Left thalamo-cortical network implicated in successful speech separation and identification

Claude Alain, Karen Reinke, Kelly L. McDonald, Wilkin Chau, Anda Pacurar, and Simon Graham

Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, Canada
Department of Psychology, University of Toronto, Ontario, Canada
Department of Psychology, Southern Illinois University, Carbondale, IL 62901, USA
Imaging Research, Sunnybrook and Women’s College Health Sciences Centre, and Department of Biophysics, University of Toronto, Ontario, Canada

Received 3 November 2004; revised 2 February 2005; accepted 8 February 2005
Available online 7 April 2005

The separation of concurrent sounds is paramount to human communication in everyday settings. The primary auditory cortex and the planum temporale are thought to be essential for both the separation of physical sound sources into perceptual objects and the comparison of those representations with previously learned acoustic events. To examine the role of these areas in speech separation, we measured brain activity using event-related functional Magnetic Resonance Imaging (fMRI) while participants were asked to identify two phonetically different vowels presented simultaneously. The processing of brief speech sounds (200 ms in duration) activated the thalamus and superior temporal gyrus bilaterally, left anterior temporal lobe, and left inferior temporal gyrus. A comparison of fMRI signals between trials in which participants successfully identified both vowels as opposed to when only one of the two vowels was recognized revealed enhanced activity in left thalamus, Heschl’s gyrus, superior temporal gyrus, and the planum temporale. Because participants successfully identified at least one of the two vowels on each trial, the difference in fMRI signal indexes the extra computational work needed to segregate and identify successfully the other concurrently presented vowel. The results support the view that auditory cortex in or near Heschl’s gyrus as well as in the planum temporale is involved in sound segregation and reveal a link between left thalamo-cortical activation and the successful separation and identification of simultaneous speech sounds.

Keywords: Auditory cortex; Speech; fMRI; Segregation; Thalamus; Attention

A “cocktail party” in which many people are talking at the same time is often used to illustrate a fundamental problem faced by the human auditory system, namely, the segregation of concurrent auditory events. In such situations, listeners must extract information from the composite acoustic wave containing all simultaneously active voices in order to discriminate between, and attend to, individual speakers. Although significant progress has been made in identifying the psychoacoustic factors that promote the perception of concurrent acoustic signals, there is no adequate neurobiological account of where the separation and identification of overlapping speech takes place.

The perception of concurrent auditory objects, such as two persons talking at the same time, is thought to involve a complex and widely distributed neural network including both cortical and subcortical components. Single-cell recordings in the ascending auditory pathway show that frequency periodicity, upon which concurrent sound segregation is partly based, is reflected within the patterns of afferent spike trains (Keilson et al., 1997; Palmer, 1990; Sinex and Li, 2002). Recordings of human event-related brain potentials (ERPs) contain an attention-independent component around 150 ms after sound onset, thought to reflect the automatic registration of concurrent auditory objects, and another component with a latency of 400 ms that is present only when participants are required to attend to the stimuli (Alain et al., 2001b; Alain et al., 2002). Recent functional Magnetic Resonance Imaging (fMRI) studies have identified several brain regions involved in sound identification, recognition, and speech intelligibility including the superior temporal gyrus, planum temporale, anterior temporal lobe, and inferior prefrontal cortex (Binder et al., 2000; Jancke and Shah, 2002; Lewis et al., 2004). Although these studies support the involvement of a distributed neural network in the processing of acoustic events, the neural network implicated in speech separation and identification has yet to be defined within the same individuals using the same set of stimuli.
The present study was designed to investigate the neural substrates of speech separation and identification by measuring changes in brain activity, as measured by event-related fMRI, in cases where listeners correctly identified two speech sounds presented simultaneously relative to when only one sound was correctly identified. Participants were presented with a mixture of two phonetically different vowels, which could have the same or different fundamental frequency \(f_0\). On each trial, participants identified the two vowels by sequentially pressing two different keys on a response pad. Tasks such as these are typically very demanding and generate a large number of errors. Such a paradigm is, therefore, ideal for examining brain areas involved in successful speech separation and identification while keeping the stimulus set constant. Given that the ability to understand spoken language is mostly dependent on the left lateralized cortical system (Binder et al., 2000; Zatorre et al., 1992), it was predicted that success in identifying concurrent vowels would recruit cortical regions in the left hemisphere.

**Methods**

**Participants**

A total of 11 right-handed participants whose native language was English were recruited for the present study. Two participants were excluded from the analysis because they performed near ceiling (89 and 95% accuracy, respectively) and consequently there were not enough incorrect trials to be analyzed. Nine participants (4 women and 5 men aged between 21 and 30 years, mean age = 26 ± 3.5 years) formed the final sample. None had any history of hearing, neurological, or psychiatric disorders. Ethical approval and informed consent were obtained according to the guidelines set out by the Baycrest Centre for Geriatric Care, Sunnybrook and Women’s College Health Sciences Centre and the University of Toronto.

**Stimuli and task**

Stimuli were four synthetic steady-state American English vowels: /i/, /ɛ/, /u/, /æ/ (Assmann and Summerfield, 1994). Each vowel was 200 ms in duration including 10 ms rise and fall time (2000 samples at a 10-kHz sample rate). Double-vowel stimuli were created by adding the digital waveforms of two phonetically different vowels. Each pair contained one vowel with an \(f_0\) set at 100 Hz; the \(f_0\) of the other vowel was set at 100, 103, 106, 112, or 126 Hz (i.e., 0, 0.5, 1, 2, or 4 semitones higher). Stimuli were randomized in blocks of 30 trials that included 25 double-vowel stimuli and 5 silent trials (which served as baseline) that contained no vowels and required no response. The interstimulus intervals ranged from 8 to 12 s (i.e., 4, 5, or 6 s prior to the next fMRI acquisition; see fMRI recording and analysis section). Stimuli were presented binaurally at 85 dB SPL via circumaural fMRI-compatible headphones (Silent Scan, Avotec, Stuart, FL), acoustically padded to suppress scanner noise by 25 dB.

Participants were given some practice at the task before entering the scanner to familiarize them with the stimulus–response mappings. Prior to the double-vowel task, participants were presented with each vowel individually (25 trials) and asked to identify the vowel by pressing the corresponding key on a response pad (Lumitouch, Lightwave Technologies, Surrey, BC, Canada). This ensured that participants could accurately identify each vowel when presented individually. None of the participants had any difficulty in identifying the single vowels, and all reached a level of 95% correct or better.

Participants performed six blocks of 30 trials. On each double-vowel trial, participants were required to identify the two vowels composing the sound mixture by sequentially pressing the two corresponding keys on the fMRI-compatible response pad. Participants received no feedback regarding their performance.

**fMRI recording and analysis**

Participants were scanned using a research-dedicated whole-body 3.0-T MRI system (Signa 3T94 hardware, VH3M3 software; General Electric Healthcare, Waukesha, WI) with a standard quadrature bird-cage head coil. Functional imaging was performed to measure brain activation by means of the blood oxygenation level-dependent (BOLD) effect (Ogawa et al., 1990) with optimal signal contrast. Functional scans were acquired by means of a long acquisition interval to minimize signal contamination from the scanner noise (Belin et al., 1999; Hall et al., 1999) using a single-shot T2*-weighted pulse sequence with spiral k-space readout (axial–oblique orientation, clustered acquisition, TR = 10,000 ms, TE = 30 ms, flip angle 80°, effective acquisition matrix = 64 × 64, number of slices = 10, voxel size = 3.125 × 3.125 × 5 mm, slice spacing = 0 mm, FOV = 20 × 20 cm) (Glover and Lai, 1998). Slices were prescribed graphically based on initial sagittal localizer scans, with the middle slice aligned collinearly with the Sylvian fissure. While this did not cover the whole brain, it allowed us to concentrate data collection on the auditory areas of interest. Reconstruction of the raw data was conducted off-line and included gridding as well as correction for magnetic field inhomogeneities and Maxwell gradient terms. For each participant, 32 time points were collected per run (run time: 5 min 20 s). The first run was always a block of single vowels to ensure that participants could reliably identify each vowel individually. This was followed by six blocks of double-vowel trials (total acquisition time: 37 min 20 s). The procedure of executing multiple runs of relatively short duration was adopted to ensure that the participants remained vigilant at both the auditory tasks and the instruction to keep their head still.

Standard high-resolution 3D T1-weighted fast spoiled gradient echo (FSPGR) images (axial orientation, TR = 7.2 ms, TE = 3.1 ms, inversion-recovery prepared TI = 300 ms, flip angle 15°, effective acquisition matrix = 256 × 192, number of slices = 124, voxel size = 0.85 × 0.85 × 1.4 mm, FOV = 22 × 16.5 cm) were obtained before fMRI to register brain structure and function. In addition to 3D T1-weighted images, T1-weighted spin echo images (TR = 500 ms, TE = 14 ms, effective acquisition matrix = 256 × 192, number of slices = 10, voxel size = 0.78 × 0.78 × 5 mm, FOV = 20 × 20 cm) with the same orientation as the functional scans were acquired and used in the registration procedure.

Data preprocessing and analyses were performed using Analysis of Functional Neuroimages software (AFNI version 2.56a) (Cox and Hyde, 1997). The first and second time points in each run, in which transient signal changes occur as brain magnetization reaches a steady state, were excluded from all analyses.

In the preprocessing stage, time series data were spatially coregistered to correct for small head motions. For each run, images acquired at each point in the series were aligned volumetrically, using the 3dvolreg program from AFNI, to a reference image acquired during the scanning session. Alignment parameters were
computed by an iterative weighted least squares fit to the reference image and then applied using 3D Fourier transform interpolation. For all participants, the peak range of head motion was less than 2.0 mm. The co-registration results were also checked visually for additional quality control.

In the analysis stage, the stimulus timing was defined for each participant, for each $f_0$ separation ($\Delta f_0$) and for correct and incorrect trials. Deconvolution using the 3dDeconvolve program from AFNI was performed to model the response for each $\Delta f_0$. The linear regression model for deconvolution incorporated lags of 4, 5, and 6 s beginning at sound onset. The stimulus-input waveform was modeled from the data, but the data itself determined the functional form of the estimated BOLD response, and as a result the shape of this response varied from voxel-to-voxel. The output included the estimated response, along with the statistical significance of the model fitting to the original functional data, for each voxel in the data set. In addition to the regression coefficients, an $F$ statistic, $t$ statistics for each response parameter, and partial $F$ statistics for each stimulus type were computed. The results for all $\Delta f_0$ values were then used for group analysis to test whether the responses varied as a function of $\Delta f_0$. Each $f_0$ separation condition was contrasted with the no-stimulus trial data. Accordingly, for the analysis examining the effects of $f_0$ separation, five activation maps were created for each participant. For the analysis of correct (both vowels identified) and incorrect (only one vowel correctly identified) trials, two activation maps were created for each participant: correct versus no-stimulus trials and incorrect versus no-stimulus trials.

The response magnitude maps were resampled to a $1 \times 1 \times 1$ mm grid, transformed into Talairach space (Cox and Hyde, 1997), and then spatially smoothed with a Gaussian filter of 6-mm full-width-at-half-maximum to account for individual variation of the anatomical landmarks and to increase the signal-to-noise ratio. These last two steps were performed to facilitate the subsequent group analysis, which consisted of a voxel-wise, mixed-model, two-factor ANOVA with participants as a random factor and group analysis, which consisted of a voxel-wise, mixed-model, two-factor ANOVA with participants as a random factor.

Across participants, the detail for each participant was correct versus no-stimulus trials and incorrect versus no-stimulus trials.

The proportion of correctly identified trials was plotted as a function of the difference in fundamental frequency ($f_0$) between the two vowels. The error bars show the standard error of the mean (±SEM) after between-subject variability has been removed as described by Loftus and Masson (1994).

**Results**

**Behavioral results**

Fig. 1 shows the group mean proportion of trials on which both vowels were correctly identified as a function of $f_0$ difference ($\Delta f_0$). Participants performed well above chance even when the two vowels shared the same $f_0$. The increase in $\Delta f_0$ led to moderate, albeit significant, improvement in vowel identification, $F(4,32) = 3.78, P < 0.05$. Pairwise comparisons indicate that only the largest difference in $f_0$ between the two vowels (i.e., 4 semitones) led to significant improvement in vowel identification, $P < 0.01$. There was no difference in identification rate from 0 to 2 semitones.

Across participants, the proportion of trials in which both vowels were correctly identified ranged from 47 to 76% (mean = 60 ± 10%). This level of performance was well above chance level, which was 17%. To determine whether some of the vowel pairs were more difficult to identify than others, the distribution of the six double-vowel stimuli was examined between the correct and incorrect trials. The analysis revealed a greater proportion of double-vowel stimuli including the vowel /i/ for the correct (83%) than incorrect (17%) trials, $F(1,8) = 142.65, P < 0.001$. The analysis of the behavioral data also indicated that participants had little difficulty in identifying at least one of the two vowels composing the sound mixture (mean accuracy = 99%, chance level = 83%). In the trials where $f_0$ differed, participants reported the vowel with the 100-Hz fundamental (51%) more often.

---

1. The linear trend in log frequency was significant, $F(1,8) = 7.26, P < 0.05$, as well as a cubic component, $F(1,8) = 15.00, P < 0.005$. 

**Fig. 1** Proportion of trials in which both constituents of double-vowel stimuli were correctly identified plotted as a function of the difference in fundamental frequency ($f_0$) between the two vowels. The error bars show the standard error of the mean (±SEM) after between-subject variability has been removed as described by Loftus and Masson (1994).
than vowels with $f_0$ set at 103 Hz (13%), 106 Hz (16%), 112 (13%), or 126 Hz (7%), $F(4,32) = 193.16, P < 0.001$.

fMRI results

The first analysis of fMRI data, illustrated in the two upper panels of Fig. 2, examined the group-averaged pattern of activation produced by two brief vowels presented simultaneously relative to silence. Fig. 2A shows those areas specific to the correct identification of both vowels while Fig. 2B shows activity related to the correct identification of one vowel only. In both cases, brief speech sounds evoked fMRI signal in thalamus bilaterally, as well as in the left and right auditory cortex in Heschl’s gyrus and the surrounding areas along the superior temporal plane. In the left hemisphere, the successful processing of double-vowel stimuli was also associated with enhanced BOLD response in the anterior temporal lobe, the inferior frontal cortex, and the supramarginal gyrus.

To identify the brain regions preferentially involved in the successful identification of both vowels, BOLD responses were compared when participants successfully separated and identified both vowels as opposed to when only one of the two vowels was correctly recognized, regardless of $\Delta f_0$. The successful segregation and identification of both vowels was associated with greater activation in left thalamus, left superior temporal gyrus, and Heschl’s gyrus extending posteriorly to auditory association cortices on the supratemporal plane (Fig. 2C). The peak activation observed in the auditory cortex was located in lateral and posterior portions of Heschl’s gyrus on the probabilistic maps developed by Penhune et al. (1996) (Fig. 3). Increased BOLD signal was also found in a network of regions in the left posterior association cortex including the superior temporal gyrus, cuneus, and inferior parietal cortex (Table 1).

Previous research has shown that participants’ likelihood of correctly identifying both vowels increases with increasing $\Delta f_0$ between the two vowels (Assmann and Summerfield, 1994; Chalikia and Bregman, 1989). In the present study, the likelihood of correctly identifying both vowels also increased as a function of $\Delta f_0$ between the two vowels. Thus, it is possible that the difference in activation maps between correct and incorrect trials could be partly confounded by $\Delta f_0$, with larger separation leading to greater activation. To rule out this explanation, changes in BOLD response were examined as a function of $\Delta f_0$ using a linear and cubic trend as well as using the behavioral data. The linear trend was not significant nor were the cubic trend and behavioral contrast, indicating that $\Delta f_0$ had little impact on the BOLD signal. When the threshold value was lowered to $P < 0.01$ uncorrected, a significant linear trend activation was found in the superior and middle temporal gyrus (not shown). However, these activations did not overlap with those observed in the contrast between correct and incorrect trials. The impact of $\Delta f_0$ was further investigated by examining the BOLD responses in auditory cortices. A region of interest analysis on auditory cortex confirmed the enhanced BOLD response when participants successfully identified the two vowels and revealed a comparable pattern for all $\Delta f_0$ levels (Fig. 4). The enhanced activity was significant only in the left auditory cortex, partly due to greater within-subject variability in BOLD responses from the right auditory cortex. This indicates that the enhanced BOLD signal for correct trials cannot be accounted for by greater $\Delta f_0$.

Another possibility is that some vowel pairs are easier to discriminate than others. As mentioned earlier, a close examination of the behavioral data revealed that participants were more likely to identify both vowels when the vowel /i/ was included in the pair. This is likely due to the fact that the vowel /i/ comprises acoustic

![Fig. 3.Overlap between the group mean focal point of activation (indicated by a small blue +) and the probabilistic map of the primary auditory cortex determined by Penhune et al. (1996). Note that the map is in MNI coordinates superimposed on the MNI template.](Image 361x586 to 489x724)
energy in the high-frequency range thereby contributing to observed differences in BOLD signal between correct and incorrect trials. To rule out this possibility, brain activation was compared for stimulus pairs that included the vowel /i/ and for those that did not. The results of this contrast are shown in Fig. 5 and reveal no significant activation in thalamus or auditory cortex along the superior temporal plane. However, there was enhanced BOLD signal in the right and left claustrum, posterior portion of the left superior and middle temporal gyrus, left angular gyrus, and left precuneus when the double-vowel stimuli included the vowel /i/. Thus, it appears that the enhanced activation of the thalamo-cortical network related to volunteers’ success at identifying the two vowels of a pair is not related to the dissimilarity of the formant frequencies of the two vowels.

Discussion

Auditory streaming is a critical stage in auditory perception that permits listeners to identify the various sound sources contained in the incoming acoustic wave. Here, using fMRI, enhanced brain activity was observed in left thalamus, superior temporal gyrus,
The enhanced fMRI signal in auditory cortex near or in Heschl’s gyrus and in the planum temporale when participants successfully identified two concurrently presented vowels as opposed to when only one of the two vowels was recognized. The results suggest that success in speech separation and identification is contingent upon the activation of this left thalamo-cortical network. This finding extends previous neuro-psychological studies that show impaired language processing following damage to thalamo-cortical projections (Kaga et al., 2000; Radanovic and Scaff, 2003). Moreover, the present findings complement previous electrophysiological studies by identifying the brain regions that may underlie changes in scalp recorded brain activity associated with perception and identification of concurrent auditory objects (Alain et al., 2001b; Alain et al., in press; Dyson and Alain, 2004).

The left thalamo-cortical activation may reflect higher levels of attention when both vowels were correctly identified as opposed to when only one vowel was recognized. For instance, participants’ attention may have wandered from trial to trial because of the long inter-trial intervals used in the present study. Hence, correct trials may simply indicate that participants’ attention was more focused on the task at hand and may have resulted in higher BOLD responses. Previous fMRI studies of auditory selective attention have found greater BOLD signals in auditory cortex for attended than unattended stimuli (Grady et al., 1997; Jancke et al., 2001; Petkov et al., 2004; Woodruff et al., 1996). However, there is an important difference between these studies and the current one. In previous studies, the effects of attention were isolated by contrasting BOLD signal elicited by attended and unattended sounds using demanding intra- or inter-modal tasks. In the present study, all stimuli were attended (i.e., task-relevant) and participants were at ceiling in identifying at least one of the two vowels (99% accuracy). Therefore, it is unclear whether such a putative difference in the amount of attention allocated on correct and incorrect trials would be reflected in the left thalamo-cortical activation. Nonetheless, it remains possible that the enhanced activation associated with success is related to top–down effects on brain mechanisms involved in segregation. Previous ERP studies have shown that concurrent sound segregation involves both automatic and schema-driven processes (Alain and Izenberg, 2003; Alain et al., 2001b; Petkov et al., 2004; Woodruff et al., 1996). The present data also extend previous findings in auditory fMRI studies by showing that brief speech signals (200 ms in total duration) evoked robust and reliable signal changes in BOLD responses. Moreover, these changes were observed in a widely distributed network of regions including the thalamus and Heschl’s gyrus extending to the anterior temporal lobe and inferior prefrontal cortex. This pattern of activation has consistently been observed in tasks involving sound identification and is consistent with the dual pathway model of auditory processing (Alain et al., 2001a; Arnott et al., 2004; Rauschecker, 1998; Rauschecker and Tian, 2000).

The neural underpinning of auditory scene analysis is attracting considerable attention in part because of the realization that deficits in parsing auditory events may contribute to speech perception problems often observed in older adults (Alain et al., 2001c; Grimault et al., 2001; Grube et al., 2003) and in individuals with dyslexia (Helenius, 1999 #176; Sutter et al., 2000). Griffiths and Warren (2002) proposed that the explicit registration and discrimination of auditory objects require computations subsequent to the initial processing of sounds in the ascending pathway and primary auditory cortex, and suggest that these computations might be carried out in the planum temporale. According to this model, neural activity in the planum temporale should vary with listeners’ likelihood of hearing one or two auditory objects. Some support for this model was found in the present data. For example, the fMRI signal was greater in the planum temporale when listeners successfully separated and identified two concurrent speech signals. However, it is unclear how the model would account for the difference in BOLD signal observed in regions other than the planum temporale. For example, if the activity in thalamus and auditory cortex solely reflects the automatic registration of acoustic data, then one would have expected comparable activity in these brain regions for correct and incorrect trials. The present study appears more compatible with a model in which the midbrain and the auditory cortex near or in Heschl’s gyrus play an active role in breaking apart the acoustic wave into separate objects and may reflect neural interactions between higher (e.g., planum temporale) and lower (e.g., primary auditory cortex and thalamus) auditory centers. Until now, this position has been supported primarily by ERP data, showing changes in brain responses as early as 30 ms after sound onset (Dyson and Alain, 2004) and suggesting an automatic registration of acoustic cues which lead to sound segregation. The results of the present study indicate that the thalamus and primary auditory cortex not only automatically register the acoustic information needed for sound segregation but also carry out some of the computation that eventually leads to successful speech separation and identification.
Acknowledgments

This work was supported by grants from the Canadian Institute of Health Research and the Natural Sciences and Engineering Research Council of Canada. We wish to thank B. Dyson and J. Snyder for their comments on the manuscript and valuable discussion. We are particularly indebted to Peter Assmann and Quentin Summerfield for providing the vowel stimuli and to Virginia Penhune and Robert Zatorre for providing the statistical maps of the primary auditory cortex.

References


