Left Posterior Temporal Regions are Sensitive to Auditory Categorization

Rutvik Desai, Einat Liebenthal, Eric Waldron, and Jeffrey R. Binder

Abstract

■ Recent studies suggest that the left superior temporal gyrus and sulcus (LSTG/S) play a role in speech perception, although the precise function of these areas remains unclear. Here, we test the hypothesis that regions in the LSTG/S play a role in the categorization of speech phonemes, irrespective of the acoustic properties of the sounds and prior experience of the listener with them. We examined changes in functional magnetic resonance imaging brain activation related to a perceptual shift from nonphonetic to phonetic analysis of sine-wave speech analogs. Subjects performed an identification task before scanning and a discrimination task during scanning with phonetic (P) and nonphonetic (N) sine-wave sounds, both before (Pre) and after (Post) being exposed to the phonetic properties of the P sounds. Behaviorally, experience with the P sounds induced categorical identification of these sounds. In the PostP > PreP and PostP > PostN contrasts, an area in the posterior LSTG/S was activated. For both P and N sounds, the activation in this region was correlated with the degree of categorical identification in individual subjects. The results suggest that these areas in the posterior LSTG/S are sensitive neither to the acoustic properties of speech nor merely to the presence of phonetic information, but rather to the listener's awareness of category representations for auditory inputs. ■

INTRODUCTION

Speech perception is shaped by the biological significance of speech in human cognition, the complex spectro-temporal structure of human vocalizations, and the categorical nature of phoneme representations. A common neuroimaging paradigm for studying the neural substrates mediating speech perception is to compare brain activation patterns elicited during the processing of speech and nonspeech sounds (Obleser et al., 2006; Uppenkamp, Johnsrude, Norris, Marslen-Wilson, & Patterson, 2006; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Davis & Johnsrude, 2003; Binder et al., 2000; Scott, Blank, Rosen, & Wise, 2000; Mummery, Ashburner, Scott, & Wise, 1999; Demonet et al., 1992). Stronger activation in the left superior temporal gyrus and sulcus (STG/STS) is typically observed for speech compared to nonspeech sounds. However, interpretation of this result is complicated by the fact that the speech and nonspeech control sounds may differ in their acoustic properties, such that differences in the pattern of activation that they elicit may reflect the differential analysis of their physical properties in auditory regions generally concerned with analysis of complex sounds. Even when the speech and nonspeech sounds are carefully matched in their spectro-temporal characteristics (Liebenthal et al., 2005), another

difficulty is that they typically differ in their familiarity to the listener, such that the differential speech versus nonspeech activation may reflect the extensive experience of humans with speech sounds rather than a specialization for phoneme perception per se.

Here we circumvent the problem of matching speech and nonspeech sounds on acoustic properties and familiarity by using sine-wave analogs of speech and speech-like sounds. Sine-wave speech analogs are tone complexes in which the time-varying center frequency and power of each speech formant are represented by a tone varying in frequency and amplitude (Remez, Rubin, Pisoni, & Carrell, 1981). A naïve listener typically perceives these sounds as nonspeech. When informed that the sounds correspond to speech, and after brief training, listeners can usually perceive them as speech (Liebenthal, Binder, Piorkowski, & Remez, 2003; Remez, Pardo, Piorkowski, & Rubin, 2001; Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989). Sine-wave analogs lack the fine-grain acoustic properties of speech such as pitch and harmonic structure and are therefore unfamiliar to the listener (whether they replicate speech or not). However, they preserve the coarse dynamic features of individual formants, which is sufficient to evoke phonetic perception.

We used functional magnetic resonance imaging (fMRI) to assess brain activation patterns associated with phonetic perception by comparing the activation with sine-wave speech analogs, before and after being in-

Medical College of Wisconsin

formed of their phonetic nature, while subjects were engaged in a discrimination task. Because identical stimuli were used in the naïve and informed conditions, any differences in activation between these scans could not be attributed to differences in the acoustic properties of the sounds. This design, contrasting activation associated with the same sine-wave speech and nonspeech analogs under naïve and informed conditions, builds upon a previous study from our group (Liebenthal et al., 2003). The previous study, however, used a demanding auditory task requiring resolution of sinewave words into their constituent tones, thereby interfering with their phonetic analysis. Here we used a three-interval, two-alternative (ABX) discrimination task requiring integral analysis of the sounds without interference with phonetic perception. This task also imposed a relatively high memory load, which was expected to promote phonetic perception in the informed condition (Crowder, 1982; Repp, Healy, & Crowder, 1979). In addition, a control condition was created using tone complexes with acoustic properties similar to the sine-wave speech but lacking phonetic information, analogous to the control sounds used in Liebenthal et al. (2005). Participants were unfamiliar with the sine-wave speech and nonspeech analogs. Thus, any changes in activation between the Pre and Post scans due simply to increased practice were expected to be similar for the phonetic and nonphonetic sounds. Finally, we tested whether awareness of the phonetic properties of the speech analogs would trigger a shift in their analysis from continuous to categorical (i.e., would enhance perceptual differences between phoneme categories and minimize perceptual differences within categories; for a review, see Harnad, 2003). We hypothesized that the level of activation in a subset of regions more responsive to speech sounds may also demonstrate sensitivity to the level of categorization of these sounds.

METHODS

Participants

Participants were 28 healthy adults (19 women), 18–43 (average 26) years old, with no known neurological or hearing impairments. All subjects were native speakers of General American English and were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). Data from four other subjects were excluded due to poor behavioral performance (overall discrimination accuracy as well as across-category (AC) discrimination accuracy of less than 55%, where chance = 50%, in the Post phonetic condition). Data from 10 other subjects, scanned between certain dates, were not used due to potential scanner artifacts. Informed consent was obtained from each subject prior to the experiment, in accordance with a protocol sanctioned by the Medical College of Wisconsin Institutional Review Board.

Stimuli

The stimuli consisted of seven-step phonetic and nonphonetic sine-wave analog continua. The phonetic items replicated a /ba/–/da/ continuum, and the nonphonetic items replicated a corresponding nonphonetic continuum created by spectrally inverting the first formant of the syllables. The third formant of the nonphonetic tokens was further manipulated to render the overall nonphonetic discrimination accuracy comparable to that of the phonetic continuum (Liebenthal et al., 2005). Tokens 3 and 5 from both continua are shown in Figure 1. Thus, the phonetic and nonphonetic continua were matched on token duration, amplitude, and spectrotemporal complexity. However, tokens in the nonphonetic continuum were not analogous to any English phoneme. The sine-wave analogs were generated using in-house sine-wave synthesis software. Sine-wave tones replicating each of the first three formants of the syllables and of the nonphonetic sounds were synthesized based on time-varying formant center-frequency and intensity values of the original speech and nonspeech sounds measured at 10-msec intervals. Frequency and intensity values at intermediate time points were computed using cubic spline interpolation. Intensity values for the second and third formants were scaled, respectively, to 95.7% and 78.6% of their value for the first formant in order to maintain the amplitude relationship between the first three formants of the original sounds. The resulting sine-wave formant analogs were sampled at 22050 Hz. The three sine-wave formants of each token were then combined into a complex tone and edited to
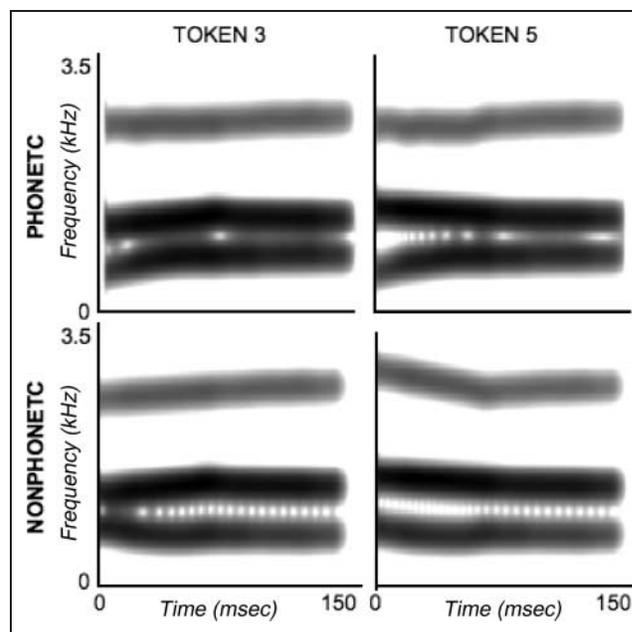

Figure 1. Spectrograms of the sine-wave sounds used in the study. Tokens 3 and 5 in the phonetic (top) and nonphonetic (bottom) continuum are shown.

150 msec duration with a 5-msec rise-decay envelope using Macromedia SoundEdit 16 (v.2.0) software.

The sounds were delivered binaurally through a stethoscopic headset with insert eartips using the Avotec SS-3100 pneumatic audio system (Jensen Beach, FL). This system provides a flat frequency response (±5 dB) at 150–4500 Hz, covering the spectral range used in this study. The sound intensity was set to a comfortable level of approximately 70 dB and was slightly adjusted between participants to accommodate for individual differences in hearing and in positioning of the eartips. For each participant, the level was kept constant throughout the session.

Sound presentation was controlled by a personal computer running PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993).

Experimental Procedure

The experimental procedure is summarized in Table 1. Prior to scanning, subjects were familiarized with the stimuli and tested with an identification task. They first listened to nine instances of each of the two anchor points (i.e., Tokens 1 and 7) of the phonetic (P) continuum and then completed 20 trials in which they were required to identify the anchor points as "sound1" or "sound2" by pressing one of two keys. For each trial, they received visual feedback in form of the correct response displayed on the computer screen. The subjects were then tested on identification of 10 presentations in random order of all seven tokens of the continuum, using the same labels ("sound1" or "sound2"). No feedback was provided. The same procedure was repeated with the nonphonetic (N) continuum. The N sounds were also labeled "sound1" and "sound2."

The subjects were then briefly familiarized with the ABX discrimination task. This was a two-alternative forced-choice task, in which the subjects heard three sounds in succession, separated by 500-msec interstimulus intervals, and decided whether the third sound (X) was identical to the first or the second sound in the preceding AB pair, by pressing one of two keys. Visual feedback was provided after each trial during training, showing the correct response. Only anchor points were used for this familiarization task.

In the scanner, subjects performed the ABX task for four runs, alternating between P and N conditions with each run. Previous research using this /ba–da/ speech continuum (Liebenthal et al., 2005) and pilot studies with the sine-wave analogs used for this experiment indicated a category boundary near Token 4 in the continuum. There were a total of 20 across-category (AC; Tokens 3–5) and 20 within-category (WC; 10 each of Tokens 1–3 and 5–7) AB pairs in each run, presented in random order. A trial consisted of three tokens presented during an otherwise silent period between image

acquisitions. No feedback was provided. There were 10 additional silent baseline trials in each run, inserted randomly, in which no stimuli were presented.

Following this first (Pre) scan session of four runs, subjects were asked whether they had noticed speech in any of the sounds presented so far during the scan. They were then informed that "sound1" and "sound2" in the P trials were actually modified versions of /ba/ and /da/ syllables. They were also informed that "sound1" and "sound2" in the N trials were computer-generated nonspeech sounds. The subjects were instructed to listen for the speech sounds /ba/ and /da/ in the sine-wave stimuli. They were then subjected to the same familiarization procedure used to introduce the test stimuli prior to scanning [listening to P anchor points (now identified as /ba/ and /da/), identification of P anchor points with

Table 1. A Summary of the Experimental Procedure

Location	Task	No. of Trials
Outside the scanner	P Iden. with feedback on anchor points	40
	P Iden. without feedback on continuum	70
	N Iden. with feedback on anchor points	40
	N Iden. without feedback on continuum	70
	P ABX with feedback on continuum	18
	N ABX with feedback on continuum	18
Inside the scanner	**Pre: P–N–P–N ABX (4 runs)**	**40 × 4**
	Inform subjects about the nature of stimuli	
	P Iden. with feedback on anchor points	30
	P Iden. without feedback on continuum	70
	N Iden. with feedback on anchor points	30
	N Iden. without feedback on continuum	70
	Post: P–N–P–N ABX (4 runs)	**40 × 4**

In the "Task" column, regular text indicates tasks performed prior to or between functional scans. **Bold** text indicates tasks performed during functional scans.

P = phonetic; N = nonphonetic; Iden. = identification task; ABX = ABX discrimination task. Tasks are listed in the order that they were administered except that the order of P and N blocks was counterbalanced across subjects.

visual feedback, followed by identification testing using the whole P continuum]. The training and testing procedures were then repeated with the N sounds, using the same "sound1" and "sound2" labels as before. This entire training and testing procedure lasted approximately 20 min and was performed while the subjects lay in the scanner.

The subjects were then scanned again (Post scans) while performing the ABX task, using exactly the same procedure as in the Pre scans.

The order of P and N runs in training, testing, and scanning was counterbalanced across subjects such that approximately half the subjects were exposed to the P stimuli first and the other half were exposed to the N stimuli first, in both Pre and Post scans.

Image Acquisition

Images were acquired on a 1.5-T GE Signa scanner (GE Medical Systems, Milwaukee, WI). Clustered (or "sparse") acquisition (acquisition time = 2100 msec) was used to collect functional image volumes separated by intervening periods of silence. T2*-weighted, gradient-echo, echo-planar images (TE = 40 msec, flip angle = 90°, NEX = 1) were collected at 8-sec intervals. Trials were positioned such that they started 1 sec after the end of each image acquisition and were followed by a silent window of approximately 3.5 sec for subjects to respond. The hemodynamic response to the last stimulus in a trial (X) was expected to peak at 4–6 sec after the onset of X, coinciding with the time of the next image acquisition. Response time (RT) was measured from the onset of X. The functional images were constructed from axially oriented 22 contiguous slices with $3.75 \times 3.75 \times 4$ mm voxel dimensions, covering the whole brain except the most dorsal fronto-parietal regions. Fifty images were acquired in each of the four Pre and four Post runs. An additional image, collected at the beginning of each run, was discarded. High-resolution anatomical images of the entire brain were obtained using a 3-D spoiled gradient-echo sequence ("SPGR"; GE Medical Systems, Milwaukee, WI), with $0.9 \times 0.9 \times 1.2$ mm voxel dimensions.

Categorical Perception Index

To assess the categorical nature of behavioral performance, logistic regression was performed on each subject's identification data. Logistic regression (Hosmer & Lemeshow, 2004) fits an S-shaped curve to the data using the maximum-likelihood method, and generates coefficient estimates for the function that is most likely to describe the observed pattern of data. Under the logistic regression framework, the probability of a /ba/ response can be modeled as

$$p(/\text{ba}/) = 1 / \left(1 + E^{-(\alpha + \beta X)} \right)$$

where X is the predictor variable (here, the position of the token or token-pair in the continuum). The coefficient β can be interpreted as the steepness or slope of the S-curve. High values of $|\beta|$ suggest a steep, step-like curve characteristic of categorical perception. Low values suggest a more linear or continuously varying response, and values close to 0 indicate a flat response curve or chance performance.

Here, a categorical perception index (CPI) was defined as the increase in β from Pre to Post scans ($\beta_{post} - \beta_{pre}$). A high CPI indicates that the perception of sounds became substantially more categorical from Pre to Post scan, whereas a CPI of 0 indicates no change.

Image Analysis

Within-subject analysis consisted of spatial coregistration (Cox & Jesmanowicz, 1999), deconvolution and voxelwise multiple linear regression (Ward, 2001) with reference functions representing four experimental conditions: pre-phonetic (PreP), pre-nonphonetic (PreN), post-phonetic (PostP), and post-nonphonetic (PostN). Individual data were smoothed with a Gaussian filter of 4 mm full width at half maximum. Anatomical scans and functional maps were projected into standard stereotaxic space (Talairach & Tournoux, 1988) using AFNI (Cox, 1996). In a random-effects analysis, individual coefficient maps were contrasted against a constant value of 0 to create group t maps. The group maps were thresholded at voxelwise $p < .03$. Clusters smaller than 732 µl (13 voxels) were removed to achieve a corrected mapwise $p < .05$ as determined by Monte Carlo simulations (Ward, 2000), which provide the probability of clusters of various sizes occurring by chance.

To examine the relation between activation and behavioral performance, individual CPI measures for P sounds were correlated with the activation in the PostP–PreP contrast of each participant on a voxelwise basis, using Spearman's rank correlation. Spearman's correlation was used because it is relatively robust to the presence of outliers in the activation or in the CPI. To gain more sensitivity, a region of interest (ROI) containing bilateral temporal lobes was defined for computing correlations, using area definitions from the Talairach Daemon in AFNI (Lancaster et al., 2000) This ROI included Heschl's gyrus, the superior, middle, and inferior temporal gyri, and the supramarginal gyrus in the left and right hemispheres. The correlation maps were thresholded at voxelwise $p < .03$, and clusters smaller than 281 µl (5 voxels) were removed to obtain a corrected $p < .05$. An identical procedure was performed for N sounds, using the PostN–PreN contrast.

RESULTS

Behavioral

Accuracy and RT data from the identification and ABX tasks are shown in Figure 2. In the identification task,
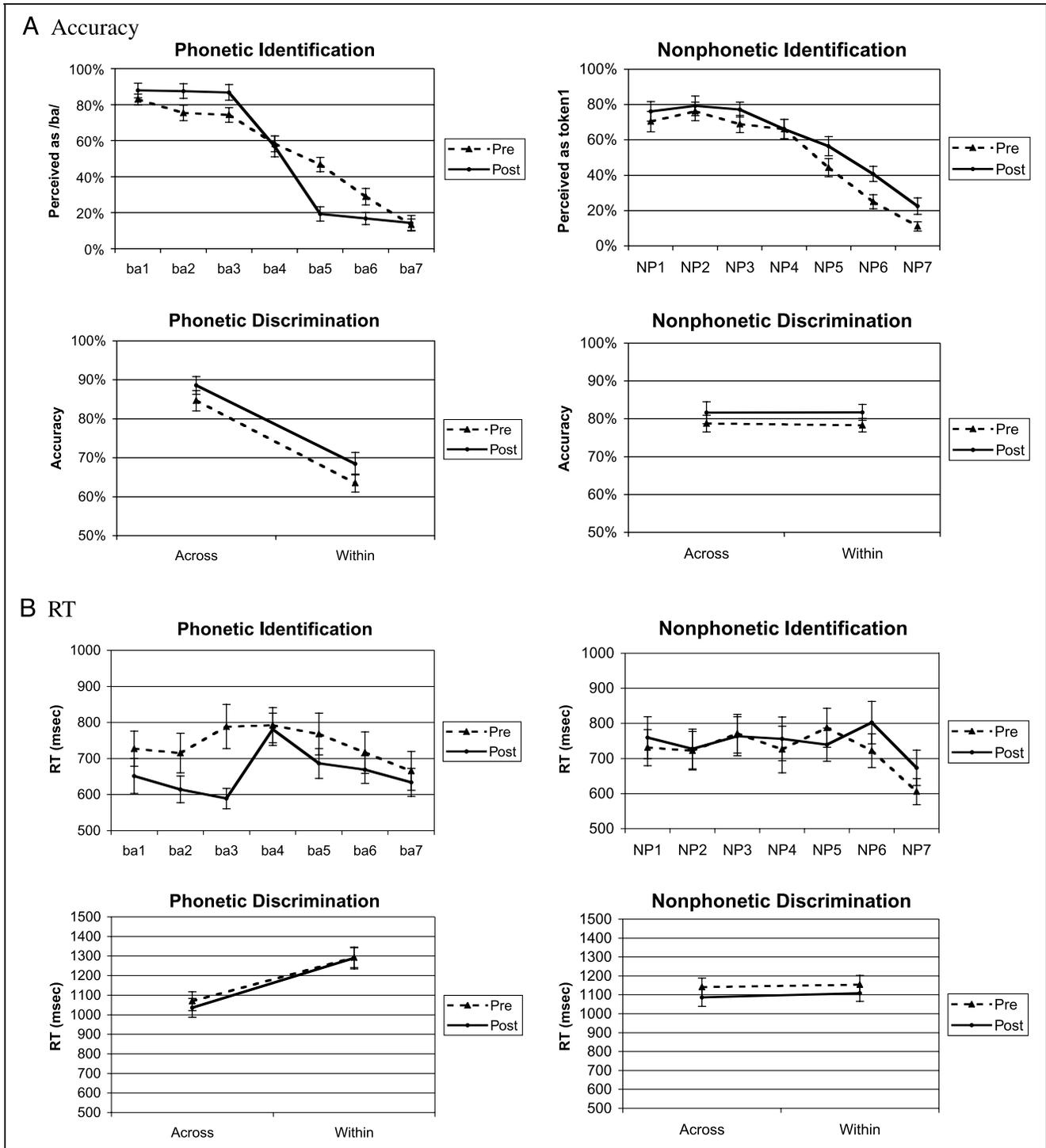
Figure 2. Performance of the subjects on the identification and discrimination tasks in Pre (naïve) and Post (informed) conditions. Error bars indicate the standard error of the mean. (A) Accuracy and (B) RT in milliseconds (msec). "Across" and "Within" represent across- and within-category discriminations, respectively.

both PreP and PreN conditions showed a continuous, mostly linear change in identification accuracy along the continuum (Figure 2A). This is consistent with the suggestion that the participants did not have discrete representations for the sounds in either Pre conditions, and they did not spontaneously perceive the P sine-wave stimuli as speech. In contrast, after the participants were informed about the phonetic nature of the P stimuli, their performance became more categorical with them, in that the two ends of the continuum were consistently identified as /ba/ or /da/. This was not the case for the PostN condition, in

which there was no significant change from the PreN performance.

The identification performances were assessed quantitatively by entering the β coefficients (slope parameters) obtained by logistic regression into a two-way repeated-measures analysis of variance with factors for training (Pre vs. Post) and sound type (P vs. N). There was a main effect of training [$F(1, 27) = 4.82, p < .037$], a main effect of sound type [$F(1, 27) = 6.78, p < .015$], and an interaction [$F(1, 27) = 6.32, p < .019$]. Post hoc comparisons with Tukey's HSD tests revealed a significant increase in β from PreP to PostP ($p < .013$) but no change in β from PreN to PostN ($p > .9$).

On the discrimination task, performance did not vary in the Post condition compared to the Pre condition for both P and NP sounds. AC accuracy was better than WC accuracy in the PostP condition, consistent with categorical perception in that condition. However, this difference was already present in the PreP condition, before subjects could categorize the P sounds. For the N sounds, AC and WC discrimination did not differ for either PreN or PostN conditions. A three-way repeated-measures analysis of variance was carried out with factors for sound type (P, N), training (Pre, Post), and contrast (AC, WC). There were main effects of sound type [$F(1, 27) = 6.15, p < .020$], training [$F(1, 27) = 7.54, p < .011$], and contrast [$F(1, 27) = 59.79, p < 10^{-6}$]. There was also an interaction between sound type and contrast [$F(1, 27) = 36.18, p < 10^{-5}$]. No other interactions were significant. Post hoc comparisons using Tukey's HSD revealed that AC accuracy was higher than WC accuracy in both PreP and PostP conditions (both $p < .0001$). The increase in AC and WC accuracy from PreP to PostP was not significant (both $p > .32$). There was no difference between AC and WC accuracy for PreN or PostN conditions, and the change in accuracy from PreN to PostN was also not significant (all $p > .76$).

The overall improvement in discrimination accuracy, from Pre to Post conditions, was the same for P and N conditions. The mean improvement in P was 8.9% ($SD = 22.9$), and in N it was 6.3% ($SD = 12.7$) ($p > .59$).

The RT results (Figure 2B) largely mirrored the accuracy data. In the identification task, RT was similar across most of the continuum in the PreP and PreN conditions. In the PostP condition, there was a reduction in RT for tokens at either end of the continuum and an increase in RT at the middle of the continuum (Token ba4), corresponding to the category boundary indicated by the accuracy data. The discrimination RT results also mirrored the accuracy data, in that RT was lower AC than WC for both PreP and PostP conditions.

In summary, the behavioral data indicate that participants did not have discrete category representations for either the P or N sounds in the Pre phase, but divided the P continuum into two perceptual categories in the Post phase. This change in categorization from nonphonetic to phonetic perception was reflected by a shift in identification but not in discrimination curves. Verbal reports of the participants after the Pre scan indicated that no participant had recognized the sounds in the Pre conditions as speech.

Functional Magnetic Resonance Imaging

The fMRI results for various conditions and contrasts are shown in Figures 3 and 4. The Appendix lists peak and activation cluster information for the contrasts. Compared to the baseline, each condition showed extensive activation that included bilateral temporal, frontal, and parietal areas (Figure 3).

PreP versus PreN

No areas were found to be more active for PreP, whereas small clusters in the left and right posterior cingulate gyrus were found to be more active for PreN (Figure 4A).

PostP versus PreP

In this critical contrast, the only area activated more for the PostP condition was in the posterior left STG/STS.
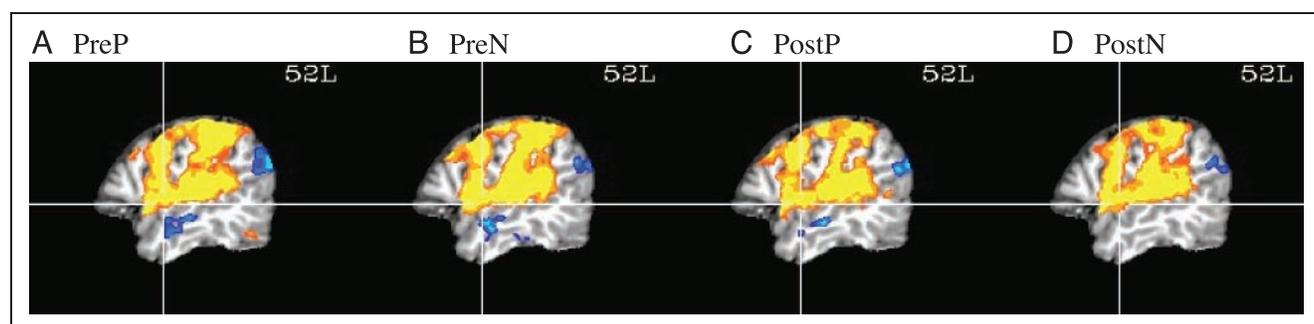

Figure 3. Activation for each experimental condition compared to the baseline, overlaid on sagittal slices of an anatomical image of one subject. Captions on each image indicate the lateral distance (mm) of the slice from the anterior-to-posterior commissure line in the right (R) or left (L) hemisphere. All maps are thresholded at corrected $p < .05$. The crosshairs show the stereotaxic y- and z-axis.
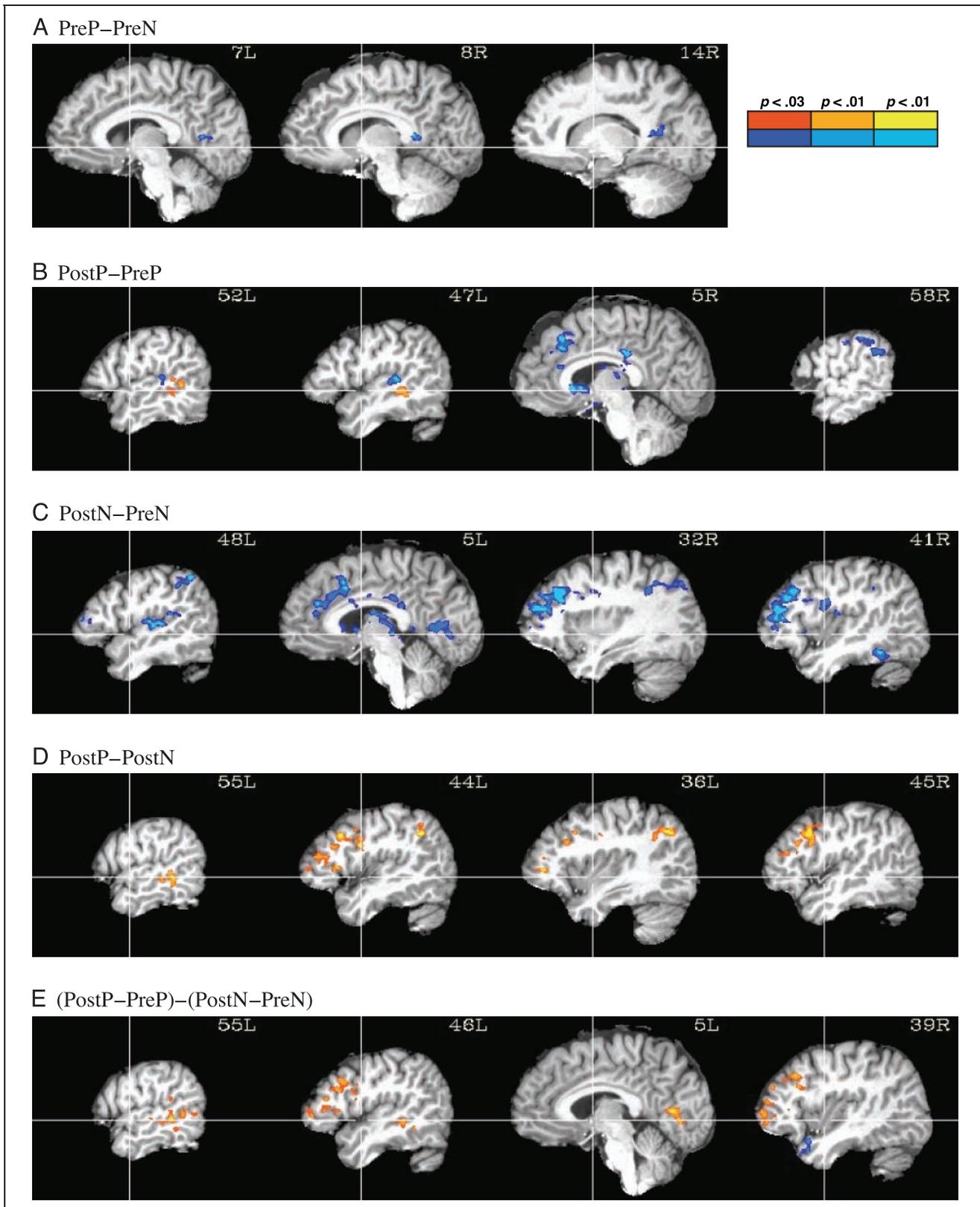
Figure 4. Activation maps for contrasts (A) PreP–PreN, (B) PostP–PreP, (C) PostN–PreN, (D) PostP–PostN, (E) Sound type × Scan interaction (PostP–PreP)–(PostN–PreN).

A number of areas showed higher activation for the PreP condition, including the bilateral posterior and anterior cingulate gyrus and the basal ganglia. The right superior and middle frontal gyri (SFG and MFG), precentral gyrus, and supramarginal gyrus (SMG), as well as a cluster on the left planum temporale, were also more active for the PreP condition (Figure 4B).

PostN versus PreN

No areas were more active for the PostN condition compared to the PreN condition. A number of areas were more active for the PreN condition, and these overlapped to a large degree with those activated for the PreP condition in the previous contrast. These included the bilateral anterior and posterior cingulate gyrus, intraparietal sulcus (IPS), basal ganglia, and MFG (right > left); the right precentral gyrus and fusiform gyrus; and the left STG/planum temporale (Figure 4C).

PostP versus PostN

The left inferior frontal gyrus (IFG), IPS, STG/STS, and precentral gyrus, as well as the bilateral MFG and precuneus, and the right SFG, were activated more in the PostP than in the PostN condition. No areas were activated more for the PostN condition (Figure 4D).

(PostP − PreP) versus (PostN − PreN)

Some of the difference in activation between Pre and Post scans could be due simply to differences in task difficulty as a result of practice and training. Some of this training effect can be removed by comparing Post–Pre activation in phonetic and nonphonetic conditions (i.e., the interaction between training and sound type). In this contrast, positively activated areas included the bilateral

MFG and IFG (right > left) and the posterior left STG and STS (Figure 4E). By comparison with the other contrasts, it is apparent that the positive values in the left STG/STS are due to a greater *increase* of activity with training for the P than for the N sounds, whereas positive values in other areas are due to a larger *decrease* in activity with training for the N sounds than for the P sounds. The right anterior STG showed negative values in the interaction contrast, due mainly to an increase in activation from the PreN to PostN condition.

Correlation with Behavioral Data

Behavioral performance on both identification and discrimination tasks varied across participants. Subjects also varied in their ability to hear the sine-wave sounds as speech, presumably leading to variation in the degree of categorical perception of the sounds. As noted in the Methods section, the CPI measures the *change* in degree of categorical identification, estimated by the change in slope of the logistic regression curve, from Pre to Post scans. High values indicate a change from continuous to categorical identification, whereas low values indicate little change. The mean CPI for the P sounds (CPI_P) was 1.24 ($SD = 2.63$), whereas the mean CPI for the N sounds (CPI_N) was −0.08 ($SD = 0.94$). CPI_P and CPI_N for individual subjects are shown in Figure 5. The individual variation in behavioral performance provided an opportunity to examine whether the degree of activation in the different brain regions was correlated with the degree of change in categorical identification exhibited behaviorally by the participants, as measured by the CPI.

We correlated the CPI_P for each participant with the level of activation in the PostP–PreP contrast in an ROI that included the bilateral lateral temporal lobes and the SMG. Two clusters in the left STG/STS and SMG were

Figure 5. Categorical perception index (CPI) for phonetic and nonphonetic sounds for individual subjects, sorted according to the phonetic CPI. CPI measures the change in categorical perception from Pre to Post scans.
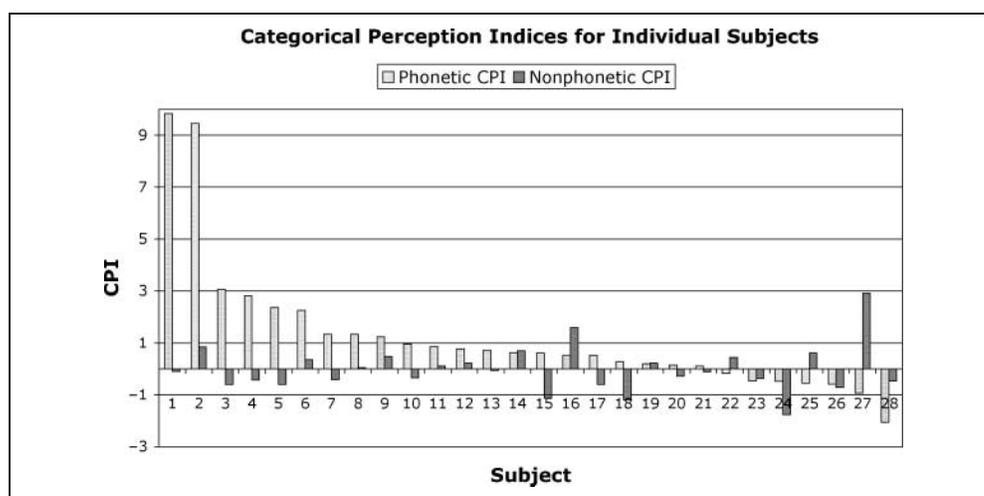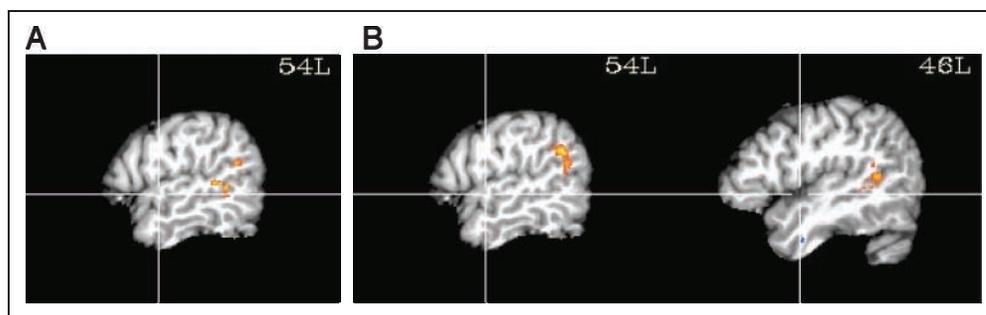
Figure 6. (A) Areas in the lateral temporal lobes in the PostP–PreP contrast that are correlated with the CPI for P sounds (CPI$_P$). (B) Areas in the lateral temporal lobes in the PostN–PreN contrast that are correlated with the CPI for N sounds (CPI$_N$).

found to be correlated with the CPI$_P$ (Figure 6A). A scatterplot of the individual training-induced activation in the PostP − PreP contrast at the maximally correlated voxel in the posterior STS (Talairach coordinates −52, −34, 7) plotted against individual CPI$_P$ values is shown in Figure 7A. Because Spearman's rank correlation was used, subjects' rank is plotted on the *x*-axis.

We similarly correlated the CPI$_N$ with the level of activation in the PostN–PreN contrast in the same ROI. Although no additional information was provided about the N sounds prior to the Post scan, some variation in CPI$_N$ was observed. The mean and variance in CPI$_N$ were not as large as those of CPI$_P$, as expected (across all subjects, β did not change significantly between PreN

Figure 7. Scatterplot of peak activation and CPI. (A) Activation at −52, −34, 7 in the PreP–PostP contrast and CPI for P sounds. (B) Activation at −56, −50, 18 in the PostN–PreN contrast and the CPI for N sounds.
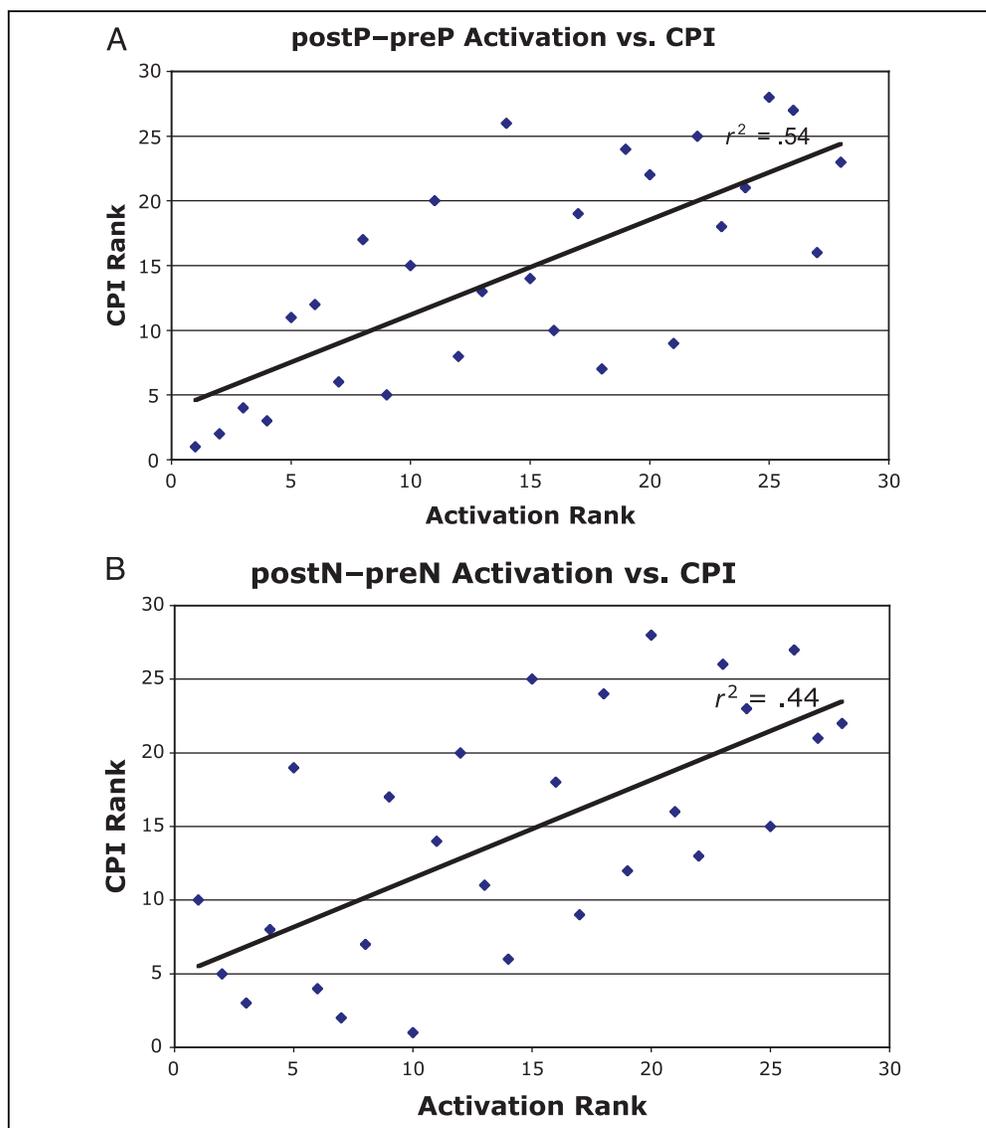
and PostN, as mentioned in the behavioral results). We were interested, however, in testing whether changes in categorical perception of nonspeech sounds might be correlated with the level of activation in temporal regions and whether the regions emerging in this analysis would overlap with those found to be sensitive to speech categorization. A cluster in the left SMG, extending into the posterior STS, was found to be correlated with the CPI_N. Smaller clusters in the left anterior MTG and right SMG were also correlated with CPI_N (Figure 6B). A scatterplot of the individual level of activation at the maximally correlated voxel in the left SMG (Talairach coordinates −56, −50, 18) plotted against the individual CPI_N is shown in Figure 7B.

We then examined whether the area in the left posterior STG/STS that was correlated with CPI_P in the PostP–PreP contrast was also correlated with CPI_N in the PostN–PreN contrast. To this end, a spherical ROI with radius of 10 mm centered at the peak of the cluster in STS (Talairach coordinates −52, −34, 7) was created. Activation in the PostN–PreN contrast was correlated with CPI_N (voxelwise $p < .03$, corrected $p < .05$) in a small cluster within this ROI (peak Talairach coordinates −51, −39, 10; cluster volume 94 μl).

Finally, we examined whether the correlations between CPI and the Pre to Post change in level of activation in the left posterior temporal region could be explained by the small general improvement in discrimination ability from Pre to Post scans rather than by changes in categorization ability. Overall change in discrimination (combining WC and AC trials) was calculated for each subject for both P and N sounds and correlated with the activation in the spherical ROI defined above, for PostP–PreP and PostN–PreN contrasts, respectively. No correlation between level of activation and overall change in discrimination was found in either analyses.

DISCUSSION

Behavioral Data

The identification functions suggest that perception of the P sounds shifted from continuous to dichotomous after subjects were informed of the phonetic potential of the sounds, indicative of categorical perception. Perception of the N sounds remained continuous in both Pre and Post conditions.

The discrimination results were consistent with categorization of the P sounds in the Post condition and continuous perception of the NP sounds in both the Pre and the Post conditions. However, the advantage for AC discrimination over WC discrimination for P sounds in the Pre condition was unexpected in light of the continuous identification function and the fact that subjects reported not hearing the sounds as speech in that condition. Interestingly, a similar effect was observed by Dehaene-Lambertz et al. (2005), who reported a small

but significant advantage for the AC sine-wave /ba/–/da/ discrimination relative to the WC discrimination in the naïve subjects. It is possible that the AC advantage in the naïve condition in both of these studies reflects a physical discontinuity in the sine-wave continuum that coincides with the phonetic category boundary. Schwab (1981) observed that naïve listeners to sine-wave speech analogs could label as accurately as informed listeners sounds in which the first and second formants (F1 and F2, respectively) spectral transitions changed in the same direction (such as in /ba/) but were less accurate when changes were in opposite directions (such as in /da/). The change in direction of the F2 transition that occurs at the boundary between /ba/ and /da/ may facilitate the discrimination between them in the naïve condition. Subjects may have been able to capitalize on this perceptual discontinuity in the phonetic continuum during the discrimination but not the identification task in the naïve condition. This is because the latter task relies on retrieval of internal representations of the sound categories and the only representations available to them for the sine-wave speech analogs in the naïve state were the trained anchor points. The discrimination results for N sounds were consistent with their continuous perception, with no accuracy or RT differences across the continuum. A perceptual discontinuity akin to that in the phonetic continuum did not occur in the nonphonetic continuum, possibly because in that continuum the F1 and F2 transitions, which contain the bulk of the sound energy and are the main cues for identification of the sounds, never closely covaried in direction. In that continuum, F1 changed from a falling pattern to a dip whereas F2 concurrently changed from a rising pattern to a falling pattern.

fMRI Data

Phonetic Perception in the Posterior STS

Compared to the PreP condition, the PostP condition more strongly activated an area in the left middle/posterior STS (approximately between Talairach $y = −30$ and $y = −40$). Activation in this general region has been reported in a number of previous studies comparing speech sounds to nonspeech sounds, but the interpretation of these differences has been problematic. For example, the lateral STG/STS (L > R) was activated for words > tones, pseudowords > tones, and reversed speech > tones contrasts in a study by Binder et al. (2000). Because the activation did not appear to depend on the phonetic intelligibility of the stimuli, these authors raised the possibility that it may have been due simply to the greater spectro-temporal complexity of the speech and reversed speech sounds compared to the tones. Liebenthal et al. (2005) subsequently compared CV syllables and nonspeech sounds of comparable acoustic complexity, and found greater

activation for the speech sounds in a similar region. In that study, the left STG/STS activation could not be attributed to the acoustic properties of the speech sounds or to differences in task demands between the speech and nonspeech conditions because the speech and nonspeech sounds were closely matched in spectro-temporal complexity, harmonic structure and periodicity, and task performance was equivalent between the conditions. However, the differential STG/STS activation could be attributed either to the linguistic nature of the speech stimuli, their categorical perception, or their familiarity. Dehaene-Lambertz et al. (2005) also found activation in the left posterior STS for sine-wave speech stimuli in the informed compared to the naïve condition. However, a nonspeech condition was not included in the Dehaene-Lambertz study to control for stimulus repetition and practice effects between the naïve and the informed scans. A similar pattern of fMRI activation was also reported by Möttönen et al. (2006) using a mixed-effects analysis in a sine-wave speech perception task. Behavioral measures of the degree of speech or nonspeech perception, however, were not reported in this study. The analysis was also restricted to a small posterior temporal ROI, so the effects in other regions were not clear. The present study goes a step further in showing unequivocally that activation in the left posterior STS region during speech perception cannot be attributed to the spectro-temporal complexity or familiarity of speech, or to practice effects.

In this study, the PostP > PostN activation highlights the difference between perception of phonetic and non-phonetic sounds, while controlling for the effects of task practice and habituation to the stimuli. This contrast revealed a focus in the left STS near the focus found for PostP > PreP. Activation in the same left posterior STG/STS region was also observed in the interaction between sound type (P, N) and scan (Pre, Post), consistent with the suggestion that it is due to perception of the sine-wave phonetic analogs as speech.

Effects of Practice and Habituation

A number of areas activated more in the PreP compared to the PostP condition, such as the anterior cingulate gyrus, basal ganglia, SFG, and MFG, have been associated with general task difficulty, attention, working memory, decision making, and response selection processes (Culham & Kanwisher, 2001; Bush, Luu, & Posner, 2000; Duncan & Owen, 2000). Very similar areas were also activated for the PreN condition compared to the PostN condition. We suggest that with practice and repeated exposure to the task and stimuli, the subjects became more efficient at the task in Post conditions compared to Pre conditions, requiring fewer resources for task performance. Similar fronto-parietal areas were also activated in the PostP > PostN contrast, in which both

conditions have similar practice effects. The PostP condition, however, is associated with additional information about sound categories. This activation likely reflects working memory and decision-making processes engaged when attempting to map sounds onto known categories, which are absent from the PostN condition.

There was also an area on the left mid/dorsal STG, including the planum temporale, which was activated more for the PreP compared to PostP condition, and also for the PreN compared to PostN condition. The decrease in activity from Pre to Post in this region may represent the habituation of early auditory processing stages due to repeated exposure to the same stimuli. This area was not activated in the PostP–PostN comparison, likely because both the PostP and PostN conditions entail similar habituation effects. Altogether, these results demonstrate a clear dissociation of function between the dorsal STG/planum temporale, which is sensitive to a wide variety of sounds and shows habituation to repeated sine-wave sounds, and the more ventral STG/STS, which is associated with representation of more abstract categorical properties of the sounds, showing an increased response when the same sounds can be mapped onto categories (Hall, Hart, & Johnsrude, 2003; Griffiths & Warren, 2002; Binder et al., 2000; Binder, Frost, Hammeke, Rao, & Cox, 1996).

Correlations with Behavioral Data

We hypothesized that if activation in the posterior STS/STG region is associated with the categorical perception of sounds, the level of this activation should be correlated with a behavioral index of the degree of categorical perception. As predicted, a voxelwise correlation analysis conducted on left and right temporal lobe ROIs showed an area in the posterior STG/STS that was correlated with CPI. The subjects with a larger CPI also showed greater activation in the PostP–PreP contrast in this region. Activation in this region was not correlated with the overall improvement in discrimination ability.

The inclusion of the N conditions in the experiment provided an opportunity to examine whether the posterior STS activation was specific to phonetic categorization, or was related to categorical perception in general. Although this study was not designed to induce changes in CPI_N and a significant systematic change was neither expected nor observed, there were small individual variations in CPI_N. Some subjects appear to have developed weak categories for the N sounds (small positive changes in the CPI_N), perhaps encouraged by the identification training and testing procedures. A few other subjects apparently suppressed these categorical representations (small negative changes in the CPI_N; Figure 5). The ROI analysis showed that an area in the posterior STS (near $y = -40$) in the PostN–PreN contrast was positively correlated with CPI_N. The magnitude

and extent of the correlated area were small, possibly due to the fact that the variation in CPI_N was also small. Nevertheless, this association of the posterior STS with categorical perception of *nonphonetic* sounds tentatively suggests that this region is not just sensitive to well-learned phonetic representations, but also to recently learned nonphonetic categories. Regions in the inferior SMG/posterior STG were also correlated with the degree of improvement in categorical identification, more strongly for N sounds. The SMG has been suggested to subserve acoustic–phonetic recoding (Hickok & Poeppel, 2000; Caplan, Gow, & Makris, 1995). SMG activation is also reported in training studies in which a nonnative sound category is learned (Golestani & Zatorre, 2004; Callan et al., 2003), and differences in the white matter volume near the SMG are associated with the ability to learn novel sounds (Golestani, Paus, & Zatorre, 2002). Along with the current results, these results are consistent with the suggestion that the SMG plays a role in representing or learning auditory categories in general, not just phonetic ones.

Conclusions

Physically identical auditory stimuli can engage different areas of the brain or engage the same area to different degrees, depending on whether they are perceived as phonetic or nonphonetic. An area in the left posterior STS, surrounding Talairach $y = -40$, is activated more when sine-wave speech analogs are perceived as speech and can be associated with learned phoneme categories. Unlike activations in most comparisons of speech and nonspeech stimuli, this activation cannot be attributed to acoustic differences between the stimuli. Activation in this region is also correlated with the degree of categorical identification of phonetic, and also to some extent, *nonphonetic* sounds. We therefore infer that in this brain region, prelinguistic representations of auditory inputs activate category representations. The left SMG also plays a role in the categorical perception of auditory information, perhaps particularly in the learning of novel sound categories.

APPENDIX

The location of activation peaks in various contrasts. The volume of the cluster (μl), the mean and maximum z-score of the cluster, the location of the peaks in the atlas of Talairach and Tournoux (1988), and approximate Brodmann's areas (BAs) of the peaks are reported. Multiple peaks are reported for some of the larger clusters.

Volume	Mean	Max	x	y	z	Structure	Approx. BA
(a) PreP–PreN							
PreP > PreN							
None							
PreN > PreP							
759	−2.5	−3.7	8	−43	10	R poCiG	29
		−3.1	17	−62	18	R Precuneus	31
742	−2.5	−3.4	−7	−60	9	L Cuneus	30
(b) PostP–PreP							
PostP > PreP							
1066	2.6	3.4	−51	−43	5	L STS	21/22
		3.3	−58	−33	5	L STG	22
		3.2	−47	−31	−1	L STS	21/22
PreP > PostP							
8470	−2.7	−5.4	7	−28	31	R poCiG	31
		−4.8	−17	14	3	L Putamen	
		−4.8	3	17	1	R Caudate	
		−3.5	−12	−20	6	L Thalamus	
		−3.4	19	−2	−21	R Uncus	34
		−2.9	1	−9	36	R CiG	24
3693	−2.6	−4.0	5	25	43	R SFG	8

Volume	Mean	Max	x	y	z	Structure	Approx. BA
		−3.5	−11	37	6	L antCiG	
		−3.3	12	30	20	R antCiG	32
		−3.3	−13	22	26	R antCiG	32
1104	−2.6	−4.3	61	−41	31	R SMG	40
		−3.4	56	−18	39	R prCG	4
1020	−2.5	−4.0	28	−36	1	R Hippocampus	
		−3.3	−22	−32	−19	L Culmen	
983	−2.5	−3.6	31	32	26	R MFG	9
847	−2.6	−3.9	−46	−29	10	L STG/HG	41

(c) PostN–PreN

PostN > PreN

  None

PreN > PostN

Volume	Mean	Max	x	y	z	Structure	Approx. BA
12683	−2.8	−4.8	35	37	26	R MFG	9
		−3.7	47	10	37	R MFG	8/9
		−3.7	39	1	21	R prCG	6/13
		−2.7	48	10	4	R prCG	13
10411	−2.7	−5.0	12	12	3	R Caudate	
		−4.9	3	−10	10	R Thalamus	
		−4.6	−21	11	9	L Putamen	
6099	−2.7	−4.7	−1	17	37	L antCiG	32
		−4.1	−7	37	24	L antCiG	32
		−3.4	−6	−6	32	L CiG	24
3533	−2.5	−4.2	7	−69	40	R Precuneus	7
		−3.7	24	−67	21	R Precuneus	31
		−3.3	31	−45	35	R IPS	40
		−3.2	24	−87	29	R Cuneus	18/19
3190	−2.7	−4.0	−50	−24	9	L STG/HG	41
		−3.7	−61	−22	26	L SMG	40
2217	−2.6	−4.1	−48	−50	47	L IPS	7/40
		−2.9	−27	−59	33	L IPS	
2100	−2.6	−4.0	−43	37	13	L IFG	46
		−3.6	−33	28	28	L MFG	9
1874	−2.5	−3.5	5	−34	25	R poCiG	23
1541	−2.6	−3.7	0	−73	−3	LG	18
817	−2.6	−3.6	41	−46	−15	R FG	37

(d) PostP–PostN

PostP > PostN

Volume	Mean	Max	x	y	z	Structure	Approx. BA
3034	2.6	4.2	46	12	37	R MFG	9

APPENDIX (continued)

Volume	Mean	Max	x	y	z	Structure	Approx. BA
1938	2.5	3.4	5	−79	40	R Precuneus	7/19
1848	2.6	4.0	−40	−57	39	L IPS	7
		3.4	−25	−71	35	L IPS	19
1776	2.6	3.9	−44	0	26	L prCG	6
1298	2.5	3.8	−37	41	6	L IFS	46
		3.5	−44	20	8	L IFG	44/45
1006	2.5	3.3	18	50	20	R SFG	10
736	2.6	3.4	−55	−36	0	L STS	21/22

PostN > PostP

  None

(e) (PostP–PreP)–(PostN–PreN)

(PostP–PreP) > (PostN–PreN)

Volume	Mean	Max	x	y	z	Structure	Approx. BA
1569	2.5	3.9	25	20	37	R MFG	8
		3.5	45	21	27	R MFG	46
		3.0	28	56	23	R SFG	10
1543	2.5	3.6	−51	−33	−1	L STS	21/22
		3.2	−52	−17	2	L STG	22
		3.1	−54	−44	6	L STS	21/22
1495	2.5	3.7	43	34	12	R IFG	46
		3.0	43	24	8	R IFG	13
979	2.5	3.4	−6	−64	8	L Cuneus	18/30
942	2.5	3.7	−36	40	6	L IFS	46
		3.3	−43	19	7	L IFG	44/45
737	2.5	3.2	−49	16	34	L MFG	9

(PostN–PreN) > (PostP–PreP)

Volume	Mean	Max	x	y	z	Structure	Approx. BA
2033	−2.5	−3.8	29	10	−31	R STG	38

(f) Temporal Lobe Areas in PostP–PreP Correlated with Behavioral CP Index

Volume	Mean	Max	x	y	z	Structure	Approx. BA
764	2.53	4.46	−52	−34	7	L poSTG/STS	22
324	2.54	3.54	−54	−47	19	L poSTG/SMG	40/22

(g) Temporal Lobe Areas in PostN–PreN Correlated with Behavioral CP Index

Volume	Mean	Max	x	y	z	Structure	Approx. BA
751	2.63	3.85	−56	−50	18	L SMG/STS	40/22
493	2.59	4.22	−41	1	−22	L antMTG	21/38
322	2.55	3.56	37	−51	24	R SMG	40/39
94	2.68	3.97	−51	−39	10	L poSTS	22

CiG = cingulate gyrus; STS = superior temporal sulcus; STG = superior temporal gyrus; HG = Heschl's gyrus; SFG = superior frontal gyrus; MFG = middle frontal gyrus; IFG = inferior frontal gyrus; IFS = inferior frontal sulcus; SMG = supramarginal gyrus; prCG = precentral gyrus; IPS = intraparietal sulcus; LG = lingual gyrus; FG = fusiform gyrus; po = posterior; ant = anterior.

REFERENCES

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception & Psychophysics, 45,* 237–250.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10,* 512–528.

Binder, J. R., Frost, J. A., Hammeke, T. A., Rao, S. M., & Cox, R. W. (1996). Function of the left planum temporale in auditory and linguistic processing. *Brain, 119,* 1239–1247.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4,* 215–222.

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage, 19,* 113–124.

Caplan, D., Gow, D., & Makris, N. (1995). Analysis of lesions by MRI in stroke patients with acoustic–phonetic processing deficits. *Neurology, 45,* 293–298.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25,* 257–271.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29,* 162–173.

Cox, R. W., & Jesmanowicz, A. (1999). Real-time 3D image registration of functional MRI. *Magnetic Resonance in Medicine, 42,* 1014–1018.

Crowder, R. G. (1982). A common basis for auditory sensory storage in perception and immediate memory. *Perception & Psychophysics, 31,* 477–483.

Culham, J. C., & Kanwisher, N. G. (2001). Neuroimaging of cognitive functions in human parietal cortex. *Current Opinions in Neurobiology, 11,* 157–163.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23,* 3423–3431.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *Neuroimage, 24,* 21–33.

Demonet, J. F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J. L., Wise, R., et al. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain, 115,* 1753–1768.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23,* 475–483.

Golestani, N., Paus, T., & Zatorre, R. J. (2002). Anatomical correlates of learning novel speech sounds. *Neuron, 35,* 997–1010.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *Neuroimage, 21,* 494–506.

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences, 25,* 348–353.

Hall, D. A., Hart, H. C., & Johnsrude, I. S. (2003). Relationships between human auditory cortical structure and function. *Audiology & Neuro-Otology, 8,* 1–18.

Harnad, S. (2003). Categorical perception. In *Encyclopedia of cognitive science*. London: Macmillan, Nature Publishing Group.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4,* 131–138.

Hosmer, D. J., & Lemeshow, S. (2004). *Applied logistic regression* (2nd ed.). New York: Wiley.

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping, 10,* 120–131.

Liebenthal, E., Binder, J. R., Piorkowski, R. L., & Remez, R. E. (2003). Short-term reorganization of auditory cortex induced by phonetic expectation. *Journal of Cognitive Neuroscience, 15,* 549–558.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15,* 1621–1631.

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage, 30,* 563–569.

Mummery, C. J., Ashburner, J., Scott, S. K., & Wise, R. J. (1999). Functional neuroimaging of speech perception in six normal and two aphasic subjects. *Journal of the Acoustical Society of America, 106,* 449–457.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., et al. (2006). Vowel sound extraction in anterior superior temporal cortex. *Human Brain Mapping, 27,* 562–571.

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

Remez, R. E., Pardo, J. S., Piorkowski, R. L., & Rubin, P. E. (2001). On the bistability of sine wave analogues of speech. *Psychological Science, 12,* 24–29.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–949.

Repp, B. H., Healy, A. F., & Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 129–145.

Schwab, E. (1981). *Auditory and phonetic processing for tone analogs of speech*. Buffalo, NY: State University of New York, Buffalo.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain*. New York: Thieme.

Uppenkamp, S., Johnsrude, I. S., Norris, D., Marslen-Wilson, W., & Patterson, R. D. (2006). Locating the initial stages of speech–sound processing in human temporal cortex. *Neuroimage, 31,* 1284–1296.

Ward, B. D. (2000). *Simultaneous inference for fMRI data*. AFNI manual.

Ward, B. D. (2001). *Deconvolution analysis of fMRI time series data*. AFNI manual.