Cerebral Cortex March 2010;20:583-590
doi:10.1093/cercor/bhp124
Advance Access publication July 2, 2009

Auditory Attentional Control and Selection during Cocktail Party Listening

Kevin T. Hill[1] and Lee M. Miller[1,2]

[1]Center for Mind and Brain, University of California Davis, Davis, CA 95618, USA and [2]Department of Neurobiology, Physiology, and Behavior, University of California Davis, Davis, CA 95616, USA

In realistic auditory environments, people rely on both attentional control and attentional selection to extract intelligible signals from a cluttered background. We used functional magnetic resonance imaging to examine auditory attention to natural speech under such high processing-load conditions. Participants attended to a single talker in a group of 3, identified by the target talker's pitch or spatial location. A catch-trial design allowed us to distinguish activity due to top-down control of attention versus attentional selection of bottom-up information in both the spatial and spectral (pitch) feature domains. For attentional control, we found a left-dominant fronto-parietal network with a bias toward spatial processing in dorsal precentral sulcus and superior parietal lobule, and a bias toward pitch in inferior frontal gyrus. During selection of the talker, attention modulated activity in left intraparietal sulcus when using talker location and in bilateral but right-dominant superior temporal sulcus when using talker pitch. We argue that these networks represent the sources and targets of selective attention in rich auditory environments.

Keywords: attention, fMRI, pitch, space, speech

Introduction

One of the brain's greatest perceptual challenges is to understand a single talker in a crowd of other voices. This remarkable ability relies on accurate processing of low-level stimulus attributes, segregation of auditory information into coherent objects (Griffiths and Warren 2004), and selectively attending to a single object at the exclusion of others to facilitate higher level processing (Alain and Arnott 2000; Shinn-Cunningham 2008). In his seminal paper framing the "cocktail party problem," Cherry addressed our ability to select a speech stream based on the perceived spatial location of the talker (Cherry 1953). Although location is undoubtedly a highly salient cue for selective attention (Spieth et al. 1954; Alho, Donauer et al. 1987; Woods et al. 2001), pitch differences can also distinguish auditory events to allow processing of a single stream (Alho, Tottola, et al. 1987; Bregman 1994; Darwin and Hukin 2000a; Woods et al. 2001). Particularly in realistic environments, pitch cues can be as robust as spatial cues (Darwin and Hukin 2000b). Numerous lines of evidence support an anatomical distinction in the auditory system between 2 classes of features, those important for identification, including pitch, and those for spatial or other sensorimotor processing such as articulation (Alain et al. 2001; Maeder et al. 2001; Clarke et al. 2002; Warren and Griffiths 2003; Arnott et al. 2004; Ahveninen et al. 2006; Degerman et al. 2006; Hickok and Poeppel 2007; Griffiths 2008; Tardif et al. 2008). This parallels the "what" and "where" pathways described in vision (Goodale and Milner 1992) and suggests that attention may act through distinct neural networks for each feature class. Despite the importance of both location and pitch in auditory scene analysis, we do not know the brain mechanisms of attention in realistic, cluttered auditory environments.

Understanding one talker in a crowd requires at least 2 attentional mechanisms thought to have distinct neural bases: attentional control and attentional selection (Egeth and Yantis 1997). Before attention can be used to selectively process stimuli, the brain must exert voluntary attentional control to shift attention to a location or spectral feature. Many studies in vision have shown a fronto-parietal network to be critical in the endogenous orienting of attention to a spatial location or other object feature such as color (Shulman et al. 1999, 2002; Corbetta et al. 2000; Hopfinger et al. 2000; Giesbrecht et al. 2003). In neuroimaging, the clearest demonstrations of top-down attentional control use designs in which attention is cued in the absence of stimuli, before selection can occur. Using such a design, Giesbrecht and colleagues showed that different elements of this network are biased toward spatial or nonspatial attentional control, with the superior parietal lobule (SPL) and dorsal precentral sulcus (DPreCS) showing greater activity in the spatial domain, and fusiform gyrus (FG) showing greater activity when directing attention toward a particular color (Giesbrecht et al. 2003). Studies in audition have been largely consistent with visual studies (Shomstein and Yantis 2006), suggesting that attentional control in the 2 modalities shares at least some common neural substrates (Shomstein and Yantis 2004). It may be that auditory attention shares the general fronto-parietal network with vision, showing analogous biases for location versus spectral features such as pitch. However, there has yet to be a comparison between the differential processing in attentional control of spatial and nonspatial auditory features in the absence of stimulus processing.

After attention has been shifted to a particular location or pitch by top-down control, the brain must use attentional selection to distinguish one object from others. Importantly for real environments, selective attention is inherently a function of limited processing capacity: If the brain is unable to evaluate all inputs, selection must occur so one set of inputs is processed to the exclusion of others (Lavie 1995; Luck et al. 1997; Alain and Izenberg 2003). Thus, in order to investigate the neural substrate of selective attention, the auditory system must be placed under high load conditions. Yet recent auditory studies addressing attention to location and other nonspatial features (including pitch, speaker identity, and timbre) tend to use low perceptual-load conditions, often with simple stimuli and presentation of a single auditory object at a time (Kawashima et al. 1999; Zatorre et al. 1999; Jancke et al.

2001; Maeder et al. 2001; Jancke and Shah 2002; Ahveninen et al. 2006; Degerman et al. 2006, 2007; Rinne et al. 2007; Salmi et al. 2007). These studies have commonly found attentional modulation early along the auditory processing hierarchy, in the superior temporal gyrus (STG). However, without the strenuous processing demands required by natural auditory environments with multiple overlapping auditory streams, these studies may reveal processes unrelated to the selection of one object among many. Therefore, in order to index the neural basis of attentional selection in realistic auditory scenes, we must use paradigms that present listeners with high processing load, as with Cherry's simultaneous, competing natural language stimuli.

Here we address the neural mechanisms of attentional control and attentional selection among multiple features in a rich, naturalistic auditory environment. Subjects perform a challenging, real-world task: to track sentences from one among several talkers perceived in external space. This creates a high processing load to engage attentional selection. The addition of a cue period in the absence of auditory stimuli allows us to differentiate the mechanisms of attentional control and selection. When the stimuli do occur, they are physically consistent across conditions, ensuring that observed effects are not due solely to the processing of stimulus attributes, but instead reflect the real-world challenge of picking out a single talker from a crowd.

Materials and Methods

Subjects

Sixteen subjects participated in the study. Mean age was 21.5 (±2.4) years. All subjects were right-handed native English speakers and had no history of neurological disorders or hearing loss. Participants gave written informed consent in accordance with procedures approved by the University of California Institutional Review Board and were paid for their participation.

Stimuli

We used "cocktail-party" stimuli consisting of 3 simultaneous talkers, all of whom spoke a continuous sequence of multiple English sentences from the 1969 Harvard/IEEE Speech Corpus (graciously provided by Qian-Jie Fu) (IEEE 1969). There were 172 sentences in the corpus (e.g., "The birch canoe slid on the smooth planks"), each with an average duration of 2.1 s. All sentences were used once before any was repeated, and all stimuli were derived from the same recordings of a female (mean pitch of 188 Hz). On every trial, each of the 3 talkers had a unique pitch and unique spatial location (Fig.1*B*). Pitch was shifted higher and lower symmetrically on an octave scale using Adobe Audition's (http://www.adobe.com) "Pitch Shifter" function. Using the same original recording for all stimuli ensured that there were no uncontrolled cues by which the experimental stimuli could be differentiated. The 3 talkers were placed in simulated 3D space using headphones and a head-related transfer function (HRTF) (Langendijk and Bronkhorst 2000). HRTFs were derived for each subject individually using short recordings from microphones (AuSIM, http://www.ausim3d.com/) placed inside the subject's ear canals while white noise was played at 5-deg increments in azimuth (–45° to +45°). This allowed us to place each talker at a unique location in perceived external acoustic space on the horizontal plane, maximizing the realistic nature of the stimuli in the confines of a functional magnetic resonance imaging (fMRI) scanner. There was always one talker at the midline, and the other 2 talkers were arranged symmetrically to the left and right. Combinations of pitch and location for each talker were controlled so that 2 adjacent trials never had the same configuration. Each of the streams consisted of a different continuous sequence of sentences that lasted the entire 6.3 s, and contained multiple sentence onsets (Fig. 1*C*). Each of the 3 sentence streams began at a random point in a sentence, ensuring that sentence onsets were not correlated between streams or between trials. We also trimmed the recordings so that the pause between sentences was no longer than the average gap within sentences, ensuring that sentence onsets were identifiable only by the contents of the stimulus and not merely acoustic modulations. The combined stimuli were presented at ~90 dB and attenuated by earplugs to reach a perceptual level of ~70 dB.

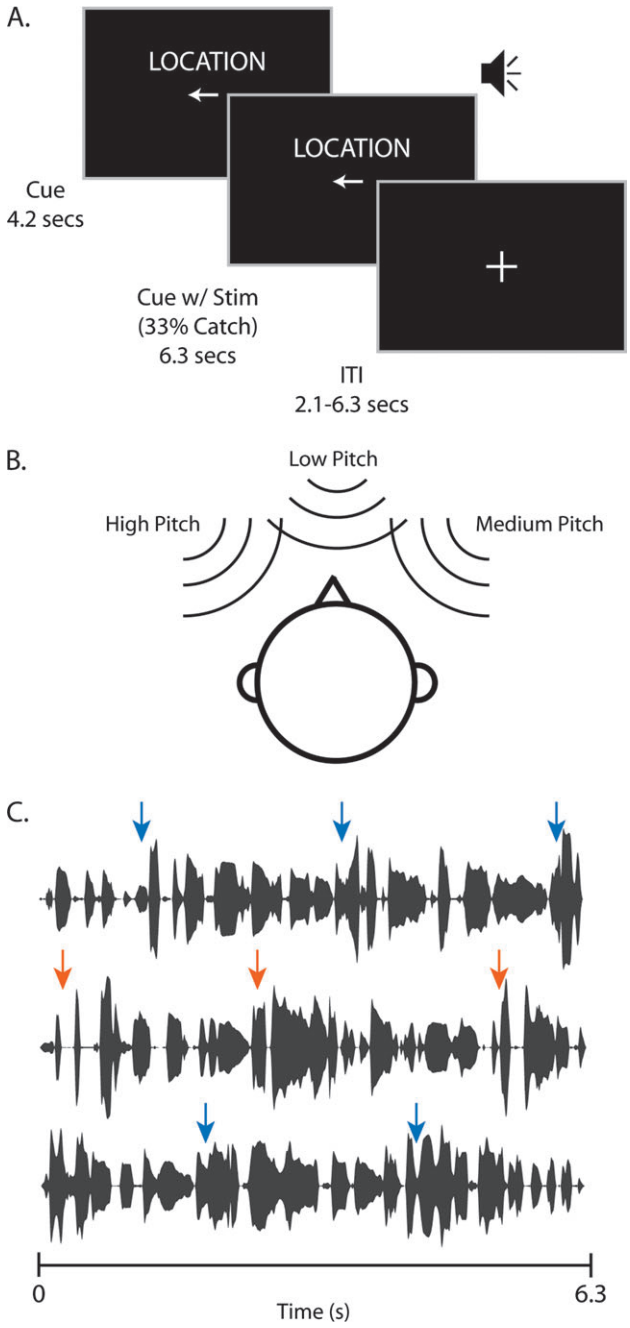

Figure 1. (*A*) A diagram of an example trial structure. Thirty-three percent of trials contained a cue period but no stimulus period (catch trials) in order to dissociate effects of cue and stimulus in the fMRI analysis. (*B*) Stimulus configuration for a single trial. Combinations of pitch and location for each talker were controlled so that 2 adjacent trials never have the same configuration. (*C*) Temporal envelopes of 3 separate exemplar speech streams. Stimuli consist of multiple continuous streams of English sentences that contained multiple sentence onsets in both the target (orange) and distractor streams (blue).

Behavioral Calibration

Equating performance across conditions is essential to ensure that the imaging results were not due to task difficulty differences. Therefore, the pitch semitone difference and the angular distance between the 3 talkers were calibrated for each subject. Pitch and location separation were calibrated using independent blocks in which the stimuli differed only in the feature being calibrated. For example, during pitch calibration, the stimuli consisted of 3 talkers that differed in pitch but were all perceived at the same spatial location. Calibration consisted of short blocks with a task similar to the main experiment (see below). Subjects were instructed to press a button whenever the middle talker (middle pitch or central location) began a sentence. Each block consisted of 4 trials of 6 s of multi-talker stimuli to ensure maximal similarity to the conditions in the experiment. All sentence targets over the 30 blocks were used to calculate a mean hit rate for the block. Performance over these blocks was calibrated in 20 blocks using an adaptive staircase that targeted a mean hit rate of 75% over each 30-s block. Spatial disparity was adapted in 5° increments, whereas pitch disparity used 0.25 semitone increments. After calibration, mean pitch separation was 1.80 ± 0.38 semitones, and mean location separation was 14.33° ± 5.30°.

Design

The experiment was organized into trials consisting of a cue and stimulus period. Between trials, subjects fixated on a cross presented in the center of the screen. During the cue period, subjects were shown both the cue type ("Location," "Pitch," or "Rest") and an icon that determined the talker to be attended. No auditory stimuli were presented during the cue period. For location trials, a left arrow denoted that the subject should attend to the left talker, a right arrow cued attention to the right talker, and a circle cued attention to the middle talker. During pitch trials, an upward arrow cued for the high pitch talker, a downward arrow for the low pitch talker, and a circle for the middle pitch talker. Rest trials contained only a fixation cross as an icon. The cue portion lasted for 4.2 s. After the cue portion, the cue text was left on the screen while the multi-talker stimuli described above came on for 6.3 s, followed by a jittered intertrial interval (Fig. 1*A*). Intertrial intervals could be 2.1, 4.2, or 6.3 s, occurring randomly with relative probability 3:2:1, respectively. In order to distinguish between blood oxygen level–dependent (BOLD) activity related to the cue period and the stimulus period of the trial, we employed a catch-trial design (Ollinger et al. 2001). On 33% of all trials, the stimuli did not play, and the trial ended. This paradigm renders the BOLD signals for the cue and stimulus trial periods independent enough for a general linear model (GLM) to disambiguate their relative contributions to the overall signal. Each experiment consisted of 6 runs lasting about 8 min. The experiment was controlled using Presentation software from Neurobehavioral Systems (http://www.neurobs.com/).

Task

During both location and pitch trials, the subject was instructed to press a button with his or her left index finger whenever the cued talker began a sentence. Therefore, there were multiple targets within each stimulus period (Fig. 1*C*). Because the contents of each sentence could not be anticipated, selective attention had to be maintained over the entire 6.3-s stimulus period to detect when a target sentence was beginning. Responses to each target were recorded as correct when the subject's button press occurred within a window 100–900 ms after the beginning of a sentence and labeled as incorrect if the subject failed to respond within that window. In all conditions, the stimuli consisted of 3 talkers with both unique pitch and location, so there were no low-level stimulus attribute differences between conditions. Subjects were instructed not to do anything during rest trials.

Imaging

Magnetic resonance imaging (MRI) experiments were carried out on a Siemens 3-T TRIO scanner with an 8-channel RF headcoil and a whole body gradient system. Foam padding was used to comfortably restrict head motion. Each session began with a series of images to determine regional anatomy, including sagittal localizer (Repetition Time (TR) = 200 ms, Echo Time (TE) = 5 ms) and T2 weighted (TR = 2000 ms, TE = 28 ms) images. Single-shot gradient-echo echoplanar images (EPIs) were acquired for 34 near-axial slices. The functional scans had the parameters: TR of 2.1 s, TE 29 ms, 64 × 64 × 64 acquisition matrix, 3.0-mm slice thickness with a 0.43-mm gap, a 220-mm field of view, bandwidth 2365 Hz/pixel, and flip angle of 84°. Auditory stimuli were presented with an audio system customized for use in high magnetic fields (MR-Confon http://www.mr-confon.de/en/). Earplugs combined with the audio system earmuffs passively attenuated the scanner noise to ~70 dB, and stimuli were played at ~90 dB. All sounds were filtered to account for the uneven frequency attenuation of the earplugs, ensuring that stimuli were perceived as intended. A high-resolution 3D magnetization-prepared rapid gradient echo image for use in intersubject coregistration was taken at the end of the session.

Data Analysis

Behavioral data were analyzed using custom in-house scripts written in Matlab 7.1 (Mathworks, http://www.mathworks.com/). fMRI data were analyzed using SPM 5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/). Motion-corrected EPIs were slice time corrected, coregistered to the subjects' T2 images, normalized to the Montreal Neurological Institute (MNI) template (Evans et al. 1993) and smoothed with an 8-mm Gaussian smoothing kernel. Before running a GLM on the data, a multiplicative linear trend for each run was removed to mitigate variability across runs. Separate GLM regressors were added for all cue periods in each condition, all stimulus periods in each condition, session mean covariates, a global intercept term, and motion parameters from both the Siemens and SPM based motion correction algorithms. A series of impulses indicating the onsets of each condition (e.g., pitch cues, location stimuli) were convolved with the canonical SPM hemodynamic response function. Even with tightly spaced events, the catch-trial design (Ollinger et al. 2001), along with jittered intertrial intervals, allows for adequate separation of the different trial components. Cortical brain maps were statistical tests across subjects on linear combinations of beta values (regression coefficients). When correcting for multiple comparisons, we used the False Discovery Rate, or FDR, which controls for the proportion of false positives among all suprathreshold voxels (Genovese et al. 2002). Region of interest (ROI) analysis was performed with the Marsbar toolbox (http://marsbar.sourceforge.net/).

Results

Behavior

Although the scanner task was difficult overall with 53.0% mean ± 12.0% standard deviation (SD) of target sentence beginnings successfully detected, this reflects an improvement in performance over the course of the trial. Although performance was quite low when a target sentence began shortly after stimulus onset, subjects were able to reach a fairly high degree of accuracy (72 ± 14.7% SD) over the 6.3-s trial (Fig. 2*B*). We should point out that our measure of accuracy is meaningful only if it is selective, with relatively few false alarms. Because the experiment had multiple targets per trial, one cannot calculate a simple false alarm rate. However, we can quantify errors by dividing the total number of incorrect responses (outside the 800-ms hit window) by the total number of targets, over the course of the experiment for each subject. This error ratio was 0.332 (±0.101 SD) indicating that subjects made only 1 extra button press for roughly every 3 target sentences, or slightly less than 1 per trial.

To ensure that all effects reported herein are not the result of difficulty differences between conditions or the particulars of which talker was attended on a given trial, we analyzed performance across trial type and target properties. Performance did not significantly differ between trial types (pitch:

51.9 ± 3.0% standard error of the mean (SEM), location: 54.1 ± 3.9% SEM) or among the attended talker's pitch or location (data not shown). There were no significant interactions between trial type and temporal progression of performance ($F(9,135) = 1.36$ $P = 0.213$, using 9 nonoverlapping temporal bins). Reaction times were also similar across trial types (pitch: 515.3 ± 5.9 ms SEM, location: 518.8 ± 5.6 ms SEM).

fMRI

Attentional Control during Cue Period

Preparatory attention strongly activated ($P < 0.01$ FDR corrected) a left-dominant fronto-parietal network that showed a large degree of similarity between cue types (Fig. 3, Table 1). These activation maps reflect the signal for each trial type (location or pitch) relative to rest cue trials. Both cues activated inferior frontal gyrus (IFG), DPreCS, intraparietal sulcus (IPS), and SPL relative to rest.

Qualitative inspection of the activation maps shows greater extent and bilaterality for different cue types in different cortical regions. This raised the question of whether these anatomically distinct regions contained biases toward one feature class or another. We therefore performed a ROI analysis on areas showing greater activity to both pitch and location cues relative to rest cues. This is formally substantiated by logically conjoining the pitch and location greater than rest thresholded maps ($P < 0.01$ FDR), and then creating an ROI for each contiguous cortical cluster in the conjunction. The results of this analysis can be seen in Figure 4. IFG showed pitch-cue biased activity, whereas DPreCS and SPL showed location-cue biased activity, and IPS showed no bias. All tests were 2-tailed paired *t*-tests with a threshold of $P < 0.05$.

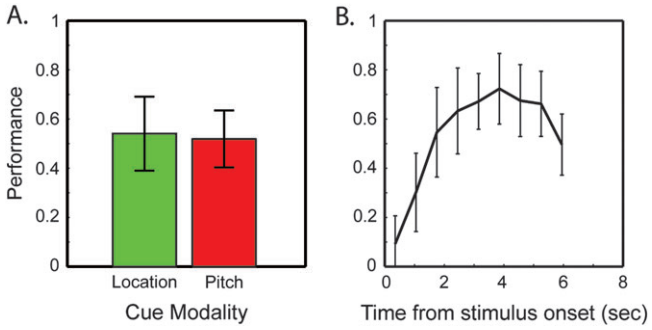

Figure 2. Behavioral data showing effect of cue type on accuracy and the time course of performance within a trial. Difficulty was equal between conditions (*A*), ensuring that imaging results do not reflect effort-related activation. The increase in performance over time (*B*) may reflect streaming segregation buildup, language processing, or both. Discrete points in B represent time-binned data, and error bars show SD.

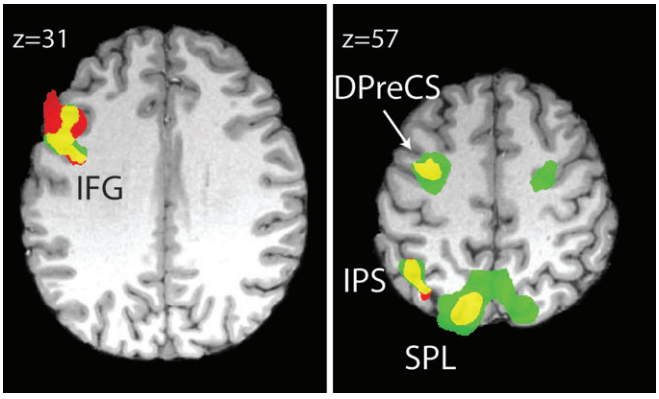

Figure 3. Activation of a left hemisphere dominated, fronto-parietal network during the cue period of both location and pitch trials relative to rest cues. Pitch cue (red) and location cue (green) show a large degree of overlap (yellow) when compared with the rest cue. Activations include IFG, DPreCS, IPS, and SPL ($P < 0.01$ FDR). In this and all subsequent imaging figures, activations are shown on a representative subject's brain.

Attentional Selection during Stimulus Period

Next, we examined attentional modulation and bias during the stimulus period of the trials, when subjects selected one talker from the others. Figure 5 shows the main effect of attention on stimulus processing. Again, red regions reflect activation in pitch stimuli greater than rest stimuli, whereas green regions reflect activation in location stimuli greater than rest stimuli. A thresholded map of $P < 0.01$ FDR reveals a large, contiguous network of activated cortical regions that include bilateral STG and Superior Temporal Sulcus (STS) (but not Heschl's Gyrus), insula, frontal and parietal cortex, the basal ganglia, and cerebellum. These broad networks show an extremely high degree of overlap between condition, and corrected whole-brain contrasts yielded no significant results between the 2 trial types of location and pitch. Some overlap may be due to the object-based nature of attention, where attentional effects spread to any salient feature of an attended object. Another, noncognitive, contributing factor may be that our task proved very difficult in the scanner for some subjects. If the task becomes too difficult, participants may attempt to compensate by attending to all available cues in an effort to understand the sentences. This would have the consequence of washing out the attention-mediated task differences between the 2 conditions for the stimulus period. The cue period however, during which no stimuli are playing, would be relatively unaffected by performance.

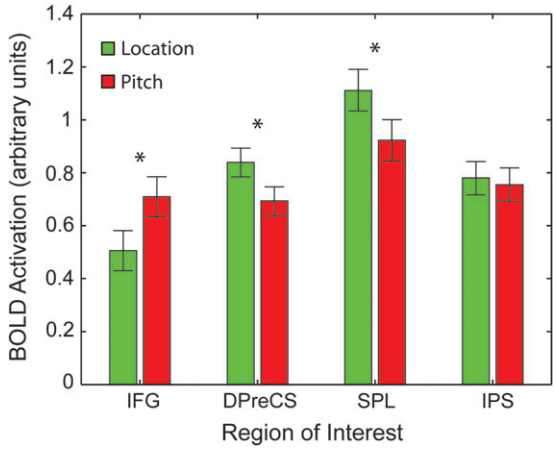

Figure 4. ROI analysis reveals biases in attentional control network for pitch and location. ROIs were defined by overlap between pitch and location cues greater than rest cues (the yellow regions in Fig. 3). IFG shows a stronger activation during pitch cues relative to *rest* cues, whereas SPL and DPreCS show greater activity for location cues. IPS shows no reliable difference. All tests are 2-tailed paired *t*-test (*$P < 0.05$ uncorrected). Error bars show standard error of the difference within subjects.

To ensure that our imaging results for the stimulus period reflected neural activity corresponding to successful attentional selection to a single feature, we selected our 10 best performing subjects (average performance > 50% correct) and ran the analyses again. The main effect of attention greater than rest was qualitatively very similar to the maps shown in Figure 5, which confirmed that our subset of subjects was representative. We were interested in identifying regions among those strongly modulated by attention that show a consistent bias between location- and pitch-based selection. To this end, we created minimum-*t*-score null conjunction tests from the main effects of attention and the difference between location and pitch. For example, to identify regions involved in auditory selection based on pitch we tested the conjunction between pitch greater than rest and pitch greater than location during the stimulus period of the trial. This gives us regions that show activity in pitch trials greater than both location and rest trials. The converse contrast allowed us to identify areas with location greater than either rest or pitch activity. Pitch-biased effects of selective attention were found bilaterally in posterior STS, and right middle STS. Location-biased effects of selective attention were found exclusively in left IPS (Fig. 6, Table 1). This region of IPS is both posterior and lateral to the region found to be active during the cue portion of the trial.

Discussion

In this study, we have shown the differential contributions of attentional control and selection mechanisms in realistic cocktail-party listening with neuroimaging techniques. The use of realistic spatial cues, natural language, and temporally overlapping stimuli makes our findings relevant to real-world situations, which involve high processing load and require selection among competing objects. The catch-trial trial design and careful control of stimulus attributes between conditions allows us to characterize both the attentional control system, which directs attention to a particular location or pitch in the absence of stimuli, and the attentional selection of auditory objects from a complex auditory scene.

Attentional Control

We found that the network responsible for auditory attentional control was active when listeners attended to either pitch or space. Although attentional control for both classes of features activated a highly overlapping, left dominated, fronto-parietal network, portions of this network showed significant activity differences depending on the feature class to which listeners directed their attention. This corroborates observations of cue-period activity during visual attentional control experiments. There is a remarkable correspondence between the areas identified in Figure 3 and those in Giesbrecht et al. (2003) showing the effects of both spatial and color based attentional control in vision. Their group also found greater activity for both cues relative to rest in a left dominated fronto-parietal network. Although direct comparison across studies is difficult, it would appear that both experiments are activating a homologous network. In the visual modality, SPL and DPreCS both showed a robust spatial attention bias and the FG showed color feature biases. Although our findings for spectral feature attention differ from studies in vision, the location of the nonspatial biased attentional control center in each modality corresponds well with known functional anatomy; the bias for visual color features occurs along the ventral visual "what" pathway, whereas bias for auditory spectral features in our language task activates regions near inferior frontal regions linked to language processing. These findings suggest an attentional control system that is well designed for modulating multi-sensory information in circumstances where coordination across modality would prove beneficial, such as spatial attention, yet also maintaining parallel independent pathways for modulation of modality-specific features (Shomstein and Yantis 2004, 2006). It is also worth noting that these findings resemble studies of auditory working memory for location and voice identity (Rama et al. 2004; Arnott et al. 2005). This may be due to common cognitive demands in the 2 tasks. Anticipatory attentional control may often rely on memory-derived representations of location or feature, and working memory tasks typically require sustained, selective attention. In vision at least, there is a high degree of overlap between networks thought to be responsible for attentional control and working memory (LaBar et al. 1999; Corbetta and Shulman 2002), and further investigation in the auditory domain is warranted.

Left IPS was unique among the attentional control areas, in that it was equally active for pitch and location cues relative to rest. This suggests that IPS may be an integrative center that coordinates attention regardless of which class of features is the focus of attention. This is supported by work in the visual domain, which has shown a linear combination of cue

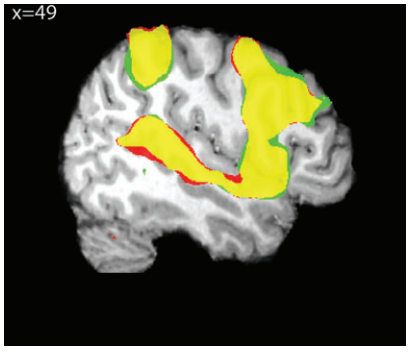

Figure 5. During the stimulus period of the trial, attention produces large changes in a host of areas throughout the brain including auditory and multimodal cortices. As in Figure 4, effects of pitch trials are shown in red and location is shown in green. Note the high degree of overlap (yellow). ($P < 0.01$, FDR).

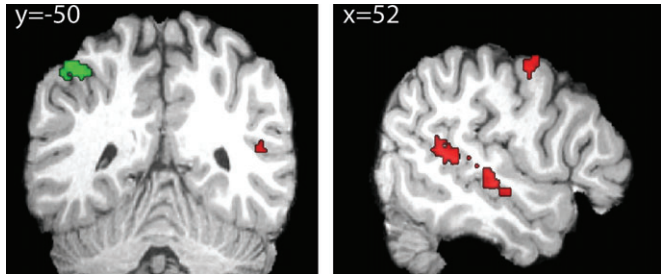

Figure 6. Analysis of the 10 subjects achieving above 50% average performance during the stimulus period reveals regions responsible for attentional selection during a cocktail-party task. Red regions represent areas with greater activation for pitch trials than either rest or location trials and contain multiple locations along STS. Green regions show greater activation for location trials relative to either rest or pitch trials, and contain only IPS. Neither condition preferentially activates any regions along STG. All tests are minimum *t*-score *t*-test against the conjunction null ($P < 0.005$).

Table 1

Region labels, MNI coordinates, uncorrected *P* values, and *t*-scores of all cortical activations for contrasts mentioned in the text

Region	Coordinates (mm)			*P* value	*T*-score
	x	*y*	*z*		
Cue period					
Location > rest					
Left SPL	−18	−74	54	<0.0001	12.98
Left IPS	−34	−50	48	<0.0001	11.35
Left DPreCS	−30	−6	54	<0.0001	10.35
Right DPreCS	24	−8	54	<0.0001	9.44
Left IFG	−40	2	34	<0.0001	7.09
Left inferior temporal gyrus	−58	−58	−12	<0.0001	5.74
Superior frontal gyrus	−4	8	64	<0.0001	5.79
Pitch > rest					
Left IPS	−38	−52	48	<0.0001	9.31
Left SPL	−12	−72	56	<0.0001	9.01
Left IFG	−46	16	28	<0.0001	8.99
Left DPreCS	−32	−2	62	<0.0001	7.93
Stimulus period					
Pitch > rest and location					
Left posterior STS	−52	−38	14	<0.0001	6.51
Right posterior STS	54	−42	10	0.0004	4.84
Right middle STS	50	−20	−6	0.0004	4.83
Right precentral gyrus	54	0	56	0.0007	4.54
Location > rest and pitch					
Right IPS	−38	−48	50	0.0012	4.16

All cue-period results are significant at FDR $P < 0.01$.

information for spatial and nonspatial feature cues in IPS during a cueing task (Egner et al. 2008), and a large body of research that has shown IPS to be critical in integrating information across a number of modalities (Calvert 2001; Miller and D'Esposito 2005; Bishop and Miller 2009). This integration may be critical to perceiving and acting upon objects that contain information distributed across modalities and feature classes within a modality, as patients with damage to IPS report problems maintaining object boundaries in both vision (Friedman-Hill et al. 1995) and audition (Cusack et al. 2000).

Attentional Selection

As evident in Figure 4, an active task in a complex acoustic scene recruits a large network of cortical areas sensitive to attentional selection in a rich environment. In agreement with previous fMRI findings (Petkov et al. 2004), we find no attentional modulation in Heschl's gyrus. This does not mean that neural activity in primary auditory cortex is not modulated by attention; it may be too weakly or differently modulated such that BOLD signal cannot capture it. Nevertheless, this suggests that primary auditory cortex faithfully relays auditory information to higher cortical regions, even in complex auditory scenes, which are then the targets of auditory selective attention in order to overcome later processing bottlenecks.

Also in accordance with previous studies, we have found separate networks that are differentially active for attention to different acoustic feature classes. Yet in contrast to previous studies, which found affects of attention along STG (Zatorre et al. 1999; Ahveninen et al. 2006; Degerman et al. 2006; Salmi et al. 2007), we observed significant bias only for regions in bilateral posterior STS, right middle STS, and left IPS. The earlier results may differ from ours due to the low processing load, with only a single auditory stimulus being played at a time. This suggests that the regions showing greater activity in the current study are the primary targets of selective attention in complex environments during "cocktail-party" listening. One alterative explanation is that the differences among studies may be due to the linguistic nature of our stimuli. For instance, there are a number of previous studies of selective attention to complex stimuli such as overlapping speech (Woods et al. 1984; Hashimoto et al. 2000; Lipschutz et al. 2002; Alho et al. 2003; Nakai et al. 2005). However, these studies were not designed to identify the networks for selective attention based on a particular feature, and their findings more closely resemble regions where we found greater activation for "both" space and pitch. Among speech studies generally, neuroimaging has identified STS as a key region responsible for processing human vocalizations (Belin et al. 2000) and shown that rSTS may play a key role in processing paralinguistic attributes of speech such as pitch (Belin et al. 2002; Kriegstein and Giraud 2004). However, other studies using simple noise stimuli in a change-detection paradigm also activated primarily regions along rSTS (Alain et al. 2001), supporting a broad role for rSTS in processing the spectral features of auditory objects. Likewise with lIPS, there is no evidence to date that suggests that the IPS represents a spatial processing region specialized for speech sounds, making it unlikely that our results are driven by speech-specific processing.

We believe the sum of evidence instead points to rSTS and lIPS representing the activation of selective attention due to a high processing load that characterizes complex auditory scenes. This agrees with behavioral and electrophysiological studies showing that selective attention acts after auditory objects are formed and not on early auditory attributes (Alain and Arnott 2000; Ihlefeld and Shinn-Cunningham 2008), similar to object-based attentional modulation for visual or auditory-visual objects (Busse et al. 2005). It is possible that this form of selective attention acts upon fundamentally different cortical networks than tasks with low processing load, but this deserves further study. Regardless, our results add to these previous findings by identifying the neural substrate responsible for this selection based on different classes of features in complex auditory scenes.

When taken in context of previous work, this study demonstrates that 1) auditory attentional control uses conserved supramodal spatial networks and specialized auditory spectral feature networks to shift attention for an auditory selection task, and 2) that STS and IPS may represent the earliest site of selective attention for complex auditory scenes. Because of the divergence of these findings using complex stimuli from others that have used simple stimuli, future studies of attention might investigate the differential contributions of these attentional networks, particularly in their sensitivity to perceptual load.

Funding

National Institutes of Health: National Institute on Deafness and other Communication Disorders (R01-DC8171 to L.M., T32-DC8072-01A2 to K.H.).

Notes

We thank Qian-Jie Fu for providing the Harvard/IEEE speech corpus. *Conflict of Interest:* None declared.

Address corresponding to Lee M. Miller, UC Davis Center for Mind and Brain, 267 Cousteau Place, Davis, CA 95618, USA. Email: leemiller@ucdavis.edu.

References

Ahveninen J, Jaaskelainen IP, Raij T, Bonmassar G, Devore S, Hamalainen M, Levanen S, Lin FH, Sams M, Shinn-Cunningham BG, et al. 2006. Task-modulated "what" and "where" pathways in human auditory cortex. Proc Natl Acad Sci USA. 103:14608–14613.

Alain C, Arnott SR. 2000. Selectively attending to auditory objects. Front Biosci. 5:D202–D212.

Alain C, Arnott SR, Hevenor S, Graham S, Grady CL. 2001. "What" and "where" in the human auditory system. Proc Natl Acad Sci USA. 98:12301–12306.

Alain C, Izenberg A. 2003. Effects of attentional load on auditory scene analysis. J Cogn Neurosci. 15:1063–1073.

Alho K, Donauer N, Paavilainen P, Reinikainen K, Sams M, Naatanen R. 1987. Stimulus selection during auditory spatial attention as expressed by event-related potentials. Biol Psychol. 24:153–162.

Alho K, Tottola K, Reinikainen K, Sams M, Naatanen R. 1987. Brain mechanism of selective listening reflected by event-related potentials. Electroencephalogr Clin Neurophysiol. 68:458–470.

Alho K, Vorobyev VA, Medvedev SV, Pakhomov SV, Roudas MS, Tervaniemi M, van Zuijen T, Naatanen R. 2003. Hemispheric lateralization of cerebral blood-flow changes during selective listening to dichotically presented continuous speech. Brain Res Cogn Brain Res. 17:201–211.

Arnott SR, Binns MA, Grady CL, Alain C. 2004. Assessing the auditory dual-pathway model in humans. Neuroimage. 22:401–408.

Arnott SR, Grady CL, Hevenor SJ, Graham S, Alain C. 2005. The functional organization of auditory working memory as revealed by fMRI. J Cogn Neurosci. 17:819–831.

Belin P, Zatorre RJ, Ahad P. 2002. Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res. 13:17–26.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature. 403:309–312.

Bishop CW, Miller LM. Forthcoming 2009. A Multisensory cortical network for understanding speech in noise. J Cogn Neurosci. 21: 1790–1805.

Bregman AS. 1994. Auditory scene analysis: the perceptual organization of sound. MIT Press.

Busse L, Roberts KC, Crist RE, Weissman DH, Woldorff MG. 2005. The spread of attention across modalities and space in a multisensory object. Proc Natl Acad Sci USA. 102:18751–18756.

Calvert GA. 2001. Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb Cortex. 11:1110–1123.

Cherry EC. 1953. Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am. 25:975–979.

Clarke S, Bellmann Thiran A, Maeder P, Adriani M, Vernet O, Regli L, Cuisenaire O, Thiran JP. 2002. What and where in human audition: selective deficits following focal hemispheric lesions. Exp Brain Res. 147:8–15.

Corbetta M, Kincade JM, Ollinger JM, McAvoy MP, Shulman GL. 2000. Voluntary orienting is dissociated from target detection in human posterior parietal cortex. Nat Neurosci. 3:292–297.

Corbetta M, Shulman GL. 2002. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci. 3:201–215.

Cusack R, Carlyon RP, Robertson IH. 2000. Neglect between but not within auditory objects. J Cogn Neurosci. 12:1056–1065.

Darwin CJ, Hukin RW. 2000a. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. J Acoust Soc Am. 107:970–977.

Darwin CJ, Hukin RW. 2000b. Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. J Acoust Soc Am. 108:335–342.

Degerman A, Rinne T, Pekkola J, Autti T, Jaaskelainen IP, Sams M, Alho K. 2007. Human brain activity associated with audiovisual perception and attention. Neuroimage. 34:1683–1691.

Degerman A, Rinne T, Salmi J, Salonen O, Alho K. 2006. Selective attention to sound location or pitch studied with fMRI. Brain Res. 1077:123–134.

Egeth HE, Yantis S. 1997. Visual attention: control, representation, and time course. Annu Rev Psychol. 48:269–297.

Egner T, Monti JM, Trittschuh EH, Wieneke CA, Hirsch J, Mesulam MM. 2008. Neural integration of top-down spatial and feature-based information in visual search. J Neurosci. 28:6141–6151.

Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM. 1993. 3D statistical neuroanatomical models from 305 MRI volumes. Proc IEEE-Nucl Sci Symp Med Imaging Conf. 3:1813–1817.

Friedman-Hill SR, Robertson LC, Treisman A. 1995. Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. Science. 269:853–855.

Genovese CR, Lazar NA, Nichols T. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage. 15:870–878.

Giesbrecht B, Woldorff MG, Song AW, Mangun GR. 2003. Neural mechanisms of top-down control during spatial and feature attention. Neuroimage. 19:496–512.

Goodale MA, Milner AD. 1992. Separate visual pathways for perception and action. Trends Neurosci. 15:20–25.

Griffiths TD. 2008. Sensory systems: auditory action streams? Curr Biol. 18:R387–R388.

Griffiths TD, Warren JD. 2004. What is an auditory object? Nat Rev Neurosci. 5:887–892.

Hashimoto R, Homae F, Nakajima K, Miyashita Y, Sakai KL. 2000. Functional differentiation in the human auditory and language areas revealed by a dichotic listening task. Neuroimage. 12:147–158.

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

Hopfinger JB, Buonocore MH, Mangun GR. 2000. The neural mechanisms of top-down attentional control. Nat Neurosci. 3:284–291.

IEEE. 1969. IEEE recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust. AU-17:225–246.

Ihlefeld A, Shinn-Cunningham B. 2008. Disentangling the effects of spatial cues on selection and formation of auditory objects. J Acoust Soc Am. 124:2224–2235.

Jancke L, Buchanan TW, Lutz K, Shah NJ. 2001. Focused and nonfocused attention in verbal and emotional dichotic listening: an FMRI study. Brain Lang. 78:349–363.

Jancke L, Shah NJ. 2002. Does dichotic listening probe temporal lobe functions? Neurology. 58:736–743.

Kawashima R, Imaizumi S, Mori K, Okada K, Goto R, Kiritani S, Ogawa A, Fukuda H. 1999. Selective visual and auditory attention toward utterances-a PET study. Neuroimage. 10:209–215.

Kriegstein KV, Giraud AL. 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage. 22:948–955.

LaBar KS, Gitelman DR, Parrish TB, Mesulam M. 1999. Neuroanatomic overlap of working memory and spatial attention networks: a functional MRI comparison within subjects. Neuroimage. 10:695–704.

Langendijk EH, Bronkhorst AW. 2000. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. J Acoust Soc Am. 107:528–537.

Lavie N. 1995. Perceptual load as a necessary condition for selective attention. J Exp Psychol Hum Percept Perform. 21:451–468.

Lipschutz B, Kolinsky R, Damhaut P, Wikler D, Goldman S. 2002. Attention-dependent changes of activation and connectivity in dichotic listening. Neuroimage. 17:643–656.

Luck SJ, Chelazzi L, Hillyard SA, Desimone R. 1997. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J Neurophysiol. 77:24–42.

Maeder PP, Meuli RA, Adriani M, Bellmann A, Fornari E, Thiran JP, Pittet A, Clarke S. 2001. Distinct pathways involved in sound recognition and localization: a human fMRI study. Neuroimage. 14:802–816.

Miller LM, D'Esposito M. 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. J Neurosci. 25:5884–5893.

Nakai T, Kato C, Matsuo K. 2005. An FMRI study to investigate auditory attention: a model of the cocktail party phenomenon. Magn Reson Med Sci. 4:75–82.

Ollinger JM, Corbetta M, Shulman GL. 2001. Separating processes within a trial in event-related functional MRI: iI. The analysis. Neuroimage. 13:218–229.

Petkov CI, Kang X, Alho K, Bertrand O, Yund EW, Woods DL. 2004. Attentional modulation of human auditory cortex. Nat Neurosci. 7:658–663.

Rama P, Poremba A, Sala JB, Yee L, Malloy M, Mishkin M, Courtney SM. 2004. Dissociable functional cortical topographies for working memory maintenance of voice identity and location. Cereb Cortex. 14:768–780.

Rinne T, Kirjavainen S, Salonen O, Degerman A, Kang X, Woods DL, Alho K. 2007. Distributed cortical networks for focused auditory attention and distraction. Neurosci Lett. 416:247–251.

Salmi J, Rinne T, Degerman A, Salonen O, Alho K. 2007. Orienting and maintenance of spatial attention in audition and vision: multimodal and modality-specific brain activations. Brain Struct Func. 212:181–194.

Shinn-Cunningham BG. 2008. Object-based auditory and visual attention. Trends Cogn Sci. 12:182–186.

Shomstein S, Yantis S. 2004. Control of attention shifts between vision and audition in human cortex. J Neurosci. 24:10702–10706.

Shomstein S, Yantis S. 2006. Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. J Neurosci. 26: 435–439.

Shulman GL, d'Avossa G, Tansy AP, Corbetta M. 2002. Two attentional processes in the parietal lobe. Cereb Cortex. 12:1124–1131.

Shulman GL, Ollinger JM, Akbudak E, Conturo TE, Snyder AZ, Petersen SE, Corbetta M. 1999. Areas involved in encoding and applying directional expectations to moving objects. J Neurosci. 19:9480–9496.

Spieth W, Curtis JF, Webster JC. 1954. Responding to one of two simultaneous messages. J Acous Soc Am. 26:391–396.

Tardif E, Spierer L, Clarke S, Murray MM. 2008. Interactions between auditory 'what' and 'where' pathways revealed by enhanced near-threshold discrimination of frequency and position. Neuropsychologia. 46:958–966.

Warren JD, Griffiths TD. 2003. Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. J Neurosci. 23:5799–5804.

Woods DL, Alain C, Diaz R, Rhodes D, Ogawa KH. 2001. Location and frequency cues in auditory selective attention. J Exp Psychol Hum Percept Perform. 27:65–74.

Woods DL, Hillyard SA, Hansen JC. 1984. Event-related brain potentials reveal similar attentional mechanisms during selective listening and shadowing. J Exp Psychol Hum Percept Perform. 10:761–777.

Zatorre RJ, Mondor TA, Evans AC. 1999. Auditory attention to space and frequency activates similar cerebral systems. Neuroimage. 10:544–554.