Behavioral/Systems/Cognitive

Auditory Cortex Encodes the Perceptual Interpretation of Ambiguous Sound

Niclas Kilian-Hütten,[1,2] Giancarlo Valente,[1,2] Jean Vroomen,[3] and Elia Formisano[1,2]

[1]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, and [2]Maastricht Brain Imaging Center, Maastricht University, 6200 MD Maastricht, The Netherlands, and [3]Department of Psychology, Tilburg University, 5000 LE Tilburg, The Netherlands

The confounding of physical stimulus characteristics and perceptual interpretations of stimuli poses a problem for most neuroscientific studies of perception. In the auditory domain, this pertains to the entanglement of acoustics and percept. Traditionally, most study designs have relied on cognitive subtraction logic, which demands the use of one or more comparisons between stimulus types. This does not allow for a differentiation between effects due to acoustic differences (i.e., sensation) and those due to conscious perception. To overcome this problem, we used functional magnetic resonance imaging (fMRI) in humans and pattern-recognition analysis to identify activation patterns that encode the perceptual interpretation of physically identical, ambiguous sounds. We show that it is possible to retrieve the perceptual interpretation of ambiguous phonemes—information that is fully subjective to the listener—from fMRI measurements of brain activity in auditory areas in the superior temporal cortex, most prominently on the posterior bank of the left Heschl's gyrus and sulcus and in the adjoining left planum temporale. These findings suggest that, beyond the basic acoustic analysis of sounds, constructive perceptual processes take place in these relatively early cortical auditory networks. This disagrees with hierarchical models of auditory processing, which generally conceive of these areas as sets of feature detectors, whose task is restricted to the analysis of physical characteristics and the structure of sounds.

Introduction

Investigations of perceptual processing are usually faced with a confounding problem when attempting to separate out subjective perception from the processing of stimulus-specific characteristics. This is because, usually, differential percepts follow differential physical stimulus characteristics, i.e., distinct stimuli elicit distinct percepts. The traditional reliance on cognitive subtraction logic, which demands the use of one or more comparisons between stimulus types, renders dissociating sensation from conscious perception impossible.

In the visual domain, this problem has been tackled using ambiguous stimuli and the phenomenon of multistable perception (for review, see Sterzer et al., 2009). In the auditory domain, however, the creation of stimuli and designs that permit comparing differential perceptual states while keeping the physical input constant seems a bigger challenge. In previous studies, temporal phenomena, such as auditory streaming (Cusack, 2005; Gutschalk et al., 2005) or illusory continuity (Riecke et al., 2009) have been investigated similarly in human subjects. However, to date, no attempt has been made to employ ecologically valid, ambiguous auditory stimuli to investigate the neural basis of differential perceptual interpretations of an auditory stimulus's identity.

Here, we use functional magnetic resonance imaging (fMRI) and pattern-recognition analysis to identify activation patterns that encode the perceptual interpretation of physically identical, ambiguous phonemes. We adhere to a principle based on the McGurk effect, called cross-modal recalibration (Bertelson et al., 2003), in which lip movements are used to disambiguate ambiguous auditory phonemes. Repeated presentation of these videos increases the proportion of corresponding responses in subsequent audio-only forced-choice trials, thus eliciting an aftereffect. This enables us to compare physically identical, yet differentially perceived, sounds, and thus allows for the investigation of a purely perceptual distinction of stimulus identity.

According to popular models of auditory processing, representations become more abstract with hierarchical distance to the primary auditory cortex (A1) along two (what/where) pathways (Scott and Johnsrude, 2003; Liebenthal et al., 2005; Rauschecker and Scott, 2009). In humans, the regions adjacent to the Heschl's gyrus, which we refer to as early auditory cortex, are supposedly restricted to the analysis of physical features and the acoustic structure of sounds. In contrast to this notion, fMRI and pattern-recognition techniques (Haxby et al., 2001; Haynes and Rees, 2005) have recently been used to demonstrate the existence of vowel representations in these regions that were invariant of the specific speaker uttering them (Formisano et al., 2008). These findings suggest the presence of abstract perceptual sound representations in early auditory areas. The same vowels, however, even when uttered by distinct speakers, still share an abundance of acoustic similarities. Whether it is these acoustic features or the

Received Sept. 1, 2010; revised Oct. 22, 2010; accepted Nov. 17, 2010.

This work was supported by Maastricht University and the Netherlands Organization for Scientific Research, Innovational Research Incentives Scheme Vidi Grant 452-04-337 (E.F.), and TopTalent Grant 021-001-077 (N.K.H.). We thank R. Goebel for valuable discussions and F. De Martino for help with data analysis.

Correspondence should be addressed to Niclas Kilian-Hütten, Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands. E-mail: niclas.kilian-hutten@maastrichtuniversity.nl.

DOI:10.1523/JNEUROSCI.4572-10.2011

 0270-6474/11/311715-06$15.00/0

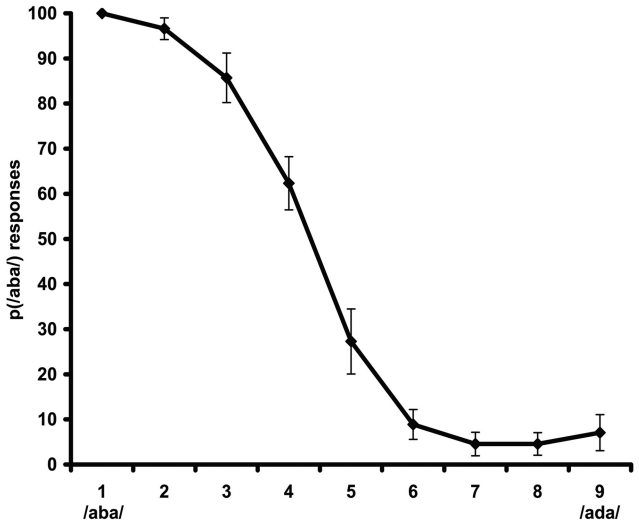

Figure 1. Results of the auditory pretest. The mean proportions (p) of /aba/ classifications across the 11 participants for each stimulus in the nine-sound continuum are given. Sound 4 was chosen as A? for eight of the participants and sound 5 for the remaining three.

perceptual categorization of the vowels that forms the basis of classification cannot be conclusively resolved by this study. The use of acoustically identical stimuli, which are perceptually categorized as distinct auditory entities, seems to be the only way to overcome this problem. Here, we trained our pattern classification algorithm on purely perceptual labels and restricted the information provided to it to voxels within the temporal lobe to investigate the role of the auditory cortex in constructive perceptual processing.

Materials and Methods

Participants. Twelve healthy native Dutch students of the University Maastricht (five males; mean age, 24.83 years) were recruited to participate in the study. One participant was left-handed. None of the participants had a history of hearing loss or neurological abnormalities. Approval for the study was granted by the Ethical Committee of the Faculty of Psychology at the University of Maastricht. One subject was discarded on the basis of a self-reported difficulty in categorizing the phonemes (perceiving them exclusively as /asa/), which was reflected as an extreme response bias in the behavioral data.

Stimuli. The stimulation entailed digital auditory and visual recordings of a male Dutch speaker pronouncing the syllables /aba/ and /ada/. The two auditory stimuli were 640 ms, with 240 ms stop closure. From these, a place-of-articulation continuum was synthesized by means of varying the F2 formant by equal steps of 39 Mel, resulting in nine different stimuli, ranging from a clear /aba/ via seven ambiguous stimuli to a clear /ada/. The audiovisual stimuli were synthesized by pairing the visual recordings of the speaker pronouncing /ada/ and /aba/, respectively, with the most ambiguous auditory stimulus, determined as such in a pretest.

Behavioral pretest. Each participant underwent an auditory pretest outside the fMRI scanner to individually determine the most ambiguous auditory stimulus from the /aba/-to-/ada/ continuum (Bertelson et al., 2003). This pretest consisted of 98 forced-choice judgments on all different stimuli from the continuum, with presentation frequency biased, so that the five central stimuli were presented in 14 trials, the second and tenth stimuli were presented in eight trials, and the first and the eleventh stimuli were presented in six trials. Stimuli were presented binaurally through loudspeakers. Participants were required to press one of two buttons if they had perceived /aba/ and the other one if they perceived /ada/. This resulted in an estimation of each participant's ambiguous auditory token (A?), which was used for the rest of the session (Fig. 1).

Experimental procedure. The experimental procedure during scanning was based on the phenomenon of cross-modal recalibration, a McGurk aftereffect (Bertelson et al., 2003). In a typical McGurk paradigm, an auditorily presented disyllable (/aba/) is paired with an incongruent visual disyllable (/aga/), pronounced by a speaker. The addition of this incongruent visual input changes the listener's auditory percept into an intermediate one (/ada/). Interestingly, when an ambiguous auditory component (A?, between /aba/ and /ada/) is used, exposure to this ambiguous stimulus dubbed onto a video of a face pronouncing /aba/ or /ada/ selectively increases the proportion of corresponding responses in subsequent audio-only forced-choice trials. This makes it possible to contrast conditions where the physical stimulus is identical, whereas the perceptual interpretation differs by comparing brain activation patterns in response to auditory posttest stimuli perceived as /aba/ versus /ada/. This logic, thus, allows for the investigation of a purely perceptual distinction of stimulus identity.

The design entailed two major elements (Fig. 2): blocks of multimodal exposure (the recalibration phase) and slow, event-related auditory posttests. In the recalibration phase, videos were presented, consisting of the individually determined ambiguous stimulus A?, dubbed on the visual recording of the speaker pronouncing either /aba/ (Vb) or /ada/ (Vd). Bimodal stimuli were presented in blocks of eight identical trials (block A?Vb or block A?Vd), with an interstimulus interval (ISI) of one repetition time (TR; 2000 ms). In each of the two runs, five A?Vb and five A?Vd blocks were run in randomized order (160 trials total). During these exposure trials, participants were required to press a button whenever a small white spot (12 pixels) appeared on the speaker's upper lip to ensure participants focused their attention on the speaker's lips. This occurred once per block at a random position.

Each block of bimodal exposure was followed by six auditory posttests (120 trials total), which, like the pretests, consisted of forced-choice /aba/–/ada/ judgments. Unlike the pretests, however, here only the A? token and the two tokens closest to it on the continuum were presented, twice each. Due to the use of a slow event-related design, the jittered ISI was six TR (12 s) on average (Fig. 2).

Scanning parameters. Functional MRI data were collected on a 3-tesla fMRI scanner (head set-up, Siemens) at the Maastricht Brain Imaging Center in Maastricht, The Netherlands. For each participant, two functional runs of 665 volumes were acquired. For later overlay, a high-resolution structural scan (voxel size, $1 \times 1 \times 1$ mm^3) was collected using a T1-weighted three-dimensional (3D) ADNI sequence [TR, 2050 ms; echo time (TE), 2.6 ms; 192 sagittal slices]. Both functional runs and the structural scan were acquired in a single session for each participant. For functional images, a blood oxygenation level-dependent (BOLD)-sensitive echo-planar imaging (EPI) sequence was used (matrix, 64×64, 27 slices; slice thickness, 3 mm; field of view, 192×192 mm^3, resulting voxel size, $3 \times 3 \times 3$ mm^3; TE/acquisition time slice, 30/55.5, flip angle, 90°). Volume acquisition was clustered in the beginning of each TR, leaving a silent delay within each TR during which stimuli were presented in the absence of EPI noise. This was done to optimize stimulus audibility, an approach which has been shown to be highly efficient in auditory fMRI paradigms (Jäncke et al., 2002; van Atteveldt et al., 2007). Hence, the effective TR was 2000 ms, including 1500 ms of sequence scanning time and a 500 ms silent delay. Stimuli were presented and synchronized with the MR pulses using the software package Presentation (Neurobehavioral Systems).

Data preprocessing. Functional and anatomical images were analyzed using BrainVoyager QX (Brain Innovation) and customized code written in MATLAB (MathWorks). Several preprocessing steps were performed: sinc-interpolated slice-time correction, 3D-motion correction to correct for common small head movements by spatially aligning all volumes to the first volume by rigid body transformations, linear trend removal, and temporal high-pass filtering to remove low-frequency nonlinear drifts of seven or less cycles per time course. Functional slices were then coregistered to the structural volume on the basis of positioning parameters from the scanner and manual adjustments to ensure optimal fit. Subsequently, they were transformed into Talairach space. All individual brains were segmented at the gray/white matter boundary using a semi-automatic procedure based on intensity values implemented in Brain-

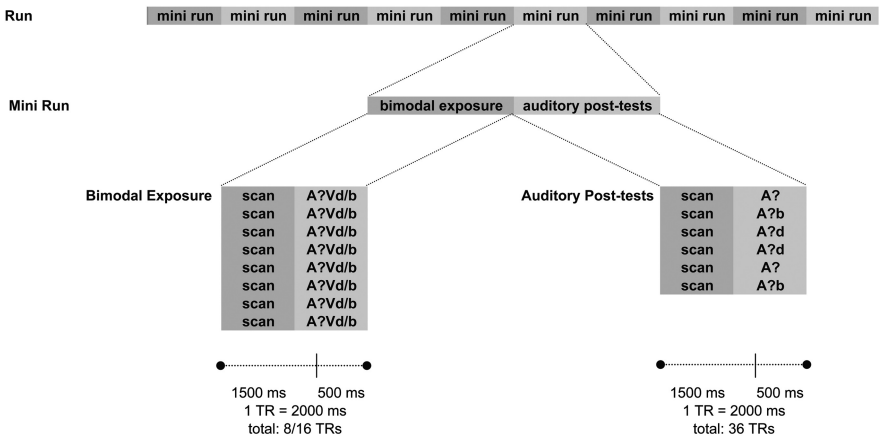

Figure 2. Schematic overview of the experimental procedure. Each run consisted of 10 mini-runs, which were each composed of a bimodal exposure block (A?Vd or A?Vb) and six auditory posttest trials. The bimodal blocks were presented within eight TRs, each of which entailed 1500 ms of scanning and 500 ms of video presentation. Between auditory posttest trials, fixation periods averaged six TRs.

Voyager QX. For all group analyses, cortex-based alignment was used to assure optimized spatial matching of cortical locations (i.e., vertices) between participants (Goebel et al., 2006).

Univariate data analysis. Functional runs were analyzed using voxel-wise multiple linear regression [general linear model (GLM)] of the BOLD-response time course. All analyses were performed at single-subject and group levels and all experimental conditions were modeled as predictors. Because the number of these depended on the individual participant's behavior (i.e., if a participant never answered incorrectly to A?Vd stimuli after an A?Vd block, this condition had zero events) and, thus, varied across runs and participants, empty predictors had to be included in all multirun and multisubject GLMs. In all GLMs, predictor time courses were convolved with a hemodynamic response function (where both the BOLD response and undershoot are modeled by a gamma function) to adjust for the hemodynamic delay (Friston et al., 1998).

Multivoxel pattern analysis procedure. We decoded multivariate patterns of BOLD activation training linear support vector machine (SVM) classifiers based solely on perceptual labels (i.e., trials perceived as /ada/ or /aba/) and classified individual posttest auditory trials accordingly. We tested our hypothesis that abstract auditory representations can be found in early auditory cortex by confining classification anatomically to areas in the temporal lobe, excluding information from other areas (Fig. 3*b*). An additional reason for this confinement is rooted in the necessity for the subjects to provide motor responses. The inclusion of motor areas would arguably increase classification performance. However, since this improvement might be based on the differentiation between motor responses corresponding to the two responding fingers instead of two percepts, it would not be informative with respect to our research question.

We estimated the response of every voxel in each trial by fitting a general linear model with two predictors, one accounting for trial response and the other accounting for the trial mean. The hemodynamic response function was optimized on a subject basis. The pattern of activation associated with a given trial consisted of the set of beta values associated with the first predictor for all the voxels considered in the analysis. The steps of the feature extraction procedure are similar to those described and validated in De Martino et al. (2008).

The ISI in the experiment presentation ranged from four to eight 8 TRs (8 to 16 s), and we removed the shortest trials (i.e., those that were followed by another posttest trial after four functional volumes) from the analysis to use a larger time window to estimate betas on the remaining trials. We retained a total of 100 trials, equally distributed among the 20 recalibration mini-blocks and two runs, from the original 120. The trials were divided into two disjoint groups (training and testing). As the two classes were generally unbalanced, we created several balanced training datasets by randomly selecting n trials from each class (with n being 80% of the total number of examples of the least represented class), leaving the remaining trials as testing dataset. We trained linear SVMs (Vapnik, 1995) on the training datasets and evaluated the generalization to new data on the test dataset. This procedure was repeated 100 times with different random trial selections.

Voxels that entered the multivariate analysis were selected using a univariate preselection by means of general linear models on the training data. We considered the design matrix obtained considering the auditory posttest used in the training data and ranked the voxels according to their activity in the two auditory posttest conditions (separately). We considered the union of the most active voxels per condition (m). The value of m was optimized for each subject, in a range from 100 to 2000, with steps of 100.

We report accuracies for the classification of unseen trials. Since we used balanced subsets for training, we expected a chance level of 50% per class. However, total accuracy (on both classes) may be influenced by the unbalanced nature of the dataset. We therefore assessed the statistical significance of our results by means of a permutation test (Nichols and Holmes, 2002) in which we performed the classification procedure after scrambling the class labels of the auditory posttest trials. To avoid bias from the voxel preselection by means of GLM, we first scrambled the auditory posttest behavioral labels, and subsequently selected the voxels on which we performed the analyses. This procedure was repeated 500 times per subject, providing an empirical distribution for the null hypotheses that the obtained accuracies derive from chance. Furthermore, we considered precision and recall as additional measures of performance of the classifier. Both these measures are widely used in assessing performances of classifiers on unbalanced data and they are insensitive of relative group size. Precision for a class is defined as the ratio between true positives and the total number of elements classified as belonging to that class. Recall for a class is defined as the ratio between true positives and the total number of elements actually belonging to that class. It can be shown that the average of both recall and precision have a chance level of 50%, regardless of differences in class sizes (Rieger et al., 2008).

To visualize the spatial activation patterns that were used for classification and to inspect consistency across participants, group discriminative maps, i.e., maps of the cortical locations that contribute most to the discrimination of conditions, were created after cortex-based alignment (Goebel et al., 2006) of binarized single-subject discriminative maps (Staeren et al., 2009). As we used a linear support vector machine, it is meaningful, at an individual map level, to rank the features (i.e., voxels) relatively according to their contribution to the discrimination and select the peaks through thresholding. Therefore, for binarizing the maps, we selected the 30% threshold level empirically and assessed the spatial (anatomical) correspondence between the peaks of the single-subject discriminative maps. In the resulting group-level discriminative maps, a cortical location (vertex) was color-coded if it was present among the 30% of most discriminative vertices in the corresponding individual discriminative maps of at least seven of the 11 subjects. To account for the multiple tests performed in creating these maps, we calculated the proportion of expected false positives (false discovery rate) that corresponded to the uncorrected p value. The q value was obtained using a statistical method that ensured robust estimates when the distribution of p values was discrete and one-sided tests were performed (Pounds and Cheng, 2006). The classification accuracy in each subject was always calculated with respect to the whole set of features and, thus, did not depend on the threshold chosen for the creation of the maps.

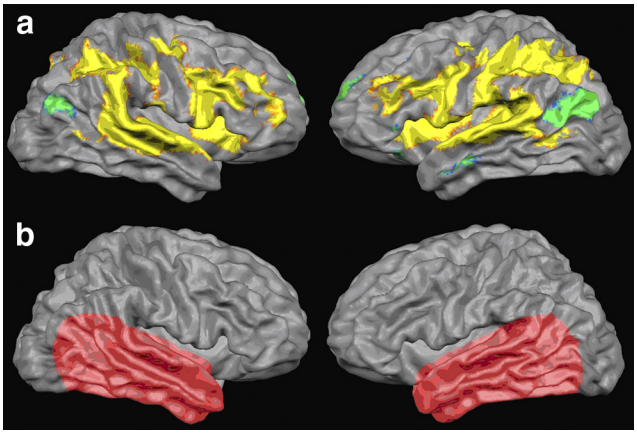

Figure 3. ***a***, Results of statistical univariate analysis. Overall auditory cortical activation in response to the ambiguous auditory stimuli as estimated with univariate random effects general linear model analysis (statistical parametric *F* map, group results of cortex-based aligned datasets, $p < 1.7 \times 10^{-11}$). The pairwise statistical comparison between the two perceptual conditions (/aba/ vs /ada/) did not yield significant results. ***b***, Illustration of the anatomical masks used. Classification was anatomically confined to the temporal lobe. The exact boundaries of the anatomical masks were determined on a subject basis to account for individual differences in anatomy.

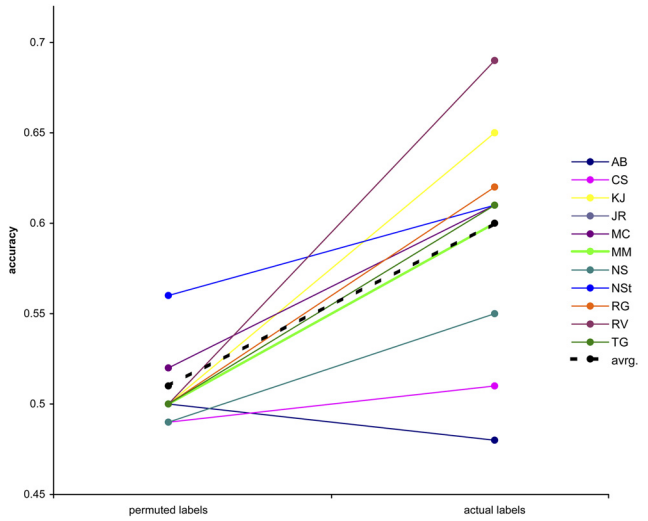

Figure 4. Classification accuracies for individual participants (identified by their initials). Accuracies obtained with the actual labels were significantly higher than the accuracies derived from the analysis using permuted class labels ($p < 5.6 \times 10^{-4}$). avrg, Average.

Results

Behavioral results

Participants were presented with individually determined, ambiguous disyllables (A?, between /aba/ and /ada/). As discussed before, it has been shown that the perceptual interpretation of these ambiguous sounds can be altered relying on cross-modal recalibration (Bertelson et al., 2003). We were able to replicate these behavioral results in an fMRI environment. An ANOVA demonstrated a highly significant difference in response (/aba/ vs /ada/) between sounds that were preceded by an A?Vb block versus an A?Vd block ($F = 75.682$, $p < 0.001$). The overall proportion of /aba/ responses was 0.31 (± 0.02) if preceded by an A?Vd block and 0.53 (± 0.02) if preceded by an A?Vb block. This significant difference was observed for all three sounds individually (A?, 0.25 vs 0.60, $F = 68.692$, $p < 0.001$; A?b, 0.05 vs 0.15, $F = 31.546$, $p < 0.001$; A?d, 0.64 vs 0.85, $F = 12.757$, $p < 0.001$). Thus, exposure to the videos significantly recalibrated later perceptual interpretation in the auditory posttest trials, such that ambiguous stimuli were more often perceived as /aba/ when preceded by an A?Vb block and vice versa.

Univariate statistical analysis

These posttest trials evoked significant fMRI responses in a network of areas consistent with previous studies on phoneme processing (Fig. 3*a*), including auditory areas in the superior temporal cortex (Heschl's gyrus), multiple regions in the planum temporale (PT), superior temporal gyrus and sulcus (STG and STS, respectively), as well as middle temporal gyrus, insula cortex, inferior parietal cortex, inferior frontal cortex, and supramarginal gyrus (Davis and Johnsrude, 2003; Liebenthal et al., 2005; Hickok and Poeppel, 2007; Desai et al., 2008). A univariate comparison between trials perceptually classified as /aba/ and those classified as /ada/ did not yield significant differences in activation ($q = 0.05$, corrected for multiple comparisons with false discovery rate).

Multivariate pattern recognition analysis

This homogeneity in activation between perceived /aba/ and perceived /ada/ is likely due to the fact that these stimuli differed only in perceptual category, but not in physical stimulus identity. To decipher subtle and supposedly distributed differences in activation and to predict the perceptual interpretation on a trial-by-trial basis, we performed multivoxel pattern analysis. As our stimuli were physically identical, the fMRI responses to individual auditory posttest trials were labeled according to their perceptual interpretation (/aba/ vs /ada/). The algorithm succeeded in learning the relationship between perceptual labels and corresponding spatial activation patterns and in exploiting this information to correctly classify the patterns associated with the remaining unlabeled trials. The mean accuracy of the classifier across the 11 subjects was 0.60. Statistical significance of the original classification was assessed by means of permutation testing (Nichols and Holmes, 2002). We performed a paired *t* test on the classification accuracies obtained on the 11 subjects using correct labels and permuted labels, and were able to reject the null hypothesis at $p < 5.6 \times 10^{-4}$ (Fig. 4). It should be noted that in one subject, the mean of the empirical null distribution was different from the theoretical value of 50%. Nonetheless, the results of the permutation tests show that classification accuracies with correct labels are significantly greater than those obtained with scrambled labels, indicating that information on pure perception is present in the selected brain regions and is efficiently exploited by the learning algorithm. A paired *t* test between the values for precision (recall) obtained on the 11 subjects and the ones obtained from the permutation tests confirmed these results, rejecting the null hypothesis at $p < 2.4 \times 10^{-3}$.

Group discriminative maps were inspected for consistency of spatial activation patterns across participants. The main patterns of discriminative voxels revealed by the group discriminative maps (threshold corresponds to $q = 2.75\ 10^{-3}$) (Fig. 5) were left-lateralized and clustered along the posterior bank of Heschl's gyrus, Heschl's sulcus, and, adjacently, in the anterior portion of PT. The exact localization of auditory cortical fields in humans is not a trivial task, since the discrepancies between cytoarchitectonic borders and macroscopical landmarks are profound and vary from brain to brain and between hemispheres (Morosan et al., 2001). Comparing our data with data from tonotopy mapping experiments (Formisano et al., 2003), it seems likely that our discrimination maps touch upon A1 and spread across posterior–

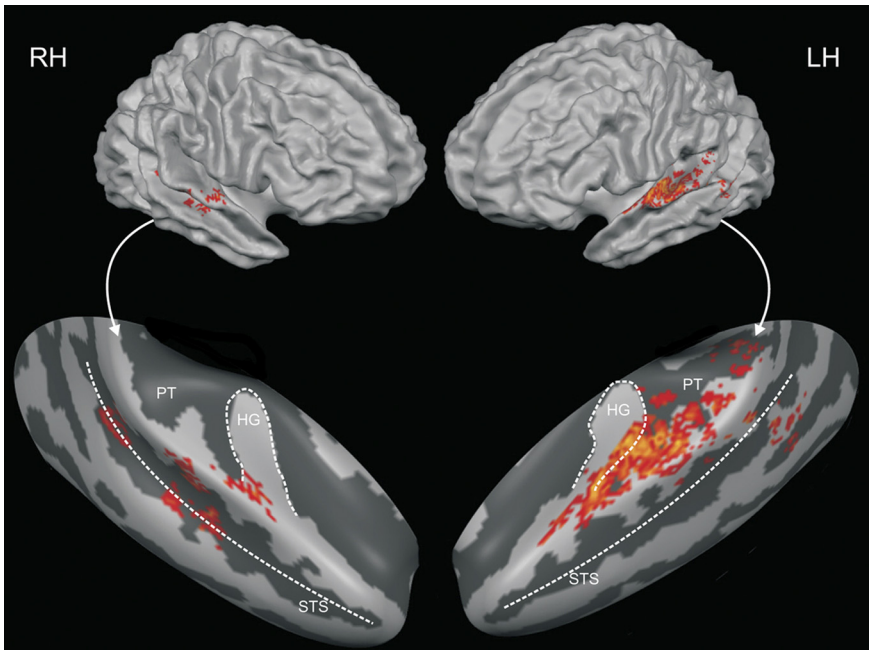

Figure 5. Discriminative map. Group map of the 30% of active voxels most discriminative for the purely perceptual difference between /aba/ and /ada/ (cortex-based aligned, smoothed). A location was color-coded if it was present on the individual maps of at least seven of the 11 subjects. This corresponds to a false discovery rate-corrected threshold of $q = 2.75 \times 10^{-3}$. Maps are overlaid on the reconstructions of the average hemispheres of the 11 subjects (top) and on inflated reconstructions of the right and left temporal lobes of these average hemispheres (bottom). RH, Right hemisphere; LH, left hemisphere; HG, Heschl's gyrus.

lateral human belt regions within area PT. However, *in vivo* neuroimaging experiments in humans currently do not allow the exact localization of maps in terms of cytoarchitectonic areas and/or single-cell response patterns. Said tonotopy mapping experiments, as well as probabilistic cytoarchitectonic maps (Morosan et al., 2001; Rademacher et al., 2001), suggest A1 itself to be centered mostly on the convexity of medial Heschl's gyrus, an area that is mostly absent from our group discriminative maps. Definite assertions remain problematic due to the aforementioned intersubject variability in macro-anatomical morphology, cytoarchitectonic make-up, and the relation between the two, as well as the differences in normalization. However, it seems likely that our maps identify mostly nonprimary auditory areas. Additional clusters of smaller extent were found at the left temporoparietal junction and, bilaterally, on middle STG and STS.

Discussion

In the present study, we used cross-modal recalibration to influence participants' perception of ambiguous speech sounds. We then trained an SVM pattern classifier on fMRI data, which was categorized using purely subjective perceptual labels. Our results show that pure perceptual interpretation of physically identical phonemes can be decoded from cortical activation patterns in early auditory areas. More specifically, our findings provide direct empirical evidence that, beyond the basic acoustic analysis of sounds, constructive perceptual information is present in regions within the anterior PT, tangent to the posterior bank of Heschl's gyrus and sulcus.

Concerning speech perception, hierarchical views of auditory processing (Rauschecker and Scott, 2009) suggest a gradient of increasing processing complexity in the anterior superior temporal gyrus (i.e., the what stream), where regions show the first clear responses to abstract, linguistic information in speech. Further, it has been suggested that "phonetic maps have some anatomical implementation in anterior temporal lobe areas" (Rauschecker and Scott, 2009). Hierarchically lower auditory regions, in contrast, are allegedly limited to low-level acoustic feature analysis. Our finding that regions in the PT adjacent to, and touching upon, Heschl's gyrus and sulcus discriminate stimuli on a purely perceptual level goes beyond such a limited feature-bound processing role. It is, however, in line with single-cell recording studies in monkeys and cats (Micheyl et al., 2005, 2007; Nelken and Bar-Yosef, 2008). For instance, object representations have been suggested to be present even as early as in A1 (Bar-Yosef and Nelken, 2007). Furthermore, Micheyl and colleagues (2005) demonstrated a strong correspondence between psychophysical findings on auditory stream formation in humans and single-unit responses in rhesus monkeys' A1, suggesting the possibility that auditory streaming percepts have a representation already in A1 in the absence of stimulus differences. However, comparing human behavior with primate neuronal responses is obviously problematic. Differences in behavior, perception, and neuroanatomy are unavoidable. Our results provide direct empirical evidence that, in humans, processing in early auditory cortex, probably corresponding to human belt areas, is not limited to low-level stimulus feature analysis.

Although across-study comparisons of the roles of human auditory areas are problematic due to differences in normalization techniques, analysis methods, and even anatomical nomenclature, our findings do seem to be in line with some previous works. Macroscopically, area PT has been described as a computational hub, which segregates spectrotemporal patterns, compares them to stored patterns, and outputs auditory objects (Griffiths and Warren, 2002), which is compatible with our results. Location and functional role of a smaller part of our clusters in PT is accordant with Spt, a functional subdivision of PT, which has been described as a sensory-motor integration region for the vocal tract motor effector (Hickok and Poeppel, 2007; Hickok et al., 2009). This is in line with the suggestion that PT is part of the dorsal stream of auditory processing (revised to include language in addition to spatial functions) and possesses the representation of templates for doable articulations, effectively disambiguating phonological information (Rauschecker and Scott, 2009). In the present experiment, this sensory-motor function of the posterior PT may serve to disambiguate ambiguous phonemes on the basis of previously seen lip movements. A functional role in cross-modal integration has been suggested before for similar regions (van Atteveldt et al., 2004). The output of an acoustic–phonetic analysis in PT may be probabilistic and represent prelexical phonemic categories, as has been suggested previously (Obleser and Eisner, 2009).

In terms of neurocomputational plausibility, our data may be accordant with reverse hierarchy theory, which states that, by default, rapid perception is based on high-level (e.g., phonological) representations alone, which are holistic and ecologically meaningful (Hochstein and Ahissar, 2002; Ahissar et al., 2009).

However, when finer discrimination between similar stimuli is needed, performance relies more on lower level activity (reverse hierarchy). This concept seems to provide an interesting interpretation of our findings. Since the task to categorize ambiguous phonemes demands scrutiny, reverse hierarchy routines might recruit low-level networks. One problem here is that, since the stimuli are identical in both perceptual conditions, there are no low-level physical features that would help discrimination. In a way, the holistic, high-level representation is the only level that provides valuable information for this task. We propose that what is reflected in the reliance on early networks in this case is not the intensified analysis of low-level stimulus features, but a perceptual bias that is stored in these regions. This bias is responsible for the behavioral (perceptual) effect and is installed by the cross-modal recalibration mechanism. Its origin may lie within higher-order areas involved in the integration of audiovisual speech signals, such as the supramarginal gyrus (Raizada and Poldrack, 2007), the intraparietal sulcus (Cusack, 2005), or even regions involved in the motor processing of speech (D'Ausilio et al., 2009). The information is then fed back from there to early auditory areas. Here, without eliciting overall differences in activation strength, the perceptual bias is stored and sensory input is transformed into more abstract entities or auditory objects. These abstract entities may be considered as the building blocks of further linguistic and vocal processing.

References

Ahissar M, Nahum M, Nelken I, Hochstein S (2009) Reverse hierarchies and sensory learning. Philos Trans R Soc Lond B Biol Sci 364:285–299.

Bar-Yosef O, Nelken I (2007) The effects of background noise on the neural responses to natural sounds in cat primary auditory cortex. Front Comput Neurosci 1:3.

Bertelson P, Vroomen J, De Gelder B (2003) Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychol Sci 14:592–597.

Cusack R (2005) The intraparietal sulcus and perceptual organization. J Cogn Neurosci 17:641–651.

D'Ausilio A, Pulvermüller F, Salmas P, Bufalari I, Begliomini C, Fadiga L (2009) The motor somatotopy of speech perception. Curr Biol 19:381–385.

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423–3431.

De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E (2008) Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage 43:44–58.

Desai R, Liebenthal E, Waldron E, Binder JR (2008) Left posterior temporal regions are sensitive to auditory categorization. J Cogn Neurosci 20:1174–1188.

Formisano E, Kim DS, Di Salle F, van de Moortele PF, Ugurbil K, Goebel R (2003) Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40:859–869.

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.

Friston KJ, Fletcher P, Josephs O, Holmes A, Rugg MD, Turner R (1998) Event-related fMRI: characterizing differential responses. Neuroimage 7:30–40.

Goebel R, Esposito F, Formisano E (2006) 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.

Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. Trends Neurosci 25:348–353.

Gutschalk A, Micheyl C, Melcher JR, Rupp A, Scherg M, Oxenham AJ (2005) Neuromagnetic correlates of streaming in human auditory cortex. J Neurosci 25:5382–5388.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430.

Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 8:686–691.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Hickok G, Okada K, Serences JT (2009) Area Spt in the human planum temporale supports sensory-motor integration for speech processing. J Neurophysiol 101:2725–2732.

Hochstein S, Ahissar M (2002) View from the top: hierarchies and reverse hierarchies in the visual system. Neuron 36:791–804.

Jäncke L, Wüstenberg T, Scheich H, Heinze HJ (2002) Phonetic perception and the temporal cortex. Neuroimage 15:733–746.

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA (2005) Neural substrates of phonemic perception. Cereb Cortex 15:1621–1631.

Micheyl C, Tian B, Carlyon RP, Rauschecker JP (2005) Perceptual organization of tone sequences in the auditory cortex of awake macaques. Neuron 48:139–148.

Micheyl C, Carlyon RP, Gutschalk A, Melcher JR, Oxenham AJ, Rauschecker JP, Tian B, Courtenay Wilson E (2007) The role of auditory cortex in the formation of auditory streams. Hear Res 229:116–131.

Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K (2001) Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13:684–701.

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

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13:14–19.

Pounds S, Cheng C (2006) Robust estimation of the false discovery rate. Bioinformatics 22:1979–1987.

Rademacher J, Morosan P, Schormann T, Schleicher A, Werner C, Freund HJ, Zilles K (2001) Probabilistic mapping and volume measurement of human primary auditory cortex. Neuroimage 13:669–683.

Raizada RD, Poldrack RA (2007) Selective amplification of stimulus differences during categorical processing of speech. Neuron 56:726–740.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718–724.

Riecke L, Esposito F, Bonte M, Formisano E (2009) Hearing illusory sounds in noise: the timing of sensory-perceptual transformations in auditory cortex. Neuron 64:550–561.

Rieger JW, Reichert C, Gegenfurtner KR, Noesselt T, Braun C, Heinze HJ, Kruse R, Hinrichs H (2008) Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. Neuroimage 42:1056–1068.

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends Neurosci 26:100–107.

Staeren N, Renvall H, De Martino F, Goebel R, Formisano E (2009) Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol 19:498–502.

Sterzer P, Kleinschmidt A, Rees G (2009) The neural bases of multistable perception. Trends Cogn Sci 13:310–318.

van Atteveldt N, Formisano E, Goebel R, Blomert L (2004) Integration of letters and speech sounds in the human brain. Neuron 43:271–282.

van Atteveldt NM, Formisano E, Blomert L, Goebel R (2007) The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. Cereb Cortex 17:962–974.

Vapnik VN (1995) The nature of statistical learning theory. New York: Springer.