Focal versus distributed temporal cortex activity for speech sound category assignment

Sophie Bouton[a,b,c,1], Valérian Chambon[a,d], Rémi Tyrand[a], Adrian G. Guggisberg[e], Margitta Seeck[e], Sami Karkar[f], Dimitri van de Ville[g,h], and Anne-Lise Giraud[a]

[a]Department of Fundamental Neuroscience, Biotech Campus, University of Geneva,1202 Geneva, Switzerland; [b]Centre de Recherche de l'Institut du Cerveau et de la Moelle Epinière, 75013 Paris, France; [c]Centre de Neuro-imagerie de Recherche, 75013 Paris, France; [d]Institut Jean Nicod, CNRS UMR 8129, Institut d'Étude de la Cognition, École Normale Supérieure, Paris Science et Lettres Research University, 75005 Paris, France; [e]Department of Clinical Neuroscience, University of Geneva – Geneva University Hospitals, 1205 Geneva, Switzerland; [f]Laboratoire de Tribologie et Dynamique des Systèmes, École Centrale de Lyon, 69134 Ecully, France; [g]Center for Neuroprosthetics, Biotech Campus, Swiss Federal Institute of Technology, 1202 Geneva, Switzerland; and [h]Department of Radiology and Medical Informatics, Biotech Campus, University of Geneva, 1202 Geneva, Switzerland

Edited by Nancy Kopell, Boston University, Boston, MA, and approved December 29, 2017 (received for review August 29, 2017)

Percepts and words can be decoded from distributed neural activity measures. However, the existence of widespread representations might conflict with the more classical notions of hierarchical processing and efficient coding, which are especially relevant in speech processing. Using fMRI and magnetoencephalography during syllable identification, we show that sensory and decisional activity colocalize to a restricted part of the posterior superior temporal gyrus (pSTG). Next, using intracortical recordings, we demonstrate that early and focal neural activity in this region distinguishes correct from incorrect decisions and can be machine-decoded to classify syllables. Crucially, significant machine decoding was possible from neuronal activity sampled across different regions of the temporal and frontal lobes, despite weak or absent sensory or decision-related responses. These findings show that speech-sound categorization relies on an efficient readout of focal pSTG neural activity, while more distributed activity patterns, although classifiable by machine learning, instead reflect collateral processes of sensory perception and decision.

categorical perception | speech sounds | encoding | decoding | multivariate pattern analysis

The discovery of spatially distributed cortical representations, exploitable for "mind reading," in all domains of cognitive neuroscience during the past decade (1–5) raises fundamental issues about the nature of neural coding in the human brain. These findings, showing that the stimuli present in our environment or mentally evoked are represented in distributed neural activity, are leading scientists even to reconsider the notion of local computational units, such as canonical microcircuits (6, 7). However, whether all the information that is encoded and decodable in our brain contributes to our perceptual representations and our decisions remains an important issue in neuroscience. The relevance of this question is exemplified by the extreme scattering and redundancy of word-meaning representations throughout the brain that was recently shown using voxel-wide modeling of fMRI data (8). Decoding models probing multidimensional statistical dependencies between experimental conditions or stimulus features and spatiotemporal activity patterns distributed across voxels/neuronal populations require careful interpretation (9–11). Distributed neural activity patterns could be taken to indicate either that the information they contain is critical to cognitive operations or simply that they could be used as such, e.g., for stimulus categorization (12). However, the sensitivity of decoding models applied to neurophysiological data and the multidimensional features they rely on to give positive results do not necessarily parallel the capacity of our brain to make use of these neural patterns and multidimensional features in specific tasks (11, 13–15). Data-driven results arising from multivariate decoding models might lead us to conclude that spatially distributed activity patterns are used for performing cognitive operations when in fact they might only follow from these operations, reflect associative processes, or arise from processing redundancy. This concern is relevant at any scale, considering that the implicit assumption behind multivariate decoding methods is that there is functional meaning in the geometry of the decoded pattern, whether this pattern is decoded across individual neurons or across voxels containing several hundred thousand neurons.

Interpreting broadly distributed spatial maps for speech sounds can be particularly difficult. Unlike visual stimuli, whose identity relies heavily on spatial encoding, speech sound identity relies mainly on temporal encoding (16, 17). Despite the relevance of hierarchical temporal processing in speech perception (18), wide cortex coverage with fMRI and more recently with electrocorticography (ECoG) indicates (*i*) that the original acoustic speech signal can be reliably reconstructed from broadly distributed high-frequency neural activity sampled cross-regionally throughout the superior temporal lobe (19–21) and (*ii*) that local phonemic-identity information in speech is poorly encoded by temporally resolved neural activity (2) but is finely represented by distributed cortical patterns covering a significant portion of the left temporal lobe (1). Because optimal decoding occurs when redundant information from contiguous but functionally distinct territories is pooled together, assigning perceptual relevance to such large-scale representations is ultimately tricky and

Significance

When listening to speech, phonemes are represented in a distributed fashion in our temporal and prefrontal cortices. How these representations are selected in a phonemic decision context, and in particular whether distributed or focal neural information is required for explicit phoneme recognition, is unclear. We hypothesized that focal and early neural encoding of acoustic signals is sufficiently informative to access speech sound representations and permit phoneme recognition. We tested this hypothesis by combining a simple speech-phoneme categorization task with univariate and multivariate analyses of fMRI, magnetoencephalography, intracortical, and clinical data. We show that neural information available focally in the temporal cortex prior to decision-related neural activity is specific enough to account for human phonemic identification.

Author contributions: S.B. and A.-L.G. designed research; S.B. performed research; S.B., V.C., R.T., and A.-L.G. contributed new reagents/analytic tools; S.B., V.C., R.T., and A.-L.G. analyzed data; S.B., V.C., R.T., A.G.G., M.S., S.K., D.v.d.V., and A.-L.G. wrote the paper; A.G.G. and M.S. recruited the patients; and S.K. designed the stimuli.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: sophie.l.bouton@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1714279115/-/DCSupplemental.

might conflict with the notion that speech sounds are first spectrotemporally encoded in auditory cortex before being more abstractly recoded in downstream areas (22). Accordingly, focal lesions of the temporal lobe can selectively impair different speech-perception processes (23), and recent studies in monkey even show that auditory decision-making causally relies on focal auditory cortex activity (24).

Listening to speech has long been known to elicit brain activity across both temporal and frontal cortices (25). Whether this activity reflects the use of distributed representations of phonemes as probed with decoding models or focal and selective hierarchical processing as probed with encoding models is incompletely understood. It is questionable whether all neural activity that contributes to spatially distributed patterns, with their specific geometry, reflects a neural code used to perform specific cognitive operations (9, 11, 13), i.e., assigning a speech sound to a category, or multiple/redundant codes that are differentially used depending on specific tasks and cognitive operations. In other words, although multivariate pattern analyses and model-based decoding techniques are promising in translational applications, for instance to decode inner speech in aphasic patients (26), their interpretation in cognitive neuroscience depends on the neurophysiological relevance of the models applied for decoding (27). For instance, demonstrating that a particular stimulus attribute or category can be decoded from regional neuronal activity does not imply that the region is performing categorical processing. In summary, there is a conceptual discontinuity between machine and brain decoding of neuronal activity. As an extreme example, demonstrating that phonemes could be classified using a multivariate machine-learning scheme applied to primary sensory afferents from the ear would not mean that the brain has yet decoded these signals but only that there is, by definition, sufficient information in this auditory input to support subsequent hierarchical decoding.

Using a combination of behavioral, fMRI, magnetoencephalography (MEG), and ECoG data, we attempted to clarify this issue by assessing the ability of a classifier to decode the stimulus category from neuronal responses at various levels in the auditory hierarchy and the ability of a linear model to estimate from neural responses the perceptual processes in a paradigm for assigning speech sounds to categories. We found that the assignment of speech sounds to categories relied on focal neural activity present in a circumscribed part of the auditory hierarchy, at specific peristimulus times, which was supported by the observation that the task could not be performed in a patient with a selective lesion of this circumscribed region. Nevertheless, multivariate machine decoding returned positive results from a large brain network including regions where no stimulus-related evoked activity could be detected, a finding that, in isolation, could suggest that category assignment involved a distributed pattern of activity.

Results

We first explored explicit phoneme recognition using a simple syllable-categorization task and measured global neural activity with fMRI and MEG in 16 and 31 healthy volunteers, respectively (*Methods* and *SI Text*). The subjects had to decide which syllable they heard in a /ba/ /da/ continuum in which the onset value of the second formant (F2) and the F2 slope linearly covaried in six steps (Fig. 1*A*). These two first experiments served to delineate at the whole-brain level those brain regions that were sensitive to (*i*) linear variations of F2 and (*ii*) perceptual decisional effort as assessed using behavior-based negative d' values (Figs. 1*B* and 2*A*) (*Methods* and *SI Text*). Critically, because the slope of the second formant is steeper for the /da/ than for the /ba/ phoneme, we expected the /da/ stimulus to activate a larger cortical surface than the /ba/ stimulus and hence to be associated with a stronger blood oxygenation level-dependent (BOLD) effect (*SI Text*). Both experiments converged to show that F2 variation was specifically tracked by neural activity in the right posterior superior temporal gyrus (pSTG), while perceptual decisional effort involved several regions of the bilateral inferior prefrontal and posterior temporo/parietal cortex and the right

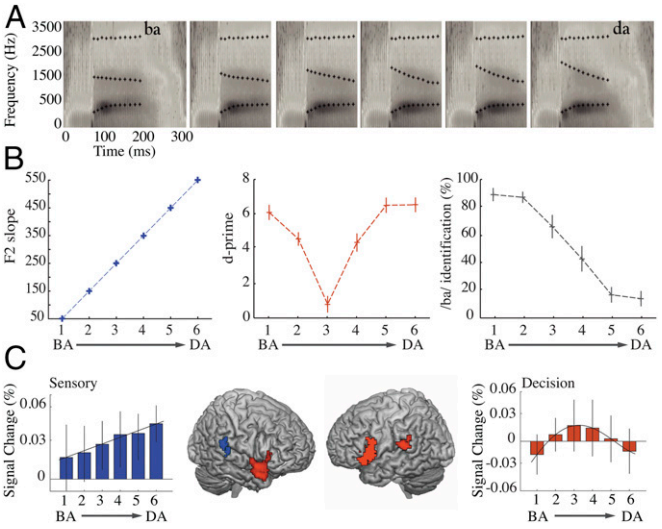

Fig. 1. fMRI results. (*A*) Spectrograms of the stimulus continuum between syllables /ba/ and /da/, synthesized with a linear increase in F2-parameters (1,650:100:2,150 Hz). Full spectrograms at the extremes of the continuum represent /ba/ (*Left*) and /da/ (*Right*) prototype syllables. Spectrograms in the middle are centered on the F2 parameters. (*B*) Values for F2 parameters (iblue; *Left*), average d' (red; *Center*), and percent of syllables identified as /ba/ (gray; *Right*). Data are shown as mean ± SEM. (*C*) Results of the regression analysis. (*Left*) Percent signal change in the right pSTG. (*Center*, right hemisphere) Spatial localization of F2 parameters for neural encoding (blue) and d' (red) in the fMRI BOLD signal, expressed as beta coefficients. Significant clusters were found in the right pSTG (peak MNI coordinates, x, y, z = 42, −34, 7, T = 3.21) for the F2 tracking and in left posterior temporo-parietal (x, y, z = −51, −28, 16, T = 4.41) and bilateral inferior prefrontal (x, y, z = 45, 17, −5, T = 5.26; x, y, z = −48, 8, 22, T = 5.29) cortices for auditory perceptual decision, d'). Images are presented at a whole-brain threshold of $P < 0.001$. (*Right*) Percent signal change in the left inferior prefrontal cortex. The BOLD signal increases with F2 parameters in the right pSTG and with auditory perceptual decision load in the left inferior prefrontal region.

anterior temporal pole (Figs. 1*C* and 2*B* and Fig. S2*D*). These activations, in particular the acoustic encoding of F2 variations, remained focal even at a lenient statistical threshold (Fig. S1). The spatial selectivity of the acoustic tracking of F2 was confirmed by a second fMRI study in which participants had to decide whether they heard /da/ or /ta/. In this case the morphed acoustic cue was no longer spectral (F2) but temporal (voice-onset time, VOT). We found that this acoustic cue was encoded in a restricted region of the left superior temporal gyrus (STG) and superior temporal sulcus (STS) (*SI Text* and Fig. S2). In short, the right pSTG was recruited for encoding the slope of the second formant in the ba–da continuum, and the left STG/STS was recruited for encoding the duration of the consonant part in the da–ta continuum, reflecting the hemispheric dominance for temporal vs. spectral acoustic processing (28).

We used dynamic source modeling of the MEG data to explore the dynamics of acoustic encoding and perceptual decision. We found neural correlates of F2 parameters encoding 120 ms post stimulus onset in the right pSTG. Auditory perceptual decision-related activity appeared in this region at 165 ms and co-occurred with a second peak of F2 encoding activity at 175 ms (Fig. 2*B*). In addition to the spectral response profile within the right pSTG and left prefrontal cortex, a Granger causality (GC) analysis across the two areas showed that neural activity related to F2 and negative d' ($-d'$) corresponded to bottom-up encoding and top-down decoding activity, respectively. Both analyses were associated with neural activity in the high-gamma band for F2 variation and in the beta band for $-d'$, confirming the generic implication of these two frequency ranges in bottom-up and top-down processes (Fig. 2*C*) (29–31). Here, we related decisional effort with the top-down process, in line with a predictive coding

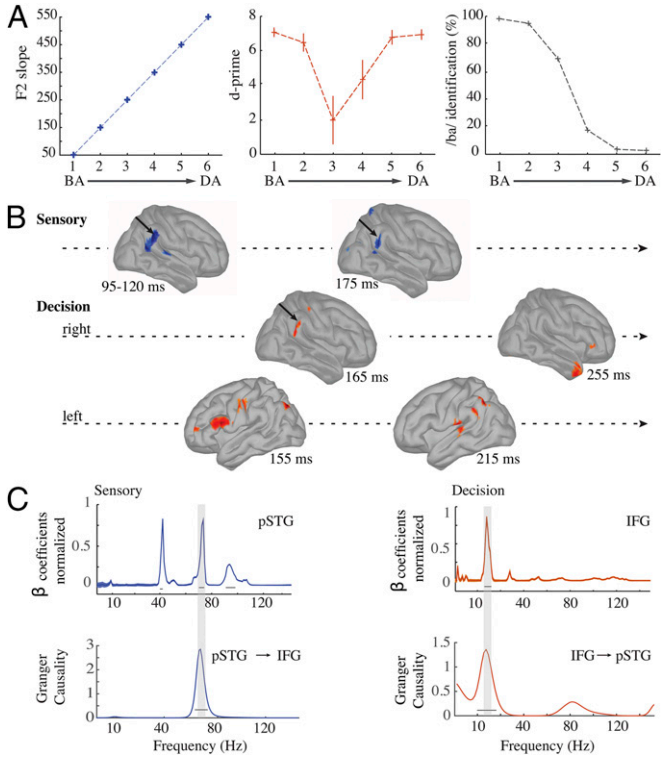

Fig. 2. MEG results. (*A*) Values for F2 parameters (blue, *Left*), average d' (red, *Center*), and percent of syllables identified as/ba/ (gray, *Right*); data are shown as mean ± SEM. (*B*, right-hemisphere) Dynamic spatial localization of the neural encoding of F2 (blue) and d' (red) in MEG signals, expressed as beta coefficients. Only the bootstrapped $P = 0.05$ significance threshold (Bonferroni-corrected) activations are represented. The right pSTG (indicated by black arrows) is first activated at 95–120 ms for encoding F2 parameters and then is reactivated for phonemic decision at ∼165 ms. (*C*, *Upper*) Spectral profile of beta coefficients from regressions between F2 values and neural response in the right pSTG (*Left*), and between $-d'$ values and neural response in the left inferior prefrontal area (*Right*). F2 was dominantly tracked by gamma and high-gamma activity, whereas decisional activity was expressed in the low beta band. Thick black lines indicate significant beta coefficients at $P < 0.05$ (Bonferroni-corrected). (*Lower*) GC results between the right pSTG and the left IFG. Thick black lines indicate significant Granger coefficients at $P < 0.05$ (Bonferroni-corrected). Shaded gray areas highlight the correspondence between beta coefficients and GC peaks: high-gamma band for bottom-up activity from the right pSTG to the left IFG (*Left*); beta band for top-down activity from the left IFG to the right pSTG (*Right*).

view of speech processing (32, 33), whereby the more ambiguous the acoustic input, the stronger are the top-down predictions. As top-down and bottom-up signals are thought to be predominantly associated with neural activity in the beta and gamma band, respectively, we probed the dominant frequency of information transfer between the inferior frontal gyrus (IFG) and pSTG using GC. Our results confirmed that beta activity dominated from the IFG to the STG and that gamma activity dominated in the other direction (29, 31, 34, 35). The MEG findings thus support the straightforward scenario in which phoneme categorical decisions could arise from a rather focal readout of the region that encodes the critical sensory cue (F2) by prefrontal regions (36–38).

Having established the global representational validity of the regions encoding the sensory features of interest (F2 variations), we then sought to examine the responses of these regions at a finer-grained scale using invasive electrophysiology. To maximize signal-to-noise ratio and spatial specificity in the exploration of coincident neural responses to F2 and to auditory perceptual decision, we acquired intracortical EEG (i-EEG) data in three epileptic patients who together had 14 electrode shafts throughout the right temporal lobe (70 contacts). Among the electrode shafts, one penetrated the right temporal cortex through Heschl's gyrus (Fig. 3*B*). The deepest contacts of this auditory shaft strictly colocalized with the region that fMRI detected for F2 variation tracking. The patients performed the same syllable-categorization experiment on a ba/da/ga continuum in which the only changing acoustic cue was the F2 parameter (Fig. 3*A*). Behavioral results show a good detection of ba and da and a slightly less frequent detection of ga (Fig. 3*C*). Strong evoked responses to syllables were present only in the auditory shaft and were more marked/consistent in its two deepest contacts (Fig. 3*D*, *Top Row*); the responses were weak to nonexistent elsewhere (Fig. 4*A*, colored plots). Significant F2 tracking was consistently detected in all auditory contacts (*Methods*), with strong and structured effects in the two deepest ones (Fig. 3*D*, *Middle Row*). Fully consistent with the MEG results, F2 values were encoded by broadband gamma activity (40–110 Hz) from about 150 ms poststimulus onset onward, i.e., 50 ms after F2 appeared in the acoustic signal. Structured and strong neural activity related to F2 tracking was not observed in any of the other contacts of the same patient (patient 1) (Fig. S5). These data suggest that the

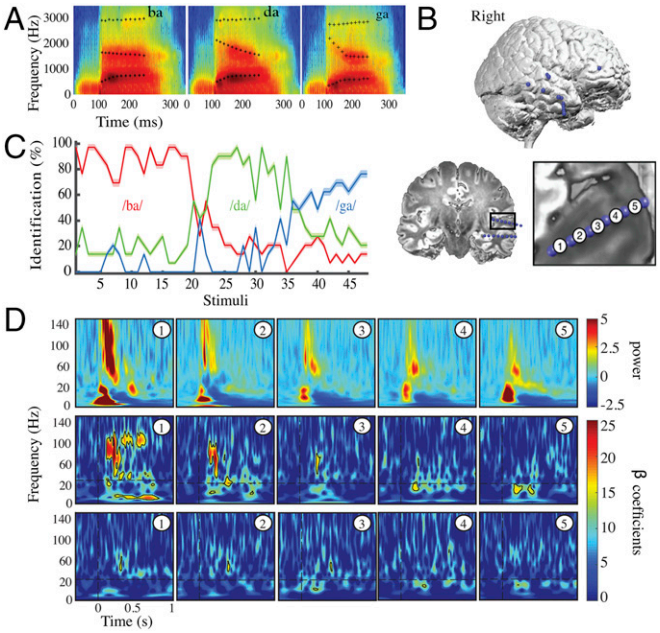

Fig. 3. i-EEG results in patient 1. (*A*) Spectrograms of /ba/, /da/, and /ga/ prototype stimuli synthesized with linear parametric F2 parameters. (*B*) Locations of i-EEG contacts in patient 1. The auditory shaft labeled shaft 1 penetrated the right pSTG and Heschl's gyrus. The patient had five other shafts distributed in the right temporal lobe. Bipolar montages from adjacent contacts are shown in the *Bottom Right* figure. (*C*) Percentage of syllable identification for each category for the three patients. The shaded zone indicates the SEM. The first 20 stimuli are categorized as /ba/, the next 13 stimuli are categorized as /da/, and the last 11 stimuli are categorized as /ga/. (*D*, *Top*) Evoked activity, averaged across stimuli, on each bipolar montage from the deepest contact (number 1) to the most external contact (number 5) in the auditory shaft (shaft 1). (*Middle*) Time–frequency representations of beta coefficients from the regression of F2 values against evoked activity on each contact of the auditory shaft. Significant F2 tracking was found in all contacts of the auditory shaft, with stronger effects in the two deepest contacts. (*Bottom*) Time–frequency representations of beta coefficients from regression of d' values against evoked activity on each contact of the auditory shaft. Decisional effects were significant on the third auditory contact about 200 ms poststimulus onset in the beta band. The vertical dashed lines indicate stimulus onset. The horizontal dashed lines indicate a change in the scaling of the oscillatory power for each time point and each frequency, with a 0.5-Hz resolution below 20 Hz and a 1-Hz resolution above 20 Hz. Black contours indicate significant t tests at $q < 0.05$ (FDR correction).

encoding parameters of the discriminant acoustic cue are available in the right pSTG for syllable recognition, confirming the spatial selectivity for F2 parameters encoding in this region.

Decisional effects were globally weak in i-EEG signals but were significant in the third auditory contact about 200 ms poststimulus onset in the beta band, in agreement with the MEG results, and in the deepest auditory contact about 350 ms poststimulus onset in the gamma band (Fig. 3*D*, *Bottom Row*). Because both fMRI and MEG showed correlates of decisional effort at several other locations in the frontal and temporal lobe, we broadened the analysis in this patient to all contacts of each shaft (Fig. S6). Perceptual decision-related effects were weak, sporadic, and inconsistent. They were significant before 500 ms poststimulus at only two other locations outside Heschl's gyrus: in the right inferior prefrontal cortex (shaft 6, contact 1; consistent with fMRI) (Fig. 1*C*) and in the anterior temporal lobe (shaft 4, contact 4; consistent with MEG) (Fig. 2*B*).

We then sought to address whether focal neural activity could afford syllable categorization. In line with previous findings based on ECoG signals (1, 2), local evoked activity from one contact was sufficiently discriminable to permit syllable categorization using a machine-learning algorithm (maximum correlation coefficient classifier; see *Methods*). Decoding was possible from all individual auditory contacts but worked best from the deepest one (Fig. 4*A* and Fig. S9*A*). Within the other electrode shafts, univariate decoding based on single-contact information was never possible. However, significant multivariate decoding from pooling all contacts in each shaft was significant for shafts 1, 2, 3, 4, and 6, even though it included nonresponsive contacts. Reciprocally, multivariate decoding was not possible in the temporal pole shaft (shaft 5), even though we detected significant perceptual decision-related neural activity in this region with fMRI and MEG.

We subsequently addressed the key question whether the information used by the classifier corresponded to that used in the human decisional process. We examined whether there was a temporal correspondence between the dynamics of decoding, as assessed by time-resolved classification (39, 40), and the presence of time–frequency neural cues that informed the subject's perceptual decision. For this analysis, to ensure the independence of the analyzed dataset (*SI Text*), we no longer probed the decisional effort (the search for information, $-d'$) but the decisional outcome. We approximated the neural cues that were critical to the decisional outcome by the difference in the time–frequency responses of correctly and incorrectly recognized prototype syllables. The correct-minus-incorrect contrast indicates the parts of the neural signal that, if missing, are associated with an erroneous perceptual decision. Note that this contrast matches, as closely as possible, the output of the maximum-correlation coefficient classifier, which tests the extent to which a linear association can correctly predict syllables from neural activity.

Significant time–frequency correlates of correct classification were found only in the three deepest contacts of the auditory cortical shaft (Fig. S7); they were sporadic and inconsistent elsewhere (red frames in Fig. S7 show significant activity for $t <$ 500 ms). In the deepest auditory contact (contact 1 on shaft 1), where both F2 tracking and univariate classification were maximal (Fig. 3 and Fig. S8), cues associated with correct perceptual decisions were present as early as 150 ms, i.e., before the first significant decoding peak at 200 ms (Fig. 4*B* and Figs. S6 and S9*B*). This important finding shows that within 150 ms the right pSTG had encoded enough information about F2 onset frequency and slope to inform correct syllable recognition by the subject and that this information could be exploited by the classifier to distinguish across syllables (*Discussion*).

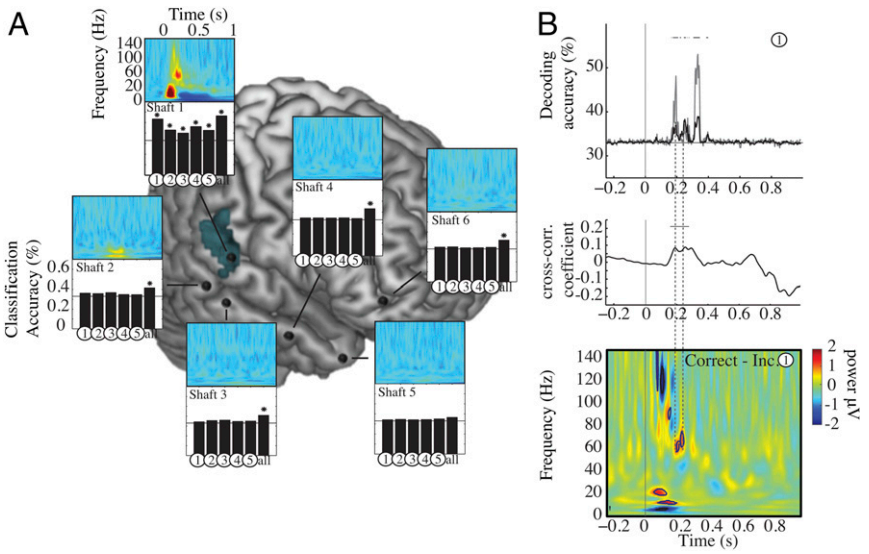

Fig. 4. Decoding in patient 1. (*A*, right hemisphere) Colored panels show time–frequency representation of evoked activity on each shaft. A strong evoked response to syllables is present only in the auditory shaft. Bar graphs show neural decoding through univariate and multivariate classifiers. Histogram bars numbered from 1–5 show the univariate classifier results based on the activity from each contact of each shaft. The bar at the far right ("all") shows the multivariate classifier results based on multidimensional information from all contacts of each shaft. Stars above the black bars signal significant classification accuracy for specific contacts within each electrode shaft ($q < 0.05$, FDR-corrected). Univariate classification was possible from all auditory contacts of shaft 1 overlapping fMRI F2 parameters activation (blue-shaded area) but worked best in the deepest one. Univariate classification failed everywhere except for these auditory contacts, whereas syllable decoding worked above chance using the multivariate approach in shafts 1, 2, 3, 4, and 6. (*B*, *Bottom*) Time–frequency differences between correct-minus-incorrect classification computed on contact 1 of shaft 1. Black borders indicate significant differences in neural activity between correct and incorrect classification scores, in comparison with a zero-mean normal distribution at $q < 0.05$, FDR-corrected. (*Top* to *Bottom*) Temporal relationship between the time course of machine decoding from the deepest auditory contact (*Top*, black line), mean univariate classification from all auditory contacts (*Top*, gray line), and the time–frequency cues used by the subject to make a correct perceptual decision (difference in the time–frequency response between correctly and incorrectly recognized syllables, *Bottom*). The thick gray lines show significant results for each time point (significant decoding accuracy, $q < 0.05$, FDR-corrected). (*Middle*) Cross-correlation coefficients between univariate decoding accuracy and significant correct-minus-incorrect time–frequency clusters. Significant effects were found in the 60- to 80-Hz high-gamma band. The horizontal black line indicates significant cross-correlation coefficients at $P < 0.05$ (Bonferroni-corrected).

Critically, the temporal coincidence between neural correlates of response correctness and machine decoding (Fig. 4*B*) was only partial. It was fairly good for the first two significant decoding peaks (<200 ms) of both single auditory contact decoding and mean univariate decoding across all auditory contacts but was poor for the latest (and strongest) peak. The first two decoding peaks coincided precisely with transient high-gamma activity in the 60- to 80-Hz range (significant zero-lag cross-correlation) (Fig. 4*B*), in line with the observation made with MEG (Fig. 2) that F2 was specifically encoded by neural activity in this frequency range and that 60- to 80-Hz activity preceded decisional activity in the left IFG. However, the third decoding peak had no matching time–frequency event in the correct vs. incorrect activity profile (Fig. 4*B*). These observations indicate that the classifier did not systematically capture those neural cues that informed the subject's decision. Thus, machine classification and human subjects did not exploit the same cues locally. Presumably, the outcome of the mean univariate classifier reflected distributed information that was no longer relevant for or assimilated into neuronal decision variables. The strongest local decoding peak occurred at 370 ms, i.e., more than 250 ms later than the first correctness effect and about 100 ms after the last one and likely reflected postdecisional choice-correlated neural activity.

So far, the results indicate that the phonemic decision was informed by focal early neural activity (<200 ms) that encodes F2 in sustained multifrequency oscillatory activity (Fig. 2). However, distributed subthreshold neural activity not detected by conventional (univariate) analyses of fMRI, MEG, and intracortical EEG data also might contribute to syllable-identity encoding. We therefore addressed whether decoding was possible even from contacts where there was no detectable F2 and from perceptual decision-related activity (Figs. S5 and S6). We broadened the analysis to the 14 shafts of the three patients, including two additional patients who had electrode shafts over the right temporal lobes (n = 14), and performed time-resolved multivariate decoding from all cortical contacts (n = 36). Significant decoding was found at 250, 300, and 600 ms, showing that syllables could be decoded from broadly distributed activity (Fig. 5). To address whether the distributed pattern of activity was driven by local auditory activity, we performed the same analysis without the contribution of the auditory shaft of patient 1. Early classification (<300 ms) dropped below statistical significance, but the latest classification peak at 600 ms remained unaffected (Fig. 5). This result demonstrates that decoding remained possible from cortical contacts that showed neither F2- nor auditory perceptual decision-related activity. We even obtained significant late decoding when deep structures, such as the amygdala and the hippocampus, were included in the multivariate analysis (n = 70 contacts). As each penetrating shaft, except the auditory one, spanned functionally different territories, from the cortex to deeper structures, these findings show that the possibility of decoding neural activity in a multivariate approach does not allow one to conclude that the regions sampled for decoding amount to a meaningful neuronal representation, defined operationally in terms of a correct perceptual categorization. Overall, classification analyses from the i-EEG data confirmed the spatial selectivity of the early critical information involved in ba/da/ga syllable categorization. They also showed that syllable classification was possible from distributed activity (Figs. 4 and 5) that occurred later than the perceptual decision-related effects, as detected with both MEG and i-EEG.

Since the decoding of i-EEG returned positive results when contacts in which no significant neural activity could be detected were pooled together, we sought to explore the spatial distribution of /ba/ and /da/ category decoding using the MEG dataset. The idea was to determine whether whole-brain decoding would be restricted to regions that showed statistically significant activation with all three approaches, MEG (Fig. 2), fMRI (Fig. 1), and i-EEG (Figs. S5 and S6), or would also be possible in regions that did not critically participate in the task. This analysis was expected to provide time-resolved information to appraise whether decoding reflects noncritical processes downstream of sensory encoding and early decisional steps. Such a finding would concur with the i-EEG results suggesting that decoding is possible from brain regions that are only collaterally involved in the cognitive process at stake.

Using whole-brain MEG sensor data and a time-resolved multivariate learning algorithm (maximum correlation coefficient classifier) (*Methods*) (41), we found that speech sound categories could be decoded from very early brain responses in a focal region of the right pSTG (Fig. 6). When we focused our

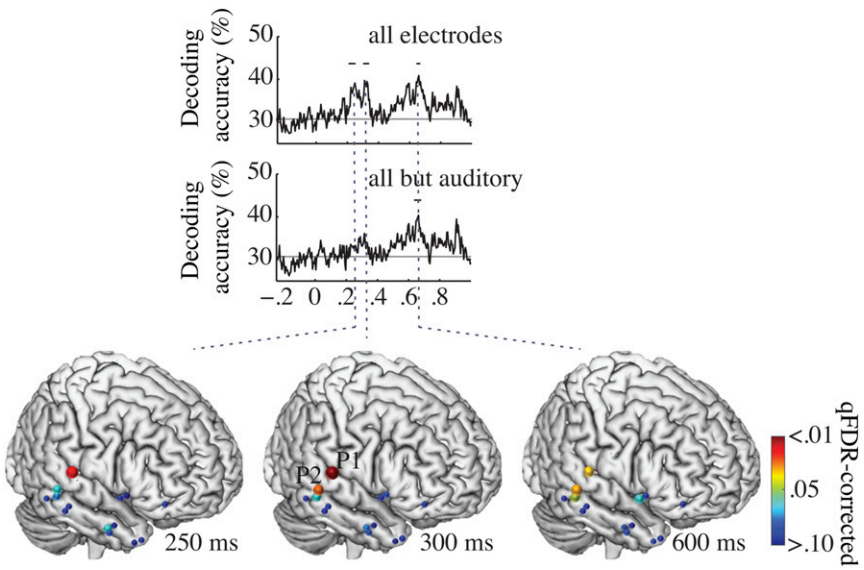

Fig. 5. Decoding in all patients. Time course of the decoding accuracy from multivariate pattern classification with all shafts (*Top*) and without the auditory shaft of patient 1 (*Middle*). Early classification dropped below statistical significance, while the latest classification peak at 600 ms remained unaffected. (*Bottom*, right hemisphere) Location of shafts (n = 14) from which neural activity was recorded during syllable identification (three patients, fixed-effects model). Colored dots show cluster-level significance from $q > 0.10$ to $q < 0.01$ (q-FDR corrected) multivariate classification performed on all shafts. Dot size is proportional to q. Significant classification was observed at 250, 300, and 600 ms, showing that syllables could be decoded from broadly distributed activity. At 300 ms, P1 refers to patient 1, and P2 refers to patient 2.

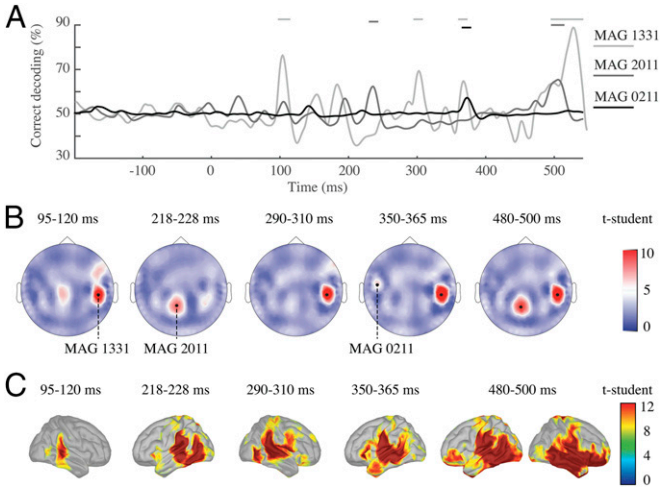

Fig. 6. Decoding of MEG data reveals bilateral temporo-frontal cortex involvement in speech-sound categorization. (*A*) Percentage of correct decoding over time for each of the magnetometers (MAG 1331, MAG 2011, and MAG 0211) showing significant decoding activity. On the *x* axis, zero corresponds to stimulus onset; on the *y* axis, 50% indicates chance performance. Horizontal dark and light gray lines indicate significant decoding ($q < 0.05$, FDR-corrected). (*B*) Sensor topographies depicting the average decoding response in magnetometers averaged within each of the five windows of interest. Black dots indicate the position of the three magnetometers showing significant decoding activity. (*C*) Source localization at key decoding times. Source localization for MEG signals are displayed on a standard cortex at 100, 220, 300, 360, and 500 ms poststimulus onset. Color bars at the right in *B* and *C* indicate Student *t*-test values.

analysis on those sensors that contained significant information about syllable identity (*Methods*), we found that the activity recorded by the sensor MAG 1331 could be categorized as early as 100 ms, with up to 78% accuracy ($t = 13.01$, $q < 0.001$, Cohen's $d = 4.7$). Critically, syllable identity could also be decoded at later time points, first at 220 ms on the sensor MAG 2011 and then at 350 ms on the sensor MAG 0211, with scores reaching 63% accuracy ($t = 7.17$, $q < 0.001$, Cohen's $d = 2.5$) and 58% accuracy ($t = 6.37$, $P < 0.001$, Cohen's $d = 2.2$), respectively. Corresponding source analyses then revealed that the decoding at 220 ms poststimulus onset arose from the left STG and STS, two regions that were not primarily responsive to acoustic cue tracking and perceptual decision (Fig. 2). Source analysis further showed that the decoding at 350 ms arose from the left IFG and thus corresponded to late decisional effects. Together, these results show that, while decoding is most accurate in the region that critically encodes the acoustic information (the right pSTG), it is also subsequently possible from noisier activity in a broad left-lateralized network that contains associative information about the selected syllable. Interestingly, significant decoding was seen again in the right pSTG at 500 ms poststimulus onset with very high accuracy (89%, $t = 15.10$, $q < 0.001$, Cohen's $d = 5.5$). This suggests that information propagation occurring across the whole language network between 100 and 500 ms contributed to improve the quality of the categorical representations at a postdecisional stage.

Discussion

In this series of studies, we addressed whether human listeners' perceptual decisions about the identity of speech sounds arose from distributed representations (42) or whether, following more closely the principles of hierarchical processing, syllable categorization was informed by the efficient readout of a restricted cortical region that contains limited but key neural information, as recently shown in monkeys (24). Our data converge to show that correct decisions about speech-sound identity were informed by local and time-limited information (<300 ms) present in the right pSTG. Even though phonemic contrasts are most often associated with activations of the left STG (43), the right STG performs low-level spectral-based acoustic processing that is relevant for speech decoding (44, 45) and categorization (46, 47). Note that the right specificity of speech-related operations can easily be missed when acoustic processing is not explicitly orthogonalized from decision-related neural activity (48). Moreover, hemispheric dominance in speech processing depends heavily on the task. Here, subjects were forced to process a single cue (spectral or temporal) to categorize ambiguous signals, which does not happen under natural listening conditions in which contextual information and redundant acoustic cues are available. Our findings confirm that spectral- and time-based analyses of speech sounds involve the right and left STG, respectively (Fig. 1 and Fig. S8). That we observed tonotopic encoding of F2 in a region contiguous to, rather than within, right Heschl's gyrus presumably reflects that the frequency range spanned by F2 is rather limited and more extensively processed in a region specialized for vocalic formants (49). Our findings also confirm that in acoustically challenging situations such as those used in the current experimental design, the left IFG is mobilized (50) and interacts with the temporal cortex in sequential loops of bottom-up and top-down processes (48, 51–53). Importantly, our fMRI and MEG results conjointly show that the focal readout of sensory encoding regions by prefrontal regions accounts for the decision variability relative to the perceptual state and that the magnitude of neural activity associated with sensory processes depends on the discrepancy between heard and expected signals (Figs. 1*C* and 2*C*) (54, 55).

Whether stimulus identity is retrieved from focal neural activity or from distributed activity patterns is a fundamental issue in neural coding theory. Our MEG and i-EEG decoding results both show that the right pSTG is critical for perceptual decisions, while distributed activations across the frontal and temporal cortices reflect the reuse of sensory information for higher-level operations, such as extraction of meaning, audiovisual integration, and so forth. It has repeatedly been observed that behavior is explained as well by taking the activity of one or a limited set of neurons as it is by considering a large-scale population (56, 57). This puzzling observation is backed up by computational models showing that pooling together information from more neurons does not improve behavioral sensitivity. This could reflect the fact that neuronal responses are strongly correlated both intrinsically by neural noise and by stimuli, so that the information contained in a neural population saturates as the number of neurons increases (58, 59). Recently, a combination of experimental and theoretical data suggests that in most cases behavior is best explained by optimal readout of a limited set of neurons (60). The authors then conclude that the neural code is redundant and that some areas are decoded nearly optimally while others are not read out efficiently. They propose that deactivating these regions should not affect behavior.

Reciprocally, as we tested in a brain-damaged patient that was carefully selected with respect to lesion extent and localization (*SI Text* and Fig. S10), deactivating regions that are optimally decoded would be expected to impair behavior. The lesion of one of the two focal regions that we identified as key for F2-based syllable categorization dramatically disrupted performance. The impairment was selective to the type of stimulus but not to the type of task, as categorization of similar syllables (da/ta) based on temporal cues, i.e., VOT, was spared. This also was expected, as syllable categorization based on VOT specifically involved the left middle temporal cortex (Fig. S2). The selectivity of the impairment with respect to the acoustic material shows that for this type of task there was no distributed rescue system. These behavioral data in a single patient remain, of course, nongeneralizable, and further lesion data will be necessary to confirm our results. Moreover, a full demonstration that, even though phoneme categories are represented in a distributed fashion, focal/early sensory information is predominantly used to categorize speech sounds would, strictly speaking require showing that distributed lesions do not impair task performance. This demonstration, however,

cannot be performed, because at the extreme, if the whole language system is injured, no linguistic operation can be achieved at all. Although, using a variety of approaches, we tried to collect as many arguments as possible to address how information is used to make perceptual decisions about speech sounds, we acknowledge that the present study can be only partly conclusive.

Critically, however, our results show that a distinction should be made between the classification of syllable-induced neuronal activity by machine learning and the neuronal decoding made by the brain to achieve syllable categorization. We found that distributed noise-level neural information, which did not carry reproducible and statistically detectable information about the sensory cue (F2 parameters) or about the categorical decision process (61), was nonetheless sufficient to inform a classifier. While this confirms previous observations that phonemic information is present in a distributed manner (1, 2), the fact that classification was possible only on late responses once the region encoding the critical acoustic cue (F2 slope) was removed from the multivariate analysis suggests that distributed phonemic representations reflect the diffusion of the information throughout the language system. While distributed information might be useful for late association-type linguistic processes, it did not appear necessary for the categorization task, as correctness effects occurred focally and early in time. These findings show that neuronal activity containing information about speech sound categories is not uniformly useful for categorizing these sounds. More generally, our findings highlight the fact that decoding algorithms, even if they can make use of distributed information that might reflect a brain state context (57) or mental association, do not indicate which regions are necessary or causal for the cognitive process at stake—here a categorical choice. These results hence suggest that distributed information about categories might reflect the redundancy of noise-level information in the brain, i.e., associative neural activity, rather than a spatial neural code that is accessed in parallel when making a speech-category perceptual decision.

Categorizing auditory syllables is a relatively low-level process, which in theory could be underpinned by distributed representations, but in our data this appears not to be the case. What then might be denoted by previously observed broadly distributed representation maps (8) and, in particular, by phonemic maps organized along articulatory features (62)? Most of these important findings (1, 52, 63) were obtained in natural listening conditions. In the work by Mesgarani et al. (1), for instance, maps were drawn from cortical surface neural activity measured 150 ms after each phoneme in connected speech, implying that activity at each time point integrated the distributed activity of several preceding phonemes and reflected coarticulation and contextual associations from the preceding words. When passively listening to connected speech, there was likely no explicit serial phoneme decoding but rather a global readout of sentences, which likely required superficially accessing many phonemic representations at once. In the same way as a computer keyboard spaces letters to minimize interference when typing them, our brain might organize the phonemic space as a function of how we need to retrieve them for production (54), i.e., following a feature-based logic. That such organizations exist does not imply that they are used as such for explicit syllable recognition, just as reading words through a keyboard spatial organization would largely be suboptimal. While the present findings do confirm the existence of distributed phonemic representations, they also question the use our brain makes of them in an explicit speech-perception context, as phoneme recognition does not seem to rely on distributed patterns of activity. Importantly however, it might be the case that, during natural speech perception, cross-hierarchical readout of redundant/correlated neural activity is genuinely exploited as a form of trade-off between accuracy of single phoneme identification (focal) and joint access to multiple representations (distributed). It will be essential in the future to address whether suboptimal decoding of large neuronal populations could be an optimal way to handle access to multiple representations.

Methods

Subjects. Twenty-eight healthy subjects participated in the MEG study (16 females; age range: 22–31 y), and 16 participated in the fMRI study (nine females; age range: 22–29 y). i-EEG was recorded in three epileptic patients (one female; ages: 44, 25, and 65 y) who underwent surgery for clinical monitoring and in one patient (female; age: 77 y) who was tested behaviorally 8 mo after an ischemic stroke. All participants were right-handed, native French speakers and had no history of auditory or language disorders. The experimental protocols for the participation in the MEG and fMRI experiments were approved by the Inserm ethics committee in France (biomedical protocol C07-28), and the protocol for studies in epileptic and stroke patients was approved by the University of Geneva Hospital in Switzerland (13–224). All participants provided written informed consent before the experiment.

Stimulus Synthesis and Behavioral Testing. High-quality speech stimuli were synthesized using an original morphing method based on a modified linear predictive coding (LPC) analysis synthesis scheme (64). Using an exponential pole-morphing approach, the second formant transition was morphed to build a linear speech sound continuum between /ba/, /da/, and /ga/ (65, 66). The initial (prototypical) /ba/, /da/, and /ga/ syllables were natural voice signals, down-sampled to 16 kHz, aligned on their burst starting time, and cut to the same length (360 ms). From the LPC analysis of natural voice sounds, the formant structure was extracted for each prototypical syllable. For resynthesis, the excitation signal from the LPC analysis was discarded, and an artificial excitation signal was used. The excitation signal consisted of a low-pass-filtered pulse train for the voiced part, additional white noise for the burst, and a small amount of white noise throughout the entire stimulus. The time dependency of the fundamental frequency, $f0 = f(t)$, was a piecewise constant and simplified version of that extracted from the original /ba/ natural voice signal. Throughout the continuum, the excitation signal, the global amplitude of the stimulus, and the first and third formant transitions were kept constant.

A six-item /ba/ /da/ continuum was presented to healthy subjects and to the stroke patient. A longer 48-item continuum, /ba/ /da/ /ga/, was used for testing epileptic patients to obtain more responses around syllable boundaries to compare correct and incorrect categorization. Note that /ba/ and /da/ categories differed only on the F2 dimensions, and hence processing this single cue was sufficient for correct perception. Another six-stimuli /da/ /ta/ continuum was used for both the behavioral and second fMRI control experiments. In that continuum, we varied the length of the VOT by deleting or adding one or several voiced periods in the signal, before or after the burst, using audio editor software.

Tasks Design. Auditory stimuli were presented using Psychophysics-3 Toolbox and additional custom scripts written for Matlab version 8.2.0.701 (MathWorks). Sounds were presented binaurally at a sampling rate of 44,100 Hz and at a comfortable hearing level individually set before the experiment via earphones. Before the experiment, each participant undertook a short session during which the minimum amplitude level leading to 100% categorization accuracy was estimated using an adaptive staircase procedure. This threshold was used to transmit the stimuli (mean 30 dB sensation level). Each continuum was delivered to participants in two independent sessions of 240 trials each for fMRI recording and 270 trials each for MEG recording. The experiment used for epileptic patients comprised 144 trials.

Participants were asked to perform an identification task. Each trial comprised one sound (randomly chosen among the 6 or 48 stimuli of the continuum), followed by 1 s of silence; then, a response screen with the written syllables "ba" and "da" (in MEG, fMRI, and behavioral sessions) or "ba," "da," and "ga" (in i-EEG sessions) was displayed. Syllables were randomly displayed from right to left on the screen to prevent motor preparation and perseverative responses. During fMRI recording, the appearance of the response screen was randomly jittered 100, 300, or 500 ms after the silence gap. Participants indicated their response on the syllable by pressing the corresponding left or right response button as quickly as possible. Subjects' responses were purposely delayed to avoid temporal overlap between perceptual processes and motor effects due to button press. Response times hence do not constitute relevant data. To limit eye movements, subjects were asked to fixate the central cross and to blink only after giving their motor response. After the response, a jittered delay varying from 3 to 5 s led to the next trial.

MEG Recording and Preprocessing. Continuous cerebral activity was recorded using an Elekta Systems MEG device, with 102 triple-sensor elements, each composed of two planar gradiometers and one magnetometer. MEG signals were recorded at a sampling rate of 1 kHz and were online band-pass filtered between 0.1 and 300 Hz. A vertical electro-oculogram was recorded simultaneously.

Before MEG recording, the headshape for each participant was acquired using a Polhemus system. After the MEG session, an individual anatomical MRI was recorded [Tim-Trio; Siemens; 9-min anatomical T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE), 176 slices, field of view = 256, voxel size = 1 × 1 × 1 mm^3]. MEG data were preprocessed, analyzed, and visualized using dataHandler software (wiki.cenir.org/doku.php), the Brainstorm toolbox (67), and custom Matlab scripts. A principal component analysis (PCA) was performed through singular-value decomposition function of numerical recipes to correct artifacts (low derivation). The first two components from the PCA were zeroed, and the signal matrix was recomputed. PCA rotated the original data to new coordinates, making the data as flat as possible. The data were then epoched from 1 s before syllable onset to 1 s after syllable offset. Another PCA was then performed on the epoched data when blinks occurred. PCA components were visually inspected to reject the one capturing blink artifacts. On average, 2.1 ± 0.7% of trials per participant (mean ± SEM) were contaminated by eye-movement artifacts and were corrected before further analyses.

fMRI Recording and Preprocessing. Images were collected using a Verio 3.0 T (Siemens) whole-body and radio-frequency coil scanner. The fMRI BOLD signal was measured using a T2*-weighted echoplanar sequence (repetition time = 2,110 ms; echo time = 26 ms; flip angle = 90°). Forty contiguous slices (thickness = 3 mm; gap = 15%; matrix size = 64 × 64; voxel size = 3 × 3 × 3 mm) were acquired per volume. A high-resolution T1-weighted anatomical image (repetition time = 2,300 ms; echo time = 4.18 ms; T1 = 900 ms; image matrix = 256 × 256; slab thickness = 176 mm; spatial resolution = 1 × 1 × 1 mm) was collected for each participant after functional acquisition. Image preprocessing was performed using SPM8 (The Wellcome Trust Centre for Neuroimaging, University College London, London, www.fil.ion.ucl.ac.uk/spm/). Each of the four scanning sessions contained 400 functional volumes. All functional volumes were realigned to the first one to correct for inter-scan movement. Functional and structural images were spatially preprocessed (realignment, normalization, smoothed with an 8-mm FWHM isotropic Gaussian kernel) and temporally processed using a high-pass filter with a cutoff frequency of 60 Hz. We then checked data for electronic and rapid-movement artifacts using the ArtRepair toolbox (cibsr.stanford.edu/tools/human-brain-project/artrepair-software.html). Artifacted volumes were substituted by linear interpolation between contiguous volumes and were explicitly modeled in the following statistical analyses. Estimated head movements were small compared with voxel size (<1 mm); 3.2 ± 0.3% of the volumes were excluded due to rapid head movements (>1.5 mm/s).

i-EEG Recording and Preprocessing. Electrophysiological activity was recorded over arrays of depth electrodes surgically implanted to identify epilepsy focus. i-EEG was recorded (Ceegraph XL; Biologic System Corps.) using electrode arrays with eight stainless contacts each (electrode diameter = 3 mm, intercontact spacing = 10 mm; AD-Tech) implanted in several brain regions in the right hemisphere (Figs. 3–5). We determined the precise electrode shaft locations by coregistering a postoperative computed tomography scan with a high-resolution anatomical MRI template. For the i-EEG recordings, we used a bipolar montage in which each channel was referenced to its adjacent neighbor. We sampled the i-EEG signal at 1,024 Hz for patient 2 and at 2,048 Hz for patients 1 and 3.

Steady-state frequency spectra were estimated using a standard Fourier transform from 1 s before to 1 s after the offset of the stimulus. Time–frequency power was defined as the single-trial square amplitude estimates of complex Fourier components. Time–frequency analyses were carried out using the Fieldtrip toolbox for MATLAB (68). The spectral power of MEG oscillations was estimated using a family of complex Morlet wavelets, resulting in an estimate of power at each time point and each frequency. We restricted the analysis to frequencies between 2 and 150 Hz, spanning the whole range of relevant brain rhythms. Note that the time–frequency transform uses frequency-dependent wavelets (from three to seven cycles per window), with decreasing time-windows with increasing frequency.

Neural Encoding of Parametric Information. We regressed out single trials of MEG, fMRI, and i-EEG signals against (*i*) the acoustic dimension, corresponding to F2 parameters [the onset value of the second formant (F2) and the F2 slope linearly covaried in six steps] or to the VOT (the voicing length before and after the consonant burst varied in six steps)], and (*ii*) the categorization difficulty dimension corresponding to the inverse of the discriminability index from signal detection theory ($-d'$). These two dimensions are naturally orthogonal ($r = 0.02$, $P > 0.20$). A general linear regression model was carried out separately for each dimension (sensory encoding and decisional effort) along the stimuli and was finally averaged across participants to produce a group-level grand average. That approach was adopted to disentangle the neural correlates of basic bottom-up perceptual processing indexing the tracking of the acoustic cue from the neural correlates of the categorization difficulty reflecting the distance of each stimulus from the phoneme identity criterion (48, 69).

fMRI: Neural Encoding of Parametric Information. Statistical parametric *t* scores were obtained from local fMRI signals using a linear multiple regression model with sensory encoding (F2 parameters or VOT value for each condition) and decisional effort ($-d'$ value reported by each subject for each trial) as covariates. Regression parameters were estimated in every voxel for each subject, and parameter estimates then were entered in a between-subject random-effects analysis to obtain statistical parametric maps. We identified brain activation showing significant contrasts of parameter estimates with a voxelwise ($T = 3.21$, $P < 0.005$, uncorrected) and clusterwise ($P < 0.05$, uncorrected) significance threshold. All reported activations survived false discovery rate (FDR) correction for multiple comparisons ($P < 0.05$) (70). Anatomical locations were determined based on automated anatomical labeling. Regressors of interest were constructed by convolving functions representing the events with the canonical hemodynamic response function. For each continuum, a categorical regressor modeled the "sound" event using a Dirac function time locked to syllable onset. Two hierarchically orthogonalized parametric regressors (referred to as "sensory encoding" and "decisional effort" regressors) were added to the sound regressor to capture the modulation of BOLD activity as a function of F2 variation tracking and categorization difficulty. For illustrative purposes (Fig. 1C and Fig. S2C), we used the rfx_plot toolbox (71) to split F2 and d' parametric regressors into six new onset regressors (simple onset regressor without a parametric modulation), each containing all events for a particular level of the stimulus continuum. Beta weights were reestimated for each of these six onset regressors and were averaged across all subjects to get the corresponding percent of signal change.

MEG: Neural Encoding of Parametric Information. We first used single-trial signals on each sensor to perform parametric regressions at successive times from −0.2 to 1 s following stimulus onset. For each participant and each sensor, we calculated the time course of beta coefficients and then computed cortical current maps with Brainstorm using the weighted minimum-norm estimation approach, meaning that the time series for each source is a linear combination of all time series recorded by the sensors (72). Sources were estimated for each subject on the basis of individual MRI images. After realignment and deformation of each subject's cortical surface, sources were projected onto the standard Montreal Neurological Institute (MNI)/Colin27 brain to perform grand mean averages. We then performed within-group statistics to show the sensitivity to sensory encoding and decisional effort dimensions. Note that both of the two transformations applied to the data (regression and source-projection) capture a linear relationship between the observed and the expected data and can thus be implemented in either order. This method was used to localize the sources of sensory and perceptual decision components and to demonstrate that sensory and decisional processing are hierarchically organized in time.

Single-trial–evoked signals on each sensor were also used to compute source current maps for each trial. The inverse operators were generated with the default MNE parameters and were applied at the single-trial level. The estimated sources were morphed to the MNI brain. We then extracted single-trial neural activity from regions of interest defined according to Destrieux's atlas (73) (G_temp_sup-Plan_tempo, G_temp_sup-G_T_transv, G_front_inf-Opercular, G_front_inf-Triangul). Single-trial–evoked responses projected on these selected sources were used in two ways:

i) For each participant, we regressed out single-trial neural activity to estimate spectral power of the beta coefficients via a standard Fourier transform. Time–frequency analyses were carried out according to exactly the same parameters defined in the previous paragraph, *i-EEG Recording and Preprocessing*. We thus estimated the trial-to-trial variability in neural signal from regions of interest at a given frequency that describe sensory encoding or decisional effort (*t* test against zero, $P < 0.05$, Bonferroni-corrected).

ii) For each participant, source activity in the pSTG and in the left IFG was used to measure GC. While GC is classically used to assess causal influence between two time series, we here computed GC for nonstationary time series, such as oscillating neural signals (74, 75). We used a nonparametric test by computing a spectral density matrix factorization technique on complex cross-spectra, obtained from the continuous wavelet transform of source-reconstructed MEG time series. We then assessed the linear directional influence between two brain areas, the pSTG and the left IFG.

GC was computed twice:

- From right pSTG to left IFG to determine whether activity in the IFG could be predicted at trial *t* by including past activity from both the pSTG and the IFG. Here we assume the information flow to be bottom-up.
- From left IFG to right pSTG to determine whether activity in the pSTG could be predicted at trial *t* by including past activity from both the pSTG and the IFG. Here we assume the information flow to be top-down.

Because we computed GC on nonstationary neural signals (i.e., evoked activity from reconstructed sources), GC spectra were obtained in a nonparametric manner by using wavelet transform, without going through the multivariate autoregressive model fitting (74), and spectral GC was intended to reveal a frequency-resolved GC. To do so, we used a spectral matrix factorization technique on complex cross-spectra obtained directly from wavelet transforms of the data. Wavelet transforms were computed at any instant (1-ms resolution) of the syllable duration (360 ms) on a trial-by-trial basis for each subject. For each subject, we then computed the mean GC across trials and the corresponding SD. The original GC spectra were standardized to obtain a vector of z-values, one for each frequency.

We tested for significant frequency peaks separately for bottom-up and top-down GC direction, directly comparing the z-transformed vectors obtained from GC spectra to a zero-mean normal distribution, and corrected for multiple comparisons with the Bonferroni method at $P < 0.05$. Our decision to focus on the left IFG was empirically motivated. Previous papers (e.g., refs. 76–78) have shown that the left IFG is consistently involved in articulatory processing during speech perception and also in lexical information retrieval, both skills that are engaged when categorizing ambiguous speech sounds, i.e., when the internal perceptual decision criterion is difficult to reach, as in the current study.

MEG: Decoding Analyses. Decoding analyses were performed with the Neural Decoding Toolbox (79), using a maximum correlation coefficient classifier on single-trial–induced responses across all MEG sensors. Data from both magnetometers and gradiometers were used. The pattern classifier was trained on the response given by the participant and computed the correlation between MEG data and the syllable identified (/ba/ or /da/) on each trial. More specifically, the classifier was trained on 80% of the data, and its performance was evaluated on the withheld 20% of the test data. The splitting procedure between training and testing data was performed 50 times to reduce the variance of the performance estimate. The reported final decoding accuracy is the average accuracy across the 50 decoding results. Classification accuracy is reported as the percentage of correct trials classified in the test set averaged over all cross-validation splits.

Additionally, an ANOVA based on the second-level test across subjects was applied to the test data to select those sensors that were significantly sensitive to syllable identity at each time point. We then assessed statistical significance using a permutation test. To perform this test, we generated a null distribution by running the decoding procedure 200 times using data with randomly shuffled labels for each subject. Decoding performance above all points in the null distribution for the corresponding time point was deemed significant with $P < 0.005$ (1/200). The first time decoding reaching significantly above chance was defined when accuracy was significant for five consecutive time points. Source localization associated with the decoding results was computed from evoked trials using the MNE source-modeling method (see above).

i-EEG: Neural Encoding of Parametric Information. We performed the same parametric regressions on i-EEG recordings from patient 1. These analyses were done only on that patient, as he was the only patient for whom one shaft showed a significant induced response to syllable perception. Shaft 1 colocalized to the site (the right pSTG) where spectral cue tracking was found with fMRI and MEG. We selected the five deepest contacts on each shaft; those contacts were located between the Heschl's gyrus and the STG on shaft 1. Parametric regressions were carried out at successive times, *t*, from −0.2 to 1 s poststimulus onset, on each selected bipolar derivation, i.e., from the deepest (1) to the most external (5) contact. We computed the power in each frequency band at each time point of each beta coefficient, with a millisecond resolution, similar to the induced power (between 2 and 150 Hz, with a 0.5-Hz resolution below 20 Hz and with a 1-Hz resolution above 20 Hz) by applying a TF wavelet transform, using a family of complex Morlet wavelets (m = 3–7). For each contact, a null distribution was computed by repeating the identical regression procedure 1,000 times with shuffled regressors. We used standard parametric tests (*t* test against zero) to assess the statistical significance of each parametric regression. The type 1 error rate (FDR) arising from multiple comparisons was controlled for using nonparametric cluster-level statistics (80) computed across contacts, time samples, and frequencies. We did not need to correct for multiple comparisons across electrode shafts, as statistical tests were run independently for each contact of each electrode shaft.

i-EEG: Decoding Analyses. We used a maximum correlation coefficient classifier [Neural Decoding Toolbox (79)] on single-trial–induced responses. The classifier was trained to classify the i-EEG data into three categories that corresponded to the three syllables identified by the patient (/ba/, /da/, or /ga/). We applied the decoding procedure on time series using a cross-validation procedure in which the classifier was trained on 90% of the trials and was tested on the remaining 10%. Our recordings consisted of three repetitions of each stimulus condition (48 stimuli from /ba/ to /ga/) for patient 1 (144 trials to be categorized) and six repetitions of each condition (288 trials to be categorized) for patients 2 and 3. The cross-validation procedure was repeated 1,000 times with a different split between training and testing datasets on each iteration, and the results were averaged across all iterations.

We estimated single-trial decoding of the neuronal response induced by different syllables using both uni- and multivariate classification. The univariate classification was applied to each bipolar derivation (i.e., to each of the five contacts of each shaft), whereas the multivariate classification was performed on neural activity from every bipolar derivation of one shaft (i.e., on the five contacts of each shaft, pooled together) and then on all bipolar derivations of patient 1 (on all contacts of the six shafts, pooled together), and finally on all three patients (on all contacts of the 14 shafts, pooled together). Single-shaft multivariate decoding was compared with mean univariate decoding computed first from each contact and then averaged. Decoding accuracy is expressed as the percent of correctly classified trials in the test set. A null distribution was computed by repeating the identical classification procedure 1,000 times with shuffled labels. We defined the number of classification repetitions with respect to the number of multiple comparisons done from each contact (FDR-corrected for univariate decoding performed on each of the five contacts, time samples, and frequencies, FDR-corrected for multivariate decoding performed on each of the six shafts, time samples, and frequencies, and FDR-corrected for multivariate decoding performed on all shafts together, time samples, and frequencies). Decoding accuracy was considered significant at $q < 0.05$ if accuracy exceeded the randomized classification at two consecutive time points.

i-EEG: Correct-Minus-Incorrect Differences. The psychometric identification function with percentage reporting /ba/, /da/, or /ga/ was defined along the corresponding continuum. Boundary separations determine the accuracy of categorical choice: The steeper the slope, the more accurate was the perceptual decision. Patient's ratings along the continuum were used to split responses into correct and incorrect trials. We subsequently computed the difference in neural activity from selected bipolar derivations between correct and incorrect conditions and then compared it with the zero-mean normal distribution thresholding at $q < 0.05$ [FDR-corrected for multiple comparisons on shafts (30 shafts tested for patient 1), time, and frequency dimensions]. This procedure was repeated 1,000 times with shuffled labels for correct and incorrect conditions.

ACKNOWLEDGMENTS. We thank Lorenzo Fontolan and Clio Coste for methods support and Aaron Schurger, Luc Arnal, Andreas Kleinschmidt, David Poeppel, Virginie van Wassenhove, Corrado Corradi Dell'Acqua, Narly Golestani, and Clio Coste for comments and useful discussions about earlier versions of this manuscript. This work was funded by European Research Council Compuslang Grant Agreement GA260347 and Swiss National Fund (SNF) Grant 320030_149319 (to A.-L.G.), SNF Grants 140332 and 146633 (to M.S.), Agnece Nationale pour la Research (ANR) Grants ANR-10-LABX-0087 IEC (Institut d'Étude de la Cognition), ANR-10-IDEX-0001-02 PSL* (Paris Sciences et Lettres) (program "Investissements d'Avenir"), and ANR-16-CE37-0012-01 (to V.C.), and SNF Grant P300P1_167591 (to S.B.).

1. Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* 343:1006–1010.
2. Chang EF, et al. (2010) Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13:1428–1432.
3. Yan Y, et al. (2014) Perceptual training continuously refines neuronal population codes in primary visual cortex. *Nat Neurosci* 17:1380–1387.
4. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–866.
5. Carota F, Kriegeskorte N, Nili H, Pulvermüller F (2017) Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cereb Cortex* 27:294–309.
6. Grill-Spector K, Weiner KS (2014) The functional architecture of the ventral temporal cortex and its role in categorization. *Nat Rev Neurosci* 15:536–548.
7. Pinotsis DA, et al. (2017) Linking canonical microcircuits and neuronal activity: Dynamic causal modelling of laminar recordings. *Neuroimage* 146:355–366.

8. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532:453–458.
9. Ritchie JB, Kaplan D, Klein C (2017) Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. bioRxiv: 10.1101/127233.
10. Weichwald S, Grosse-Wentrup M (2017) The right tool for the right question—Beyond the encoding versus decoding dichotomy. arXiv:1704.08851.
11. Hebart MN, Baker CI (August 4, 2017) Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, S1053-8119(17)30652-3.
12. Eger E, Ashburner J, Haynes J-D, Dolan RJ, Rees G (2008) fMRI activity patterns in human LOC carry information about object exemplars within category. *J Cogn Neurosci* 20:356–370.
13. Anderson ML, Oates T (2010) A critique of multi-voxel pattern analysis. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (Cognitive Science Society, Austin, TX), pp 1511–1516.
14. Dilks DD, Julian JB, Paunov AM, Kanwisher N (2013) The occipital place area is causally and selectively involved in scene perception. *J Neurosci* 33:1331–1336a.
15. Anderson ML (2015) Précis of after phrenology: Neural reuse and the interactive brain. *Behav Brain Sci* 39:1–22.
16. Shamma S, Lorenzi C (2013) On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. *J Acoust Soc Am* 133:2818–2833.
17. Moon IJ, et al. (2014) Optimal combination of neural temporal envelope and fine structure cues to explain speech identification in background noise. *J Neurosci* 34: 12145–12154.
18. de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE (2017) The hierarchical cortical organization of human speech processing. *J Neurosci* 37:6539–6557.
19. Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322:970–973.
20. Ley A, Vroomen J, Formisano E (2014) How learning to abstract shapes neural sound representations. *Front Neurosci* 8:132.
21. Pasley BN, Knight RT (2013) Decoding speech for understanding and treating aphasia. *Prog Brain Res* 207:435–456.
22. Chevillet MA, Jiang X, Rauschecker JP, Riesenhuber M (2013) Automatic phoneme category selectivity in the dorsal auditory stream. *J Neurosci* 33:5208–5215.
23. Mirman D, et al. (2015) Neural organization of spoken language revealed by lesion-symptom mapping. *Nat Commun* 6:6762.
24. Tsunada J, Liu ASK, Gold JI, Cohen YE (2016) Causal contribution of primate auditory cortex to auditory perceptual decision-making. *Nat Neurosci* 19:135–142.
25. Rauschecker JP (1998) Cortical processing of complex sounds. *Curr Opin Neurobiol* 8: 516–521.
26. Martin S, et al. (2016) Word pair classification during imagined speech using direct brain recordings. *Sci Rep* 6:25803.
27. Santoro R, et al. (2017) Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc Natl Acad Sci USA* 114:4799–4804.
28. Zatorre RJ, Belin P (2001) Spectral and temporal processing in human auditory cortex. *Cereb Cortex* 11:946–953.
29. Fontolan L, Morillon B, Liegeois-Chauvel C, Giraud A-L (2014) The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat Commun* 5:4694.
30. Arnal LH, Giraud A-L (2012) Cortical oscillations and sensory predictions. *Trends Cogn Sci* 16:390–398.
31. Fries P (2015) Rhythms for cognition: Communication through coherence. *Neuron* 88: 220–235.
32. Friston KJ, Trujillo-Barreto N, Daunizeau J (2008) DEM: A variational treatment of dynamic systems. *Neuroimage* 41:849–885.
33. Di Liberto GM, Lalor EC, Millman RE (2018) Causal cortical dynamics of a predictive enhancement of speech intelligibility. *Neuroimage* 166:247–258.
34. Morillon B, Baillet S (2017) Motor origin of temporal predictions in auditory attention. *Proc Natl Acad Sci USA* 114:E8913–E8921.
35. Schoffelen JM, et al. (2017) Frequency-specific directed interactions in the human brain network for language. *Proc Natl Acad Sci USA* 114:8083–8088.
36. Donoso M, Collins AGE, Koechlin E (2014) Human cognition. Foundations of human reasoning in the prefrontal cortex. *Science* 344:1481–1486.
37. Blumstein SE, Myers EB, Rissman J (2005) The perception of voice onset time: An fMRI investigation of phonetic category structure. *J Cogn Neurosci* 17:1353–1366.
38. Rogers JC, Davis MH (2017) Inferior frontal cortex contributions to the recognition of spoken words and their constituent speech sounds. *J Cogn Neurosci* 29:919–936.
39. Marti S, King J-R, Dehaene S (2015) Time-resolved decoding of two processing chains during dual-task interference. *Neuron* 88:1297–1307.
40. King J-R, Dehaene S (2014) Characterizing the dynamics of mental representations: The temporal generalization method. *Trends Cogn Sci* 18:203–210.
41. Isik L, Meyers EM, Leibo JZ, Poggio T (2014) The dynamics of invariant object recognition in the human visual system. *J Neurophysiol* 111:91–102.
42. Du Y, Buchsbaum BR, Grady CL, Alain C (2014) Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proc Natl Acad Sci USA* 111:7126–7131.
43. Liebenthal E, Sabri M, Beardsley SA, Mangalathu-Arumana J, Desai A (2013) Neural dynamics of phonological processing in the dorsal auditory stream. *J Neurosci* 33: 15414–15424.
44. Obleser J, Eisner F, Kotz SA (2008) Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *J Neurosci* 28:8116–8123.
45. Liégeois-Chauvel C, Giraud K, Badier J-M, Marquis P, Chauvel P (2001) Intracerebral evoked potentials in pitch perception reveal a functional asymmetry of the human auditory cortex. *Ann N Y Acad Sci* 930:117–132.
46. Alain C, Snyder JS (2008) Age-related differences in auditory evoked responses during rapid perceptual learning. *Clin Neurophysiol* 119:356–366.
47. Alain C, Snyder JS, He Y, Reinke KS (2007) Changes in auditory cortex parallel rapid perceptual learning. *Cereb Cortex* 17:1074–1084.
48. Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD (2004) Neural correlates of sensory and decision processes in auditory object identification. *Nat Neurosci* 7:295–301.
49. Moerel M, De Martino F, Formisano E (2014) An anatomical and functional topography of human auditory cortical areas. *Front Neurosci* 8:225.
50. Giraud A-L, et al. (2004) Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cereb Cortex* 14: 247–255.
51. Lee Y-S, Turkeltaub P, Granger R, Raizada RDS (2012) Categorical speech processing in Broca's area: An fMRI study using multivariate pattern-based analysis. *J Neurosci* 32: 3942–3948.
52. Arsenault JS, Buchsbaum BR (2015) Distributed neural representations of phonological features during speech perception. *J Neurosci* 35:634–642.
53. Blank H, Davis MH (2016) Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biol* 14:e1002577.
54. Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E (2014) Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *J Neurosci* 34:4548–4557.
55. Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. *J Neurosci* 32:8443–8453.
56. Kok P, Jehee JFM, de Lange FP (2012) Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron* 75:265–270.
57. Weichwald S, et al. (2015) Causal interpretation rules for encoding and decoding models in neuroimaging. *Neuroimage* 110:48–59.
58. Zohary E, Celebrini S, Britten KH, Newsome WT (1994) Neuronal plasticity that underlies improvement in perceptual performance. *Science* 263:1289–1292.
59. Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. *Phys Rev E Stat Nonlin Soft Matter Phys* 64:051904.
60. Pitkow X, Liu S, Angelaki DE, DeAngelis GC, Pouget A (2015) How can single sensory neurons predict behavior? *Neuron* 87:411–423.
61. Çukur T, Nishimoto S, Huth AG, Gallant JL (2013) Attention during natural vision warps semantic representation across the human brain. *Nat Neurosci* 16:763–770.
62. Cheung C, Hamiton LS, Johnson K, Chang EF (2016) The auditory representation of speech sounds in human motor cortex. *Elife* 5:e12577.
63. Correia J, et al. (2014) Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *J Neurosci* 34:332–338.
64. Kroon P, Kleijn WB (1995) *Linear-Prediction Based Analysis-by-Synthesis Coding* (Elsevier Science, Amsterdam).
65. Riede T, Suthers RA (2009) Vocal tract motor patterns and resonance during constant frequency song: The white-throated sparrow. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 195:183–192.
66. Goncharoff V, Kaine-Krolak M (1995) Interpolation of LPC spectra via pole shifting. *Proc IEEE Int Conf Acoust Speech Signal Process* 1:780–783.
67. Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM (2011) Brainstorm: A user-friendly application for MEG/EEG analysis. *Comput Intell Neurosci* 2011:879716.
68. Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:156869.
69. Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA (2005) Neural substrates of phonemic perception. *Cereb Cortex* 15:1621–1631.
70. Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878.
71. Gläscher J (2009) Visualization of group inference data in functional neuroimaging. *Neuroinformatics* 7:73–82.
72. Hämäläinen MS, Ilmoniemi RJ (1994) Interpreting magnetic fields of the brain: Minimum norm estimates. *Med Biol Eng Comput* 32:35–42.
73. Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53:1–15.
74. Dhamala M, Rangarajan G, Ding M (2008) Analyzing information flow in brain networks with nonparametric Granger causality. *Neuroimage* 41:354–362.
75. Dhamala M, Rangarajan G, Ding M (2008) Estimating Granger causality from fourier and wavelet transforms of time series data. *Phys Rev Lett* 100:018701.
76. Lyu B, Ge J, Niu Z, Tan LH, Gao J-H (2016) Predictive brain mechanisms in sound-to-meaning mapping during speech processing. *J Neurosci* 36:10813–10822.
77. Yoo S, Lee K-M (2013) Articulation-based sound perception in verbal repetition: A functional NIRS study. *Front Hum Neurosci* 7:540.
78. Klein M, et al. (2015) Early activity in Broca's area during reading reflects fast access to articulatory codes from print. *Cereb Cortex* 25:1715–1723.
79. Meyers EM (2013) The neural decoding toolbox. *Front Neuroinform* 7:8.
80. Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* 164:177–190.