two-dimensional filters were selective to different combinations of spectral and temporal modulations, and were centred at different frequencies along the tonotopic axis. Mathematically, the combined cortical filters performed a complex wavelet decomposition of the auditory spectrogram. The magnitude of this decomposition yielded a phase-invariant measure of the modulation content. We used two-dimensional filters that were tuned to six spectral modulation frequencies ($\Omega=0.5$, 0.7, 1.1, 1.7, 2.6 and 4 cycles per octave) and to 20 temporal modulation frequencies ($\omega=(\log_2(1):20:\log_2(50))$ Hz). The filters had a constant Q, and were directionally selective to either upward- or downward-drifting frequency sweeps.

We obtained the above auditory spectrogram and the corresponding modulation content of our stimuli using the NSL Tools package (available at http://www.isr.umd.edu/Labs/NSL/Software.htm) and using customized MATLAB codes. Afterwards, we reduced the multidimensional sound representations by first calculating the average across all time bins and by reducing the 128 tonotopic frequency bins from the auditory spectrogram into 60 frequency bins with constant bandwidth in octaves. Ultimately, the acoustic energy of each stimulus was represented by 14,400 sound features in total (6 spectral modulations × 40 temporal modulations (upward and downward) × 60 frequency bins). All of the processing steps described above were applied to all 120 stimuli, resulting in an ($N\times F$) feature matrix, where S represents the acoustic energy of all the sounds, N is the number of sounds (120) and F the number of features (14,400; see Supplementary Fig. 2a for a visualization of the transformation of the sounds).

Estimation of speaker and phoneme modulation profiles on the basis of the stimuli. The modulation profiles for the individual targets, as discussed in the 'Speaker and phoneme modulation profiles based on the stimuli' section, were obtained as follows. The feature matrix S, described above, was first normalized across all sounds (z-score), thus providing a standardized measure of the acoustic variation for a given sound feature. Then—for every sound feature—we calculated whether the sounds belonging to a specific target contained significant acoustic variation (t-test (d.f. = 39) with $P<0.05$, uncorrected). Target-specific modulation profiles were obtained by selecting the significant sound features, and by calculating the average of the acoustic variation of these features across the different sounds belonging to that target (these modulation profiles are shown in Fig. 1a). The profiles for each target class (speakers or phonemes) were obtained by calculating the average of the profiles across the different speakers or across the different phonemes. From these profiles, we created one-dimensional (frequency-unspecific) modulation profiles by calculating the average along the two irrelevant acoustic dimensions. Given that the full sound representation model included upward and downward temporal modulations, the one-dimensional temporal profiles also involved calculating the average across corresponding upward and downward modulations (for example, across −4 Hz and 4 Hz upward and downward modulation rates, respectively; the modulation profiles of each target class are shown in Fig. 1b).

Delineation of anatomical ROIs. We manually labelled the six following auditory ROIs (as described in a previous publication[30] and using an anatomical criteria described previously[56]): HG, PT, PP, antSTG, midSTG and postSTG.

HG corresponded to the first transverse temporal gyrus in the superior temporal plane. Its anterior-medial border was defined by the first transverse sulcus (FTS), and its posterior–lateral border was defined either by HS, or by the sulcus intermedius (SI) when one was present. Its medial border was confined by the circular sulcus of the insula (CSI).

The PT is a triangular area posterior to HG, along the superior temporal plane[56]. Its anterior-medial border was confined by HS or by the SI when one was present. Medially, its border was defined as the deepest point of the Sylvian fissure from the medial origin of HS until the posterior point of the STG at the temporoparietal junction. Laterally, its border was defined by the lateral rim of the superior temporal plane.

The PP is an area anterior to HG along the superior temporal plane, directly adjacent to the insula and the frontal operculum[56]. Its medial border was confined by the CSI. Its lateral border was defined by following the FTS until the anterior tip of HS or SI. From here, the lateral border became the lateral rim of the superior temporal plane.

The STG corresponds to the lower bank of the Sylvian fissure, and was defined anteriorly from the temporal pole extending posteriorly to the end of the Sylvian fissure at the temporoparietal junction. The STG was divided into an anterior, middle and posterior part. The borders of midSTG were defined by using the anterior-lateral and posterior-medial ends of HG as references points.

For all of the participants, all anatomical ROIs were labelled on the cortical surface reconstructions of each hemisphere obtained with BrainVoyager. The labelled ROIs were projected into the volume space of each participant to obtain three-dimensional masks. All of the masks were corrected for mislabelled voxels and for voxels that overlapped between the different ROIs by visual inspection and using customized MATLAB codes (the resulting ROIs are visualized in Fig. 3).

Estimation of fMRI responses to individual speech sounds. We computed responses of voxels to individual speech sounds using a two-phase procedure implemented with customized MATLAB codes. For each voxel i, the response vector $\mathbf{Y}_i$ (which consists of $[N\times 1]$, where N is the number of sounds) was obtained in two steps (as described previously[30]). First, a HRF common to all stimuli was estimated using a GLM analysis in which all speech sounds were treated as a single condition. Then, using this HRF and one predictor per sound, we computed the beta-weight for each speech sound. We implemented a fourfold cross-validation across the four sets of stimuli that are described in the 'fMRI measurements' section. The HRF was estimated using the training data, and beta-weights were computed separately for the training and test data (90 training sounds and 30 test sounds per cross-validation).

Further analyses were performed on voxels that had a significant positive response (beta-weights) to the training sounds within the anatomically defined ROIs ($P<0.05$, uncorrected). The responses of the voxels to the speech sounds were modelled twice—once on the basis of the data obtained during the speaker task and once on the basis of the data obtained during the phoneme task. This resulted in separate multi-voxel response patterns for each task.

Estimation of the linear decoders. For each task, a linear decoder was trained to estimate the relationship between the multi-voxel fMRI responses within a specific ROI and the cortical representations of all the sounds. For each participant, we trained linear decoders to predict the acoustic energy for every sound feature of the feature matrix S (described above). Separate linear decoders were trained for each feature as follows (as described previously[30]; see Supplementary Fig. 2b for a visualization). The sound feature $S_i[N_{train}\times 1]$ was modelled as a linear transformation of the multi-voxel response pattern $\mathbf{Y}_{train}$ $[N_{train}\times V]$, plus a bias term (b_i) and a noise term (no), according to the following equation:

$$S_i=\mathbf{Y}_{train}\mathbf{C}_i+b_i\mathbf{1}+no \quad (1)$$

where N_{train} is the number of sounds in the training set, V is the number of voxels per ROI, 1 is an $[N_{train}\times 1]$ vector of ones, and C_i is a $[V\times 1]$ vector of weights for which the elements c_{ij} quantify the contribution of voxel j to the encoding of feature i. Approximately 500 to 1,500 voxels were used for each linear decoder (see Supplementary Fig. 6 for the number of voxels per ROI). For each participant, the number of voxels used for decoding was equalized across the two tasks for corresponding ROIs.

The solution to equation (1) was computed by means of kernel ridge regression using a linear kernel[57]. We wanted to optimize the prediction accuracy per sound feature rather than to establish comparable linear transformations between the different tasks, therefore, the regularization parameter λ was determined independently for each feature and for each task by generalized cross validation[58]. The search grid included 32 values between $10^{0.5}$ and 10^{11}, logarithmically spaced with a grid grain of $10^{0.33}$ (see Supplementary Fig. 7b for the distribution of the selected λ values for each ROI). Training and testing of the decoders was performed using fourfold cross-validation across the four stimulus sets (90 training sounds and 30 test sounds per cross-validation).

Estimation of ROI-specific MTFs for each participant. We modelled sound encoding during performance of the tasks by estimating ROI-specific MTFs separately for the two tasks. For each participant, an MTF was modelled as follows. The linear decoders (described above) were used to predict the acoustic energy for the sound features of unseen test sounds in a fourfold cross-validation scheme (30 sounds per cross-validation). We combined the predictions of the fourfold cross-validations by concatenating the predictions of the sound features for all of the test sounds (120 sounds in total). This resulted in a feature matrix S^{pred} for each task, representing the predicted acoustic energy of all the sounds, where N is the number of sounds (120) and F the number of features (14,400). The prediction accuracy for each sound feature was then assessed by computing r between the predicted acoustic energy and the actual acoustic energy of the sounds for that sound feature, as represented in feature matrix S. This resulted in 14,400 correlation coefficients (equal to the total number of sound features), which together represent the MTF of a certain region (see Supplementary Fig. 2c,d

for a visualization of these steps). The MTFs were computed by calculating the correlations between each sound feature in the actual feature matrix S and each sound feature in the predicted feature matrix S^{pred} for each task separately. MTFs were computed for every participant, and in the next step we computed group-averaged MTFs.

Estimation of ROI-specific MTFs at the group level. We created group-averaged MTFs by combining the MTFs that were computed for each participant, and nonparametric random-effects group analysis was used to assess statistical significance for each group MTF ($MTF^i >$ chance) as follows. First, we computed the null distribution for every correlation coefficient in the MTF of each participant. These distributions were obtained by randomly permuting (5,000 times) the stimulus labels of each predicted sound feature (S_i^{pred}) and—for each permutation—we recomputed the correlation between the permuted sound feature (S_i^{pred}) and the actual sound feature (S_i). The empirical chance score of a feature's correlation coefficient (r_{chance}) was defined as the mean correlation value of this null distribution.

We then assessed whether the observed correlation values were significantly above chance at the group level using one-tailed exact permutation testing. These tests were one-sided because negative correlations indicated unplausible predictions of acoustic energy. Before statistical assessment, all correlation values were normalized by applying the Fisher z-transformation. The test statistic was the group average of the difference between the observed correlation and the obtained chance score ($d = r - r_{chance}$) of each participant. The null distribution for the group-average difference was obtained by changing the sign of d for a randomly selected subset of participants, and then recalculating the group-average difference score[59]. This procedure was repeated for all possible permutations of the sign change ($2^{13} = 8{,}192$). The P value of the test statistic was computed as the proportion of the null distribution that yielded a group difference equal to or more extreme than the observed one.

We used a cluster-size threshold procedure to correct for multiple comparisons across the different features within an MTF[60]. For this, we calculated the cluster size of the false-positive rate (α) for every permutation in the null distribution of the group-average difference using an initial uncorrected threshold of $\alpha_{in} = 0.05$. The minimum cluster size that yielded a cluster level of $\alpha_{clu} = 0.05$ was then used as a threshold for the actual correlation values within the MTF.

The whole statistical procedure described above was repeated for every ROI independently, resulting in 12 group MTFs for each task (6 for each ROI and 2 for each hemisphere). For each task, we statistically tested for hemispheric differences within each ROI (Supplementary Results 1), and the MTFs were averaged across hemispheres when no such differences were found. The resulting group MTFs are discussed in the 'Neural encoding of speech sounds in auditory ROIs' section and are visualized in Fig. 4. Group analyses were performed on all participants ($n = 13$), except for the antSTG ($n = 12$). For this ROI, we had to exclude one participant because—for this person—the sounds did not evoke significant fMRI responses during performance of the phoneme task.

Estimation of task differences between group MTFs. To test for task differences in the MTFs of corresponding ROIs, we computed pairwise comparisons using the following conjunction analyses (that is, combined contrasts): $(MTF^{speaker} > MTF^{phoneme}) \cap (MTF^{speaker} >$ chance) and $(MTF^{phoneme} > MTF^{speaker}) \cap (MTF^{phoneme} >$ chance). Note that for the conjunction test to be significant, both contrasts need to be significant.

The procedure for obtaining the P values of the contrasts ($MTF^{speaker} >$ chance) and ($MTF^{phoneme} >$ chance) is described in the above section. Similarly, the P values for the contrasts ($MTF^{speaker} > MTF^{phoneme}$) and ($MTF^{phoneme} > MTF^{speaker}$) were obtained with the same exact permutation test. The test statistic used for these contrasts was the group average of the individual task differences ($d = r_{speaker} - r_{phoneme}$ for the $MTF^{speaker} > MTF^{phoneme}$ contrast and $d = r_{phoneme} - r_{speaker}$ for the $MTF^{phoneme} > MTF^{speaker}$ contrast). These MTFs were also corrected for multiple comparisons across the whole MTF using the cluster-size threshold procedure described above. We used the MTFs that were averaged across hemispheres only when no hemispheric differences were found for both tasks (Supplementary Results 1). This was the case for HG, PP, antSTG and postSTG. In the case of PT and midSTG, we computed the contrasts separately for the RH and LH. The task differences that we found between the MTFs are discussed in the 'Analysis of task differences using single sound features' section, and are visualized in Fig. 5.

Comparison of task-specific spectral and temporal modulation profiles. We further explored the differences in the MTFs of the speaker and phoneme tasks across multiple ROIs by creating task-specific spectral and temporal modulation profiles, as follows. For every participant, the modulation profiles for the speaker task were obtained by calculating the average of the reconstruction accuracies of the significant sound features across the MTFs that showed task effects for the speaker task (that is, the significant features within HG, PT (RH), PP, midSTG (LH) and postSTG). Similarly, for each participant, the modulation profiles for the phoneme task were obtained by calculating the average of the reconstruction accuracies of the significant sound features across the MTFs that showed task effects for the phoneme task (that is, the significant features within midSTG (RH) and postSTG). We then created one-dimensional modulation profiles by calculating the average across the two irrelevant acoustic dimensions (for the temporal profiles we also averaged across upward and downward temporal modulations). We tested for task differences between these one-dimensional modulation profiles at each spectral and temporal modulation using t-tests (d.f. = 5 for spectral scales and d.f. = 19 for temporal rates with $P < 0.05$, Bonferroni-corrected for the number of spectral or temporal modulations). The results of these comparisons are visualized in Fig. 6.

We further tested for the relationship between the neural and the stimulus modulation profiles. For this, we computed r, across participants, between the spectral and temporal profiles of the speaker or phoneme task obtained from the neural data with the spectral and temporal profiles of the speaker or phoneme targets class obtained from the stimuli (Fig. 1). The differences in the correlation coefficients were assessed using t-tests (d.f. = 12 with $P < 0.05$). Before statistical assessment, all correlation values were normalized by applying the Fisher z-transformation. These results are shown in the 'Analysis of task differences using single sound features' section.

Analysis of target separability with the use of linear classification. We tested for task-related modulations in target separability using the sounds that were reconstructed from the fMRI responses. For this, we applied a SVM algorithm (implemented in MATLAB) on the predicted feature matrix S^{pred} of each task. For each participant, classification was performed on S^{pred} of each ROI and each hemisphere. For every S^{pred}, classifiers were trained to discriminate between the sounds belonging to the different speakers, and classifiers were trained to discriminate between the sounds containing the different target phonemes. The three-class problem was transformed into binary classification using a pairwise scheme (speaker 1 versus speaker 2, speaker 1 versus speaker 3 and speaker 2 versus speaker 3 for the different speakers, and /p/ versus /t/, /p/ versus /k/ and /t/ versus /k/ for the different phonemes). For each pairwise classification, all of the sounds belonging to a specific target were used (40 sounds per target); this resulted in a classification matrix of 80 sounds by 14,400 predicted sound features. Only sound features that were positively predicted were included; sound features with negatively predicted acoustic energy were set to 0. Training and testing of the SVM was performed using a leave-one-out scheme (79 training sounds and 1 test sound per iteration). Classification accuracy was assessed by computing the number of correct classifications of the test sounds divided by the total number of test sounds (80 in total). For each participant, overall classification accuracies for all speakers and for all phonemes were obtained by calculating the average of the classification performance of the three pairwise classifications.

The empirical null distributions of classification performances were obtained by randomly permuting (200 times) the target labels and repeating the training and testing procedure. For each classifier, the empirical chance score of classification performance was defined as the mean classification accuracy of the null distribution. The P values of the mean prediction accuracies were assessed using t-tests (d.f. = 12, $P < 0.05$, Bonferroni-corrected; Fig. 7a). To examine whether target separability changed as a function of task performance, we performed a repeated-measures ANOVA on the mean classification accuracies of all of the participants. The results of these analyses are discussed in the 'Analysis of target separability using multiple sound features' section, and are visualized in Fig. 7.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The stimuli and the sound representations of the stimuli (feature matrix S) and the estimated fMRI responses (beta-weights) from a subset of the participants from this study are available as Supplementary Audio Files, Supplementary Data 1 and 2.

Code availability

The code that support the findings of this study is available from the corresponding author upon reasonable request.

Received: 21 August 2018; Accepted: 3 June 2019;
Published online: 8 July 2019

References

1. Belin, P., Fecteau, S. & Bedard, C. Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* **8**, 129–135 (2004).
2. Leonard, M. K. & Chang, E. F. Dynamic speech representations in the human temporal lobe. *Trends Cogn. Sci.* **18**, 472–479 (2014).
3. Davis, M. H. & Johnsrude, I. S. Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hear. Res.* **229**, 132–147 (2007).
4. Leonard, M. K., Baud, M. O., Sjerps, M. J. & Chang, E. F. Perceptual restoration of masked speech in human cortex. *Nat. Commun.* **7**, 13619 (2016).

5. Gaskell, M. G. & Marslen-Wilson, W. D. Integrating form and meaning: a distributed model of speech perception. *Lang. Cogn. Process.* **12**, 613–656 (1997).
6. McClelland, J. L., Mirman, D. & Holt, L. L. Are there interactive processes in speech perception? *Trends Cogn. Sci.* **10**, 363–369 (2006).
7. Chi, T., Ru, P. & Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **118**, 887–906 (2005).
8. Santoro, R. et al. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.* **10**, e1003412 (2014).
9. Schonwiesner, M. & Zatorre, R. J. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proc. Natl Acad. Sci. USA* **106**, 14611–14616 (2009).
10. Theunissen, F. E., Sen, K. & Doupe, A. J. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci.* **20**, 2315–2331 (2000).
11. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
12. Atiani, S., Elhilali, M., David, S. V., Fritz, J. B. & Shamma, S. A. Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron* **61**, 467–480 (2009).
13. David, S. V., Fritz, J. B. & Shamma, S. A. Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc. Natl Acad. Sci. USA* **109**, 2144–2149 (2012).
14. Fritz, J., Elhilali, M. & Shamma, S. A. Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. *J. Neurosci.* **25**, 7623–7635 (2005).
15. Fritz, J., Shamma, S., Elhilali, M. & Klein, D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* **6**, 1216–1223 (2003).
16. Golestani, N., Hervais-Adelman, A., Obleser, J. & Scott, S. K. Semantic versus perceptual interactions in neural processing of speech-in-noise. *Neuroimage* **79**, 52–61 (2013).
17. von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J. & Griffiths, T. D. How the human brain recognizes speech in the context of changing speakers. *J. Neurosci.* **30**, 629–638 (2010).
18. Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).
19. Holdgraf, C. R. et al. Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nat. Commun.* **7**, 13654 (2016).
20. Nourski, K. V., Steinschneider, M., Oya, H., Kawasaki, H. & Howard, M. A.III. Modulation of response patterns in human auditory cortex during a target detection task: an intracranial electrophysiology study. *Int. J. Psychophysiol.* **95**, 191–201 (2015).
21. Nourski, K. V., Steinschneider, M., Rhone, A. E. & Howard, M. A.III. Intracranial electrophysiology of auditory selective attention associated with speech classification tasks. *Front. Hum. Neurosci.* **10**, 691 (2016).
22. Steinschneider, M. et al. Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. *Front. Neurosci.* **8**, 240 (2014).
23. Bonte, M., Hausfeld, L., Scharke, W., Valente, G. & Formisano, E. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *J. Neurosci.* **34**, 4548–4557 (2014).
24. Formisano, E., De Martino, F., Bonte, M. & Goebel, R. "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* **322**, 970–973 (2008).
25. Kilian-Hutten, N., Valente, G., Vroomen, J. & Formisano, E. Auditory cortex encodes the perceptual interpretation of ambiguous sound. *J. Neurosci.* **31**, 1715–1720 (2011).
26. Ley, A. et al. Learning of new sound categories shapes neural response patterns in human auditory cortex. *J. Neurosci.* **32**, 13273–13280 (2012).
27. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
28. Miyawaki, Y. et al. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* **60**, 915–929 (2008).
29. Moerel, M., De Martino, F. & Formisano, E. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *J. Neurosci.* **32**, 14205–14216 (2012).
30. Santoro, R. et al. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl Acad. Sci. USA* **10**, e1003412 (2017).
31. Baumann, O. & Belin, P. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychol. Res.* **74**, 110–120 (2010).
32. Mesgarani, N., David, S. V., Fritz, J. B. & Shamma, S. A. Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.* **123**, 899–909 (2008).
33. Chi, T., Gao, Y., Guyton, M. C., Ru, P. & Shamma, S. Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* **106**, 2719–2732 (1999).
34. Saenz, M. & Langers, D. R. Tonotopic mapping of human auditory cortex. *Hear. Res.* **307**, 42–52 (2014).
35. Fritz, J., Elhilali, M. & Shamma, S. A. Adaptive changes in cortical receptive fields induced by attention to complex sounds. *J. Neurophysiol.* **98**, 2337–2346 (2007).
36. Yin, P., Fritz, J. B. & Shamma, S. A. Rapid spectrotemporal plasticity in primary auditory cortex during behavior. *J. Neurosci.* **34**, 4396–4408 (2014).
37. Anton-Erxleben, K., Stephan, V. M. & Treue, S. Attention reshapes center-surround receptive field structure in macaque cortical area MT. *Cereb. Cortex* **19**, 2466–2478 (2009).
38. Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J. & Ekelid, M. Speech recognition with primarily temporal cues. *Science* **270**, 303–304 (1995).
39. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
40. Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T. & Medler, D. A. Neural substrates of phonemic perception. *Cereb. Cortex* **15**, 1621–1631 (2005).
41. Ahissar, M., Nahum, M., Nelken, I. & Hochstein, S. Reverse hierarchies and sensory learning. *Phil. Trans. R. Soc. Lond. B* **364**, 285–299 (2009).
42. Giraud, A. L. & Poeppel, D. in *The Human Auditory Cortex*, chapter 9 225–260 (eds Poeppel, D. et al.) (Springer-Verlag, 2012).
43. Moore, B. C. J. *An Introduction to the Psychology of Hearing* 4th edn (Academic, 1997).
44. Griffiths, T. D. & Warren, J. D. The planum temporale as a computational hub. *Trends Neurosci.* **25**, 348–353 (2002).
45. Formisano, E. et al. Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron* **40**, 859–869 (2003).
46. De Angelis, V. et al. Cortical processing of pitch: model-based encoding and decoding of auditory fMRI responses to real-life sounds. *Neuroimage* **180**, 291–300 (2017).
47. Griffiths, T. D. & Hall, D. A. Mapping pitch representation in neural ensembles with fMRI. *J. Neurosci.* **32**, 13343–13347 (2012).
48. Zatorre, R. J., Evans, A. C., Meyer, E. & Gjedde, A. Lateralization of phonetic and pitch discrimination in speech processing. *Science* **256**, 846–849 (1992).
49. Bitterman, Y., Mukamel, R., Malach, R., Fried, I. & Nelken, I. Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature* **451**, 197–201 (2008).
50. Pasley, B. N. et al. Reconstructing speech from human auditory cortex. *PLoS Biol.* **10**, e1001251 (2012).
51. Da Costa, S., van der Zwaag, W., Miller, L. M., Clarke, S. & Saenz, M. Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex. *J. Neurosci.* **33**, 1858–1863 (2013).
52. De Martino, F. et al. Frequency preference and attention effects across cortical depths in the human primary auditory cortex. *Proc. Natl Acad. Sci. USA* **112**, 16036–16041 (2015).
53. Marques, J. P. et al. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *Neuroimage* **49**, 1271–1281 (2010).
54. Gallichan, D., Marques, J. P. & Gruetter, R. Retrospective correction of involuntary microscopic head movement using highly accelerated fat image navigators (3D FatNavs) at 7T. *Magn. Reson. Med.* **75**, 1030–1039 (2016).
55. Goebel, R., Esposito, F. & Formisano, E. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum. Brain Mapp.* **27**, 392–401 (2006).
56. Kim, J. J. et al. An MRI-based parcellation method for the temporal lobe. *Neuroimage* **11**, 271–288 (2000).
57. Bishop, C. *Pattern Recognition and Machine Learning* (Springer, 2006).
58. Golub, G., Heath, M. & Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223 (1979).
59. Menke, J. & Martinez, T. Using permutations instead of Student's t distribution for p-values in paired-difference algorithm comparisons. In *Proc. IEEE International Joint Conference on Neural Networks* 2, 1331–1335 (2004).
60. Forman, S. D. et al. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.* **33**, 636–647 (1995).

Acknowledgements

We thank the staff at the Center for Biomedical Imaging EPFL, Vaud, Switzerland for access to the imaging platform, and W. van der Zwaag for facilitating data collection; J. Gonzalez for helping with auditory recording; F. Zay for reading the stimuli; C. Türk for assisting during data collection and L. Ermacora for the phonetic segmentation of the stimuli; F. De Martino for providing code for analysing the data; V. de Angelis and N. Disbergen for helping with data analysis; G. Valente for helping with the statistical analysis and D. Gallichan for motion correction of the anatomical images. This work was supported by the Swiss National Science Foundation (grant numbers PP00P3_133701, PP00P3_163756 and 100014_182381 awarded to N.G.) and the University of Geneva Language and Communication Research Network. E.F. was supported by

The Netherlands Organisation for Scientific Research (VICI grant number 453-12-002) and the Dutch Province of Limburg. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All authors contributed to the conception and design of the experiment. N.G. and E.F. supervised the study. S.R. created the behavioural task and stimuli, programmed the fMRI experiment, collected, analysed (including writing code) and interpreted the data, and wrote the manuscript. R.S. helped to program the fMRI experiment and to analyse the data (including writing code for it). A.H.-A. helped to create the stimuli and to implement the behavioural task. E.F. supervised the data analysis (including writing code for and implementing it), guided data interpretation and helped write the manuscript. N.G. helped to create the stimuli, to guide the data analysis and interpretation and to write the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0648-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.R.

Peer review information: Primary Handling Editor: Mary Elizabeth Sutherland.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Sanne Rutten

Last updated by author(s): May 31, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	PsychoPy
Data analysis	BrainVoyager 20.2, BrainVoyager QX 2.8 and Matlab

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The stimuli and the sound representations of the stimuli (feature matrix S) and the estimated fMRI responses (beta weights) from a subset of the participants from this study have been made available as Supplementary Files attached to this article.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is a within-subject design were we repeatedly measured brain responses to evaluated the quantitative difference in fMRI responses.
Research sample	The study included right-handed native French adults (6 females and 7 males, mean age 23 years, sd = 4 years), mostly University undergraduate students, but not exclusively . The participants were selected based on self-reported normal hearing and reading abilities, and given the nature of the task manipulations that we used, the participants should had normal reading abilities and no extensive musical experience.
Sampling strategy	Participants were recruited through advertisements, mostly within the University environment. We used the sample size of another study that used an similar modeling approach as a reference.
Data collection	The fMRI experiment consisted of two sessions, which were performed up to one week apart. During data collection, the participants were presented with identical speech stimuli twice, once while performing a speaker identification task on these sounds, and once while performing a phoneme identification task. During the speaker identification task, the participants indicated whether the presented sound was spoken by speaker 1, speaker 2 or speaker 3. During the phoneme identification task the participants indicated whether the presented sound contained either a /P/, /T/ or /K/ sound. Prior to the first session and outside the scanner, the participants were familiarized with the these two tasks. Behavioral performances were recorded with the use of a computer. There was nobody else present during the experiment besides the participant and the research team. The researchers were not blind to the experimental condition and the study hypotheses during data collection.
Timing	Data was collected in three different sampling cohorts, with the following acquisition dates: cohort I > October 2014 - December 2014 cohort II > June 2015 cohort III > December 2016 - January 2017
Data exclusions	Data was excluded from 2 participants, due to extensive head movement during data acquisition. This exclusion criteria was pre-established.
Non-participation	2 participants dropped out during the experiment, because of discomfort within the scanner.
Randomization	We used a within-subject design, therefore the participants were not allocated into specific experimental groups; instead they participated in all experimental conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology
☐	☐ Animals and other organisms
☐	☒ Human research participants
☐	☐ Clinical data

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited through advertisement, which were advertised mostly, but not exclusively, within the University environment. Consequently our sample mainly consists out of undergraduate students. This possibly could have had some influences on general task performances, however given that it is a within-subject design we do not think this has had an impact on the interpretation of the results.
Ethics oversight	The approval for the study was granted by the Cantonal Ethics Committee of the Vaud Canton (Switzerland).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*

Outcomes — *Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.
For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission — *Provide a list of all files available in the database submission.*

Genome browser session
(e.g. UCSC)
Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates — *Describe the experimental replicates, specifying number, type and replicate agreement.*

Sequencing depth — *Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*

Antibodies — *Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Peak calling parameters — *Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*

Data quality — *Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*

Software — *Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation — *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*

Instrument — *Identify the instrument used for data collection, specifying make and model number.*

Software — *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance — *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

Gating strategy — *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	Fast event-related
Design specifications	Every subject underwent 2 sessions. Per subject there were 16 blocks (8 per session), 960 trials in total (480 per session, 60 per block). One block lasted about 8 minutes, one trial was 1 TR (2.6 sec). Sounds were presented within silent gaps between acquisitions, with a randomized inter-stimulus-interval of 2, 3 or 4 TRs (5.2 sec, 7.8 sec or 10.4 sec).
Behavioral performance measures	Performance was measured using button presses on cued trials (128 trials). We used mean percentage correct responses to assess task performance.

Acquisition

Imaging type(s)	Functional and structural
Field strength	7 Tesla
Sequence & imaging parameters	Functional T2*-weighted images were collected using a GRAPPA acceleration X3 EPI sequence. FOV= 222 x 222 mm, matrix size = 148 x 148, slice thickness: 1.5; orientation; anterior-posterior, TE= 20 ms, TR = 2600 ms, TA= 1250 ms, flip angle = 90 degree. Anatomical T1-weighted images were acquired using the MP2RAGE sequence, derived from a modified magnetization-prepared rapid gradient-echo (MPRAGE) sequence that generates two image sets at different inversion times for bias field compensation, with the following parameters: resolution 0.6 x 0.6 x 0.6 mm, TR MP2RAGE= 6000 ms, TE= 2.05 ms, TI1 = 800 ms and TI2 = 2700 ms.
Area of acquisition	For the functional measurements, each volume consisted of 35 slices covering the superior temporal plane. The region was determined by a fixed slice positioning from the center of the brain, which was obtained with a localizer sequence.

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	BrainVoyager 20.2 and BrainVoyager QX 2.8 preprocessing steps involved: - slice scan-time correction (cubic spline interpolation) - 3D-motion correction (trillinear/sinc interpolation)) - temporal high-pass filtering to remove non-linear drifts of maximum seven cycles per time course. - Volumes were spatially smoothed (kernel width 2 mm)
Normalization	Images were normalized (non-linear) by first aligning the functional images to the anatomical images using gradient-based affine transformation (9 parameters: translation, rotation, scale). Anatomical images were normalized into Talairach space through sinc interpolation.
Normalization template	We used Talairach normalization
Noise and artifact removal	We did not use noise or artifact removal
Volume censoring	We did not use volume censoring

Statistical modeling & inference

Model type and settings	The main analysis is based on a recently developed model-based multivariate decoding approach (see Santoro et al., 2017). fMRI-responses to individual stimuli (sounds) were estimated in a two-phase procedure. For the two tasks separately, for each voxel, we first estimated the hemodynamic response function (HRF) common to all sounds, by fitting a fixed-effect GLM with all the sounds in one condition. Then, the beta weights to individual sounds were estimated by fitting a fixed-effect GLM having one predictor per sound. We implemented a 4-fold cross validation. For the HRF estimation only training data (three stimulus sets, N = 90 sounds) were used and beta weights were computed for training and testing sounds (one stimulus set, N = 30 sounds) separately. Further analyses (described below) only included voxels with significant positive responses (P < 0.05 uncorrected) to the training sounds within specific ROIs.
Effect(s) tested	We tested for differences in the encoding of speech sounds as a consequence of performing a speaker task or a phoneme task. The dependent variable that was used was the reconstruction accuracy of sound features from the brain data. Sound features were defined based on a computational model of auditory processing. In this model, each feature represented the acoustic energy in the stimuli at a specific combination of frequency, spectral modulation and temporal modulation. For the speaker task and the phoneme task, we trained a multivoxel decoder (described two points below) to reconstruct the acoustic energy for individual sound features. The reconstruction accuracy of the sound feature was

assessed by computing the Pearson's correlation coefficient between the actual acoustic energy in the test stimuli and the energy as predicted by the decoder.

We tested for task effects using a univariate approach and a multivariate approach. For the univariate contrasts we tested the following conjunctions: (Speaker Task > Phoneme Task) ∩ (Speaker Task > chance) and (Phoneme Task > Speaker Task) ∩ Phoneme Task > chance). Note that "Speaker Task" or "Phoneme Task" here indicates the reconstruction accuracies during the respective tasks. We tested for statistical significance of these differences by performing random-effects one-tailed non-parametric testing (described below).

For the multivariate approach we applied linear classification (SVM) on the sound features that were reconstructed from the fMRI responses during the speaker task and the sound features that were reconstructed from the fMRI responses during the phoneme task. For each task we trained different classifiers to differentiate the reconstructed sounds on speaker identity (speaker 1, speaker 2 or speaker 3) and on phoneme identity (/P/, /T/ or /K/). We tested for task effects by comparing classification accuracies. We specifically tested whether classification performance on speaker identity based on the reconstructed sounds of the speaker task differed from classification performance on speaker identity based on the reconstructed sounds of the phoneme task. And similarly for phoneme identity.

Specify type of analysis: ☐ Whole brain ☒ ROI-based ☐ Both

Anatomical location(s): We defined the following anatomical regions: Heschl's gyrus, planum temporale, planum polare, superior temporal gyrus. The latter region was additionally divided into a anterior, middle or posterior part.

Statistic type for inference
(See Eklund et al. 2016)

We used 'sound feature'-wise statistical inference by assessing the significance of the prediction accuracy per sound feature (Pearson's correlation coefficient). We used non-parametric random-effects group analyses for assessing statistical significance per sound feature (Pearson's correlation of sound feature i > chance), per subject. The null-distribution was obtained with 5000 permutations, and the empirical chance level of a feature's correlation was defined as the mean of this null-distribution. Per permutation, the Pearson's correlation was recomputed between randomly permuted acoustical labels of the reconstructed sound features and the actual sound features.

We then assessed whether the Pearson's correlation was significantly above chance at the group level. For this we used random-effects one-tailed non-parametric tests (exact permutation). The test statistic was the group average of each subjects difference between the observed correlation and the empirical chance level. The null-distribution of the group-average difference was obtained by changing the sign of the difference score for a random subset of subjects, and then recalculating the group averaged difference. This was repeated for all possible permutations of the sign change (2^13=8192). The P-value of the test statistic was computed as the proportion of the null distribution that yielded a group difference equal to or more extreme than the observed one. We repeated this procedure for every sound feature, per ROI and separately for the two tasks.

For the multivariate approach we used repeated-measures ANOVAs to test for task difference in classification accuracies between the two tasks.

Correction

We used cluster-size threshold procedure to correct for multiple comparisons across all the sound features within the reduced sound-representation model. We defined the cluster size of the false-positive rate (alpha) for every permutation in the null-distribution (described above) with an initial uncorrected threshold of alpha = 0.05. The minimum cluster size that yielded a cluster-level of alpha = 0.05 was used as a threshold for the actual correlation values found for all sound features within the sound-representation model.

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Independent variables were voxel's fMRI-responses to individual sounds.

Voxels within the anatomically defined ROI which had significant positive responses (P < 0.05 uncorrected) to the training sounds were selected. We did not use dimension reduction.

The stimulus feature Si [Ntrain x 1] resulting from the computational model (see effects tested) was expressed as a linear function of the multi-voxel response pattern Ytrain [Ntrain x V], plus a bias term (bi) and a noise term (n), according to the following equation:

Si = YtrainCi + bi1 + n,

where Ntrain is the number of sounds in the training set, V is the number of voxels per ROI, 1 is an [Ntrain x 1] vector of ones, and Ci is a [V x 1] vector of weights whose elements cij quantify the contribution of voxel j to the encoding of feature i.

The solution to equation 1 was computed by means of kernel ridge regression using a linear kernel. The regularization parameter λ was determined independently for each feature by generalized cross-validation. The search grid included 32 values between 10^0.5 and 10^11, logarithmically spaced with a grid grain of 10^0.33. Estimations were done with a 4-fold cross validation across the four stimulus sets.