Detection of Sounds in the Auditory Stream: Event-Related fMRI Evidence for Differential Activation to Speech and Nonspeech

Athena Vouloumanos, Kent A. Kiehl, Janet F. Werker, and Peter F. Liddle

Abstract

■ The detection of speech in an auditory stream is a requisite first step in processing spoken language. In this study, we used event-related fMRI to investigate the neural substrates mediating detection of speech compared with that of nonspeech auditory stimuli. Unlike previous studies addressing this issue, we contrasted speech with nonspeech analogues that were matched along key temporal and spectral dimensions. In an oddball detection task, listeners heard nonsense speech sounds, matched sine wave analogues (complex nonspeech), or single tones (simple nonspeech). Speech stimuli elicited significantly greater activation than both complex and simple nonspeech stimuli in classic receptive language areas, namely the middle temporal gyri bilaterally and in a locus lateralized to the left posterior superior temporal gyrus. In addition, speech activated a small cluster of the right inferior frontal gyrus. The activation of these areas in a simple detection task, which requires neither identification nor linguistic analysis, suggests they play a fundamental role in speech processing. ■

INTRODUCTION

In the acoustic chaos of the external world, one sound to which humans attend effortlessly and automatically is spoken language. Do speech signals trigger different neural processors than do other environmental sounds? In this study, we addressed this question by investigating the neural substrates mediating the initial processing of speech. We examined cortical activation during listeners' detection of rare stimuli in a sound sequence by comparing the detection of a speech stimulus to that of carefully matched nonspeech stimuli.

Processing auditory input as speech is a crucial first step before further linguistic analysis (e.g., phonetic, semantic, or syntactic) can be performed. One class of theories holds that speech, like all sounds, is initially processed by general psychoacoustic mechanisms (e.g., Cole & Jakimik, 1980). Another class of theories claims that language processing involves specialized linguistic mechanisms (e.g., Chomsky, 1986, 2000). According to strong modularity theories, speech is instantaneously shunted into a different processing pathway than other acoustic stimuli (e.g., Liberman, 1996; Fodor, 1983). Evidence from duplex studies supports this hypothesis: When the steady-state vocalic base of a syllable is presented to one ear, and rapidly changing formant transitions are presented to the other, subjects simultaneously hear speech (integrating the transitions with the base to perceive a consonant–vowel syllable) and nonspeech (the formant transitions alone sound like chirps) (Whalen & Liberman, 1987). Duplex perception therefore suggests that different processors act in parallel on the same auditory input to perform distinct, but simultaneous, operations. This raises the possibility that unique neural substrates mediate these parallel processes.

Despite extensive studies using a number of different approaches, the neural substrates involved in language processing are incompletely identified. As early as the 19th century, studies on aphasia revealed a functional asymmetry in that language processing relies preferentially on the left hemisphere. Divided along a crude dichotomy, receptive language depends on the posterior area of the superior temporal gyrus (STG), or Wernicke's area, while productive language relies on the inferior frontal lobe, or Broca's area (for a review, see Damasio & Geschwind, 1984). This functional asymmetry is complemented by a structural asymmetry favoring the left hemisphere, predominantly that of the temporal regions associated with language functions (Galaburda, Sanides, & Geschwind, 1978; Geschwind & Levitsky, 1968). Recent neuroimaging studies on language processing have suggested a more distributed cortical network whose extent and pattern of activation vary across different studies (for a relevant discussion on the variability of substrates activated during phonetic processing, see Démonet, Fiez, Paulesu, Petersen, & Zatorre, 1996;

University of British Columbia

Poeppel, 1996). There is still disagreement, however, about the ''precise'' neural substrates that perform specific linguistic operations. This arises in part from the limited resolution of neuroimaging techniques, and in part from the inherent difficulty in isolating one or another aspect of the linguistic process (e.g., semantic, syntactic, phonetic) from the other mental operations, linguistic and nonlinguistic (e.g., attention, memory), that are simultaneously performed.

The processing of sound as speech is fundamental to all levels of analysis of spoken language. In contrast to higher-order linguistic operations, this perceptual step is more easily isolated through careful manipulation of the physical properties of the signal. By honing in on the functional neuroanatomy of speech perception, neuroimaging studies provide evidence for the contribution of different neural substrates in the steady-state processing of speech. Neuroimaging studies comparing speech to simple nonspeech foils such as noise bursts (Binder et al., 1994, 2000; Zatorre, Evans, Meyer, & Gjedde, 1992) or pure tones in passive listening (Celsis et al., 1999; Binder, Frost, Hammeke, Rao, & Cox, 1996, Binder et al., 2000; Fiez et al., 1995), and active decision-making tasks (Binder et al., 1996; Fiez et al., 1995; Démonet et al., 1992) generally reveal bilateral activation to both speech and nonspeech in the STG, with some areas in the left STG significantly more activated by speech stimuli (Binder et al., 1996; Démonet et al., 1992; Zatorre et al., 1992). The difference in activation between speech and nonspeech seems to be somewhat more pronounced in the active tasks (Binder et al., 1996; Fiez et al., 1995).

However, tones and white-noise bursts are unlike speech on many spectral and temporal dimensions and, as a result, they are arguably imperfect nonspeech controls. In more recent studies, researchers have used nonspeech foils that are more closely matched in terms of the temporal and spectral properties that characterize speech. One approach has been to use reversed speech, which is acoustically matched in terms of duration, amplitude, and spectral properties, but lacks the distinct temporal attributes of speech. Here, both isolated word tokens (Binder et al., 2000; Hickok, Love, Swinney, Wong, & Buxton, 1997; Price et al., 1996) and entire reversed sentences (Wong, Miyamoto, Pisoni, Sehgal, & Hutchins, 1999) have been contrasted. Another approach has been to use signal-correlated noise, which preserves the amplitude envelope, tempo, rhythm, and syllabicity of speech, but lacks spectral information (Mummery, Ashburner, Scott, & Wise, 1999). One recent study has compared speech to musical nonspeech counterparts of systematically varying complexity matched in duration and amplitude (Benson, Whalen, Clark, & Liberman, 2000). Results from these studies using more closely matched stimuli are more heterogeneous, with authors reporting speech-specific activation in the left superior temporal sulcus (STS) either posteriorly (Bensen et al., 2000; Mummery et al., 1999) or anteriorly (Price et al., 1996), in the left posterior middle temporal gyrus (MTG) (Price et al., 1996), bilaterally in the ventral STS and STG (Binder et al., 2000; Mummery et al., 1999) and the right supramarginal gyrus (Wong et al., 1999).

The heterogeneity of these findings might stem from the fact that the nonspeech foils were matched to speech either spectrally or temporally, but not both. As of yet, no study has compared speech with nonspeech controls that are matched in both spectral and temporal attributes, despite the fact that these dimensions together convey the identifying characteristics of natural speech. To contrast speech with a nonspeech stimulus that preserves many of the spectral and temporal characteristics of speech without actually sounding like speech, we created complex sine wave analogues. These sine wave analogues consist of time-varying sinusoidal waves that track the resonant center frequencies of natural speech and reproduce the changes in these frequency peaks across time. While eliminating characteristics of the voicing source, the broader band formant information, and parts of the harmonic spectrum, these analogues preserve the critical frequency and temporal information of speech. The overall pattern of change in these energy peaks resembles the resonance changes produced by the human vocal tract when articulating speech (Remez, Rubin, & Pisoni, 1983).

Sine wave analogues are the ambiguous figures of the speech world. The defining characteristics of speech are so well preserved in these analogues that human listeners can be led into perceiving them as either speech or nonspeech. In a classic series of speech perception studies, Remez, Rubin, Pisoni, and Carrell (1981) demonstrated that sine wave analogues of ''continuous'' speech could be perceived as speech, allowing listeners to recover the message. Yet, this perception depends critically upon the listener's expectations; listeners who were not instructed to expect speech rarely heard the analogues as such (Remez et al., 1983). More recent studies show that sine wave analogues of ''isolated'' words are even less likely to be heard as speech (Remez, Pardo, Piorkowski, & Rubin, 2001).

As nonspeech counterparts, our use of sine wave analogues of isolated nonsense words is therefore ideal: While the fidelity of the sine wave analogues to the speech signal is such that analogues can be processed as speech under certain experimental conditions, by presenting analogues of ''nonsense'' words in isolation to naive listeners, we ensured that our analogues were not perceived as speech.[1]

To date, most previous studies have imaged brain function during speech perception tasks using blocked design tasks (for notable exceptions, see Hickok et al., 1997, and more recently Fiez & McCandliss, 2000). The block design is effective at describing differences in steady-state processing of speech versus nonspeech.

To observe neural activation in response to the detection of a stimulus in the auditory stream, the event-related design is more appropriate (for a discussion of the relative merits of using event-related designs, see D'Esposito, Zarahn, & Aguirre, 1999; for the specific implementation of this design to auditory tasks, see Belin, Zatorre, Hoge, Evans, & Pike, 1999; Hickok et al., 1997). This type of presentation allows modeling of the hemodynamic response to each individual stimulus presentation, thus offering a smaller and more precise window into the initial processing of speech. Moreover, it reduces the effects of possible confounds such as habituation or anticipation (Dale, 1999; Rosen, Buckner, & Dale, 1998).

In this study, we used an event-related fMRI design to investigate the neural substrates activated when a listener detects speech in comparison to a spectrally and temporally matched nonspeech stimulus (see Figure 1). In a secondary comparison we contrasted cortical activation elicited by these complex stimuli with that elicited by simple tones. Because this task only requires listeners to indicate when they detect a stimulus that is different from the background, it does not require in-depth analysis of the signal. Differences in the patterns of neural activation to speech versus complex nonspeech analogues should thus reflect the operation of the distinct processors activated during listeners' initial processing of the different stimuli.

RESULTS

Behavioral Performance

Accuracy for speech and nonspeech detection was near perfect, with participants attaining an average score of 99.6% correct motor responses for speech, and an average score of 99.8% correct for nonspeech. Participants' performance accuracy in the tone detection run was equally high with an average score of 99.6% for high tones and 100% for low tones.

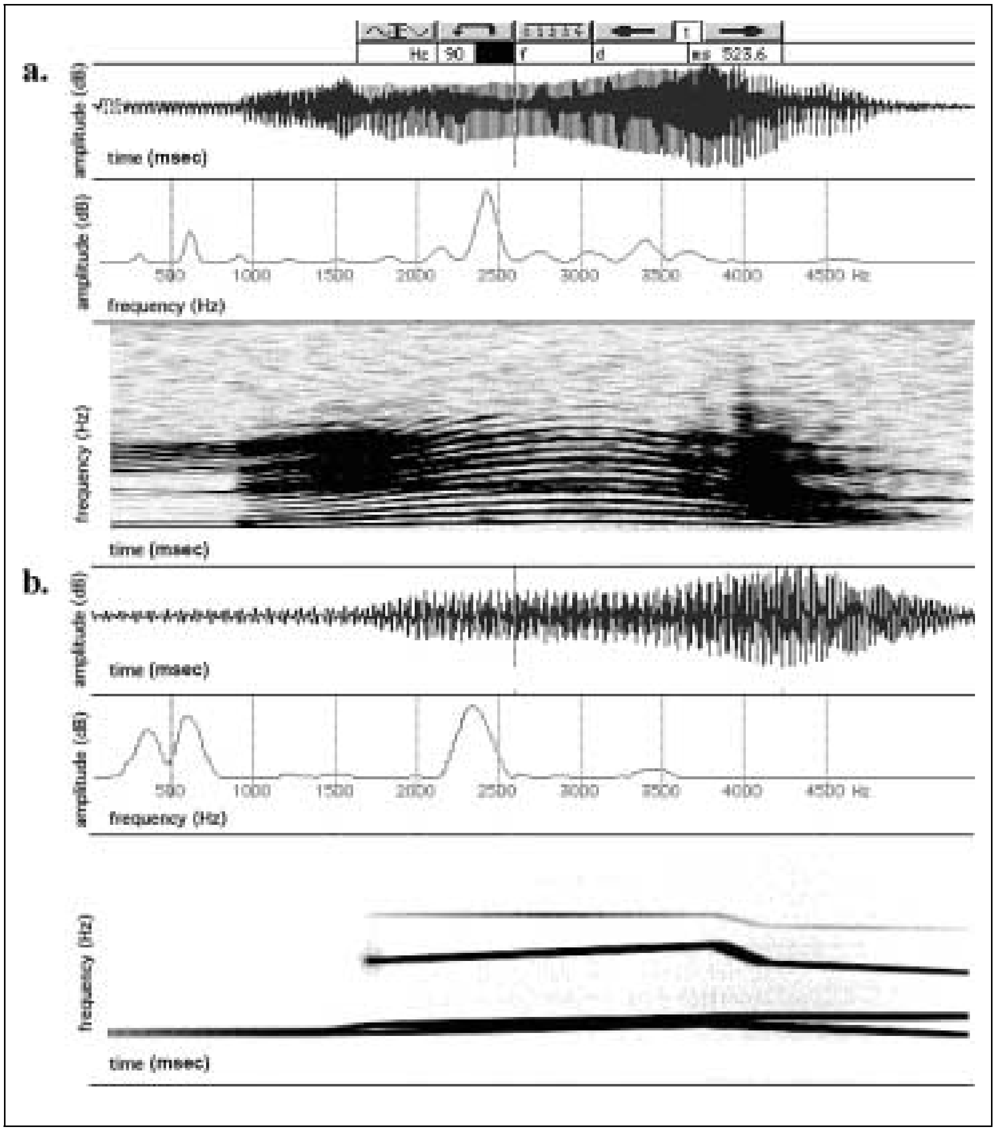

Figure 1. Speech (a) and complex nonspeech (b) stimuli. Similarities between the two types of stimuli are illustrated in waveform diagrams, spectra depicting the relative amplitudes of different frequencies, and spectrograms showing changes in frequency across time.

Participants responded fastest to speech (M = 314 msec, SE = 10), slower to high (M = 326 msec, SE = 21) and low tones (M = 328 msec, SE = 18), and slowest to complex nonspeech (M = 337 msec, SE = 13). A series of paired two-tailed t tests revealed that the only significant difference was between speech and complex nonspeech, $t(1, 14) = 2.71$, $p < .02$. No other comparisons were significant.

Cortical Activation

Speech Versus Complex Nonspeech

Speech stimuli elicited significantly greater activation than complex nonspeech stimuli in the STG and MTG (see Figure 2 and Table 1a-i). The individual hemodynamic responses confirm that activation was present in all listeners (illustrated in Figure 3). The differential activation of the MTG was bilateral, but the extent of differential activation was greater in the left hemisphere (45 vs. 24 voxels exceeding a cluster-level height threshold of $z = 4.63$, $p < .05$ corrected). A one-tailed paired t test run with the individual contrasts within the omnibus fixed-effects analyses was used to isolate the number of voxels activated for each listener. This analysis confirmed that the hemispheric advantage was in the expected direction, but did not reach significance (LH: M = 67 voxels, SE = 15, vs. RH: M = 51, SE = 9, $t(1, 14) = 1.156$, $p = .13$). There was a marked asymmetry in the topography of STG activation between hemispheres: The area of differential activation in the left hemisphere focused around a peak in the posterior STG ($y = -44$), while that in the right hemisphere was centered anteriorly and ventrally around the middle STG ($y = -16$). In addition to the temporal lobe activation, speech stimuli, but not complex nonspeech, activated a small cluster in the right inferior frontal gyrus (IFG; see Table 1a-i). There were no areas that were more activated for complex nonspeech relative to speech. An exploratory random-effects analysis confirmed the robustness and generalizability of these results (see insert in Figure 2 and Table 1a-ii). This more stringent analysis confirmed the bilateral activation of the MTG, and highlighted the degree of left hemisphere lateralization. Differential activation of the STG was observed only in the left hemisphere, in the posterior area

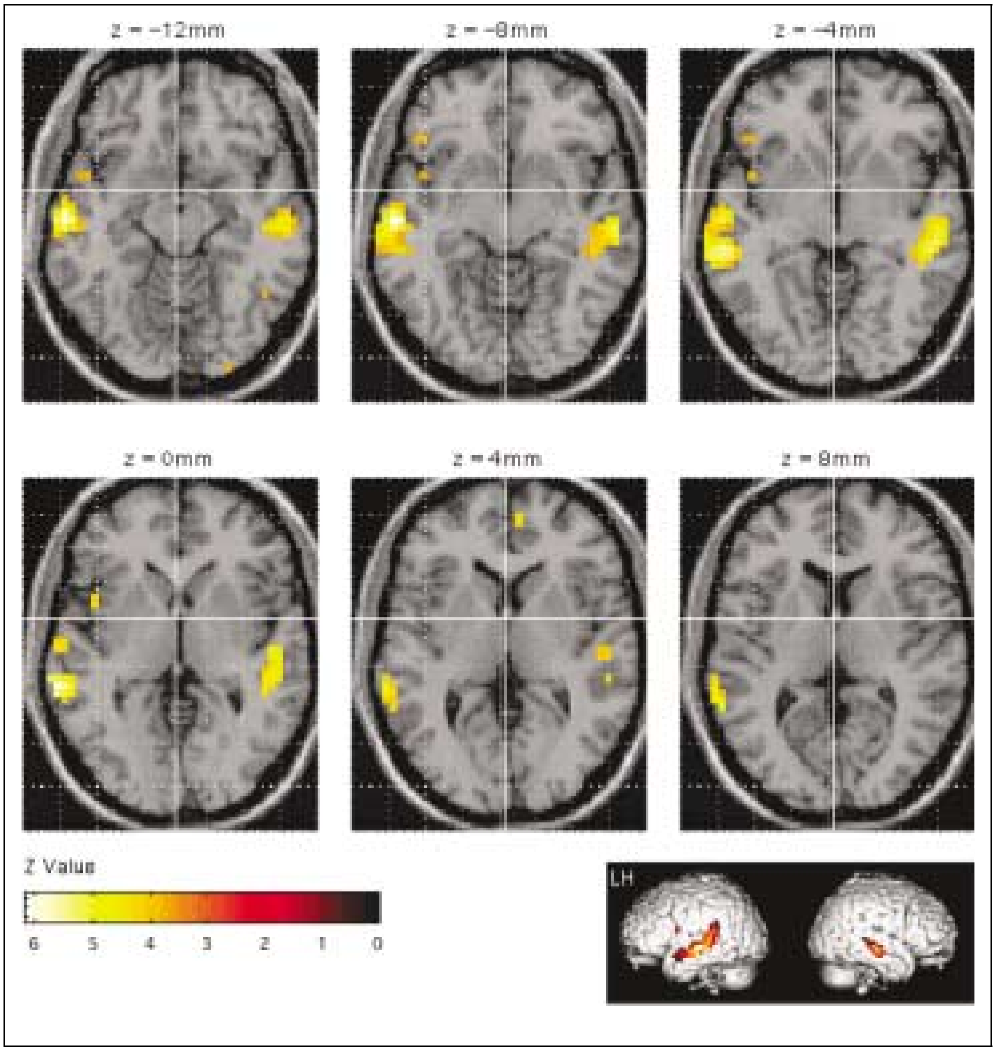

Figure 2. Speech versus complex nonspeech. Axial images illustrating cortical activation to speech relative to complex nonspeech are shown at 4-mm intervals (fixed-effects analysis; display threshold $z = 3.72$; left hemisphere is on the left; color bar indicates corresponding z score). Insert: Cortical surface rendering of areas activated for speech versus complex nonspeech using a random-effects analysis (display threshold $z = 3.72$; color bar indicates corresponding z score).

Table 1. Areas of Significant Activation for Three Comparisons of Interest

		Talairach Coordinates (mm)			
Comparison of Interest	*Region*	*x*	*y*	*z*	*z Score*
(a-i) Speech versus complex nonspeech	L MTG	−60	−16	−8	6.17***
	L MTG	−64	−36	0	5.76***
	L STG	−64	−44	12	4.72*
	R MTG	56	−28	−4	5.32**
	R MTG	52	−20	−16	5.31**
	R STG	52	−16	0	4.63*
	R IFG	40	24	16	4.66*
(a-ii) Speech versus complex nonspeech (random-effects)	L STG	−64	−32	4	5.07**
	L MTG	−60	−20	−8	4.83*
	R MTG	52	−20	−8	4.80*
(b) Speech versus simple tones	L MTG	−60	−16	−8	9.49****
	L MTG	−64	−24	4	8.64****
	L STG	−44	−32	8	5.73****
	R MTG	56	−12	−12	9.09****
	R MTG	56	−28	−4	8.16****
	R IFG	48	16	24	5.52***
	R STG	52	−48	8	5.37**
	R Insula	32	24	12	4.70*
(c) Complex nonspeech versus simple tones	L STG	−64	−24	4	6.44****
	L STG	−60	−16	−4	5.48***
	R MTG	56	−12	−12	6.39****
	R MTG	60	−28	−4	4.94*
	R STG	64	−16	0	4.87*

(a) Speech compared with complex nonspeech sounds using (i) a fixed-effects analysis and (ii) a random-effects analysis, (b) speech compared with simple tones, and (c) complex nonspeech compared with simple tones. Coordinates (x, y, z) are reported in the modified Talairach space used by SPM99. L = left hemisphere; R = right hemisphere.

*p < .05.

**p < .01.

***p < .001.

****p < .0001.

classically associated with receptive language. The activation observed in the right inferior frontal cluster did not survive the random-effects analysis.

Complex Stimuli Versus Simple Tones

The relative cortical activation to speech and complex nonspeech was compared with that elicited by the simple tones. Compared with simple tones, speech activated the STG and MTG bilaterally (see Figure 4a and Table 1b). The extent of this differential activation was greater than that observed when speech was compared to complex nonspeech. Compared to simple tones, complex nonspeech activated bilateral foci in the STG and MTG, which were smaller in extent and in magnitude than the differential activation to speech (see Figure 4b and Table 1c). There were no areas that were relatively more activated for simple tones compared to either speech or to complex nonspeech.

DISCUSSION

The present study demonstrates that even in a simple detection task, speech elicits greater and topographically different cortical activation than complex nonspeech analogues. Specifically, the detection of speech activated the MTG bilaterally, a unique locus in the left posterior STG in the vicinity of Wernicke's area, and a small locus in the right IFG (see Figures 2 and 4a). This pattern of results suggests commonalities in the neural substrates processing complex auditory stimuli, as well as some degree of functional specialization for speech even at this early processing stage.

Recruitment of Classic Receptive Language Areas: A Specialization for Speech Detection From the Initial Stages of Processing

When compared with complex nonspeech, the detection of speech elicited activation in classic receptive language areas along the Sylvian fissure, including the auditory association cortex (Brodmann's area (BA) 22) of the left STG, the posterior part of which is classically referred to as Wernicke's area. In addition, speech elicited differential activation bilaterally in the MTG (BA 21/22). There was a trend for more extensive activation to speech in the left MTG compared with this area's right hemisphere homologue. This greater extent of differential activation in the left temporal lobe, and the unique locus of differential activation in the posterior STG, suggest a left hemisphere lateralization for speech processing.

The bilateral, but left hemisphere-weighted, activation of the temporal lobes that we observed during speech detection is consistent with the results of many studies directly comparing speech and nonspeech processing. Most studies have reported both bilateral cortical activity, as well as loci of activation lateralized to the left hemisphere. Bilateral activation of the temporal lobes has been reported in comparisons of speech with noise bursts (Binder et al., 1994, 2000; Zatorre et al., 1992), signal-correlated noise (Mummery et al., 1999), musical nonspeech counterparts (Benson et al., 2000), and reversed speech (Binder et al., 2000). Many of these

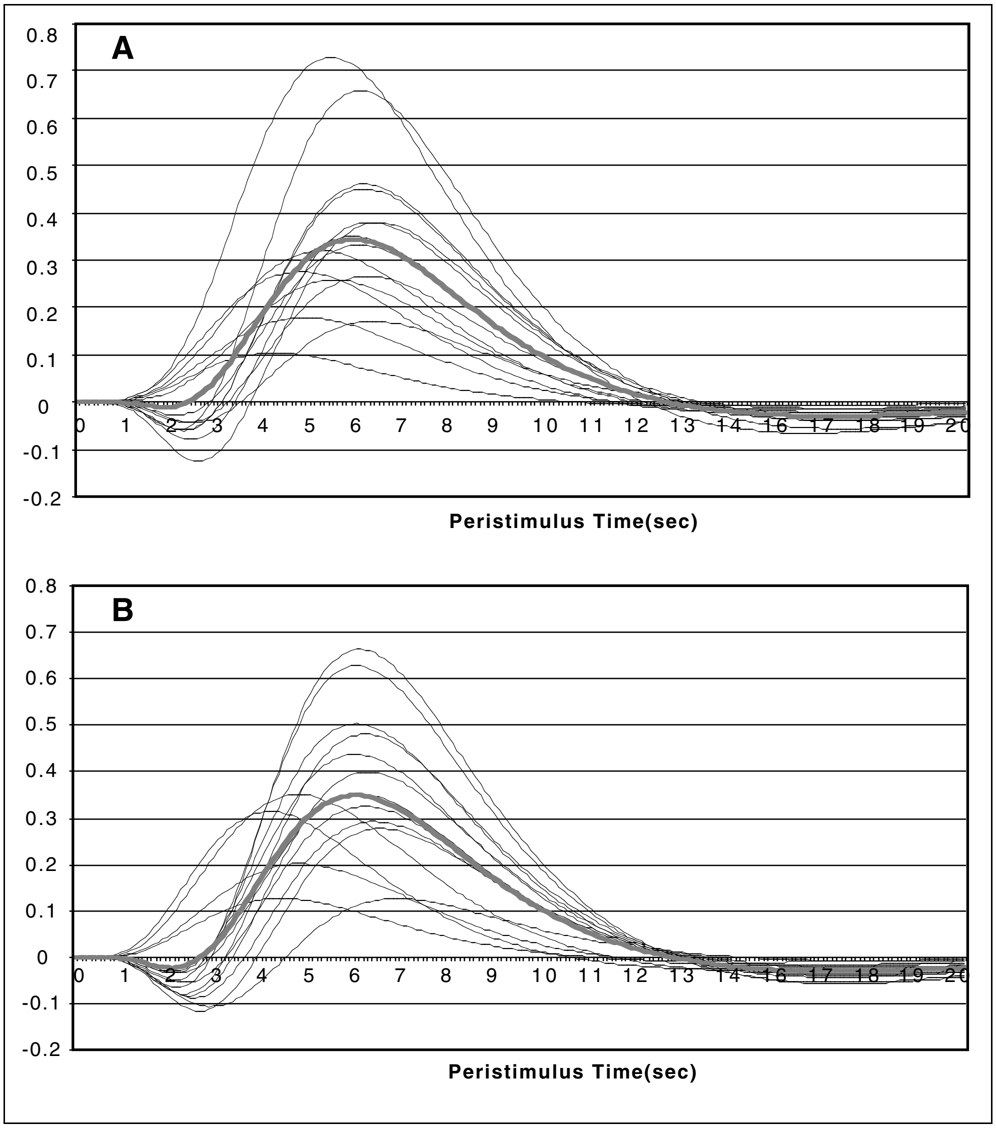

Figure 3. Modeled hemodynamic response to speech stimuli for individual listeners from a voxel of peak activation in the: (A) left MTG (modified Talairach coordinates: −60, −16, −8) and (B) left posterior STG (modified Talairach coordinates: −64, −44, 12) (mean response in **bold**). Hemodynamic responses are plotted in arbitrary units.

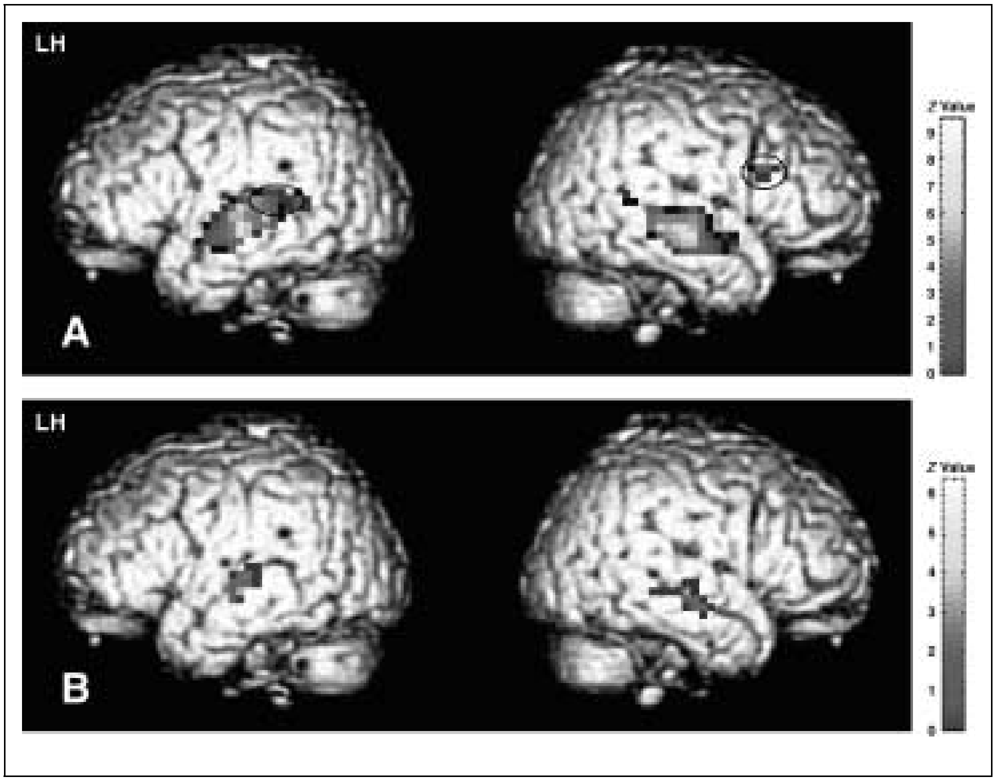

Figure 4. Complex sounds versus simple sounds. Cortical surface rendering of differential activation for the following: (A) Speech relative to simple tones. Circumscribed areas map onto the posterior STG in the left hemisphere, and the IFG in the right hemisphere. (B) Complex nonspeech relative to simple tones (all comparisons: fixed-effects analysis; display threshold z = 4.63; gradient bars indicate corresponding z score).

studies also report the lateralization of a unique locus in the left hemisphere that is differentially activated by speech though the exact location is variable: Some studies report activity in the supramarginal gyrus (Benson et al., 2000; Celsis et al., 1999), others activate the posterior STG (Mummery et al., 1999; Binder et al., 1994; Zatorre et al., 1992), or the anterior STG (Price et al., 1996). The more anterior activation observed in the latter study might be due to their use of real words (which presumably activate additional processes beyond simple speech detection) compared with the nonsense words used in other studies, including our own.

This study corroborates previous neuroimaging and neuropsychological research indicating that human language is processed by unique neural substrates. The recruitment of the left posterior STG in a variety of speech processing tasks (see Table 2) suggests that it plays an important role in that process. The activation we observe in this study using a simple oddball detection task that (1) does not require overt identification of the stimulus as speech, and (2) does not require in-depth linguistic analysis of the speech stimuli suggests that the role the left posterior STG plays is fundamental during the first steps of linguistic processing. This pattern of findings is inconsistent with theories in which the recognition and analysis of speech is like the perception of all sounds (e.g., Diehl & Kluender, 1986). Instead, these results argue for modular theories of

Table 2. Studies Contrasting Speech and Nonspeech Processing

			Talairach Coordinates		
Study	*Task*	*Nonspeech Comparison*	*x*	*y*	*z*
Benson et al. (2000)	passive listening	musical nonspeech	−62	−23	16
Binder et al. (1994)	passive listening	white noise	n.a.	n.a.	n.a.
Binder et al. (2000)	nonspecific button press	tones	−52	−42	6
Celsis et al. (1999)	change detection	tones	−30	−52	30
Mummery et al. (1999)	passive listening	signal-correlated noise	−54	−38	8
Zatorre et al. (1992)	nonspecific button press	white noise	−58	−21	8
Vouloumanos et al.	simple detection	complex matched nonspeech	−63	−42	13

Coordinates (x, y, z), in standard Talairach space, of the peak activation area in the posterior STG of the left hemisphere.

speech perception (e.g., Liberman, 1996; Fodor, 1983), which claim that the processing of language is specialized from the initial detection of linguistic stimuli. Though our results strongly suggest a neural specialization for processing human speech, we will discuss several additional accounts that could enrich the interpretation of these findings.

Potential Effects of Attention and Familiarity

One account of the differential cortical activation elicited during speech detection speaks to the increased attentional resources recruited during its processing. Though attention is likely to play some modulatory role on the magnitude of the hemodynamic response during auditory processing in our task, we think it unlikely to fully account for the pattern of differential activation we observe. In designing our task as one of simple detection with both speech and complex nonspeech as infrequent oddballs embedded among background tones, we effectively equated task-related attentional demands between the two types of oddball stimuli. Listeners were required to monitor the auditory stream and perform the same operation irrespective of stimulus type. Moreover, in using the event-related design that allows for pseudorandomization of stimuli of interest, listeners' attention levels should have been maintained throughout as there is no reliable way for them to anticipate the next stimulus type (standard or oddball), as may be the case in block designs.

Although this study explicitly controlled task-related attentional demands, it could be argued that speech stimuli recruited greater attentional resources through their status as familiar sounds in auditory space. We have attempted to minimize this confound by: (a) using a nonsense word, which was novel to the listener, and (b) comparing speech to high and low tones, which are familiar sounds frequently used as dial tones, busy signals, or alerting signals at traffic crossings adapted for persons with disabilities. Comparisons of speech to relatively novel complex nonspeech and to relatively familiar tones yielded a similar differential pattern of activation. This suggests that familiarity, if implicated at all, would only contribute modestly to the pattern of activation we observe.

The results of neuroimaging studies investigating auditory attention during speech processing corroborate the modesty of its potential contribution. Selective attention tasks report patterns of temporal lobe activation in part similar to the pattern we observe in our speech detection task, and in part notably different. As in our study, tasks of selective attention revealed a left hemisphere advantage in the MTG for attended conditions (Hashimoto, Homae, Nakajima, Miyashita, & Sakai, 2000; Pugh et al., 1996). However, in the STG, the attention-related activation elicited by these tasks was bilateral (Hashimoto et al., 2000; Grady et al., 1997; Pugh et al., 1996), in stark contrast to the left-lateralized activation locus we observe in the posterior STG (see Figure 2). The differential activation in the posterior STG of the left hemisphere elicited during speech processing is therefore unlikely to be modulated by attention.

On the other hand, attentional mechanisms are likely to modulate the activation we observe in the right IFG. This area has reliably been recruited in studies investigating attention and general arousal (Hashimoto et al., 2000; Stevens & Schwartzreich, 2000; Tzourio et al., 1997; Pugh et al., 1996). Although this activation might be speech-specific, a number of alternative explanations have been proposed. In a different task that preserves some features of the "change detection" aspect of our task, Celsis et al. (1999) also observed a small region in the right frontal gyrus that was activated by a deviant sequence containing a square tone (compared to a sine tone). The authors suggest pitch monitoring of deviants (consistent with the results of Pugh et al., 1996; Zatorre et al., 1992; Zatorre, Evans, & Meyrer, 1994), and differences in spectral content between stimuli of interest (square > sine) as possible explanations. Another interesting possibility is that this region is involved in processing nonphonetic ("voice") aspects of natural speech, but not in processing words themselves (Stevens & Schwartzreich, 2000; but see Belin, Zatorre, Lafaille, Ahad, & Pike, 2000).

Is There a Ventral Specialization for Speech Processing?

It has been proposed that the temporal lobe is structured around a functional dichotomy, in which the dorsal and lateral surfaces of the STG are involved in unimodal auditory processing (Galaburda & Sanides, 1980), whereas the ventral aspect receives and integrates input from many modalities (Mesulam, 2000; Rauschecker, 1998; Baylis, Rolls, & Leonard, 1987). Since language processing often integrates information from multiple modalities (e.g., bimodal speech perception; McGurk & MacDonald, 1976), language functions are likely to be carried out in areas where different streams converge (Geschwind, 1965). The results of some neuroimaging studies comparing speech and nonspeech processing support this hypothesis. Studies comparing speech to white noise (Binder et al., 1994, 2000; Zatorre et al., 1992), tones (Binder et al., 2000), and signal-correlated noise (Mummery et al., 1999) reported nonspecific activation in the dorsal aspect of the STG to both speech and nonspeech, whereas activity in the ventral STG, and in the STS, was more closely correlated with speech alone.

Our study does not provide the level of anatomical resolution required to address this issue conclusively. However, within our limited resolution, our results seem to be consistent with these studies in suggesting that the detection of speech is more closely associated with ventral processing streams. Compared with complex nonspeech, speech activated a large region of the MTG, and areas of the STG along the STS (see Figure 2). A comparison of speech to simple tones revealed a similar, but more extensive, pattern of activation including massive recruitment of the MTG and the ventral aspect of the STG (see Figure 4a). Thus, our results seem to be consistent with a more ventral specialization for speech processing. Clearly, further converging research is required to confirm this ventral precedence.

Is Differential Activation a Matter of Complexity?

A possible explanation for the differential activation to speech may be that speech stimuli are acoustically more complex than the sine wave analogues (despite matching on coarse spectral content, speech contains broadband frequency information and harmonic spectra that are lacking in the nonspeech analogues), and therefore recruit more of the same cortical resources during processing. Indeed, the overlapping regions of maximal activation observed in the MTG during processing of speech and complex nonspeech would point to at least some common processing mechanisms (compare Figure 4a and b). The differences in topography of STG activation (particularly the lateralized locus in the left posterior STG) argue against this. Moreover, recent evidence from a study comparing processing of speech and musical nonspeech stimuli of systematically varying complexity indicates that nonspeech complexity is reflected in corresponding activation of Heschl's gyrus and the areas immediately adjacent, and not in regions posterior to secondary auditory cortex (Benson et al., 2000). For these reasons, it is unlikely that stimulus complexity can fully account for the differential pattern of activation.

Is Lateralization Due to the Rapid Transitions of Speech?

Temporal processing theorists suggest that left hemisphere lateralization for speech processing results from a specialization for processing all rapidly changing acoustic information, not for speech per se (e.g., Tallal, Miller, & Fitch, 1993). Although the left hemisphere does possess an advantage for processing rapidly changing temporal information (e.g., Belin et al., 1998; Johnsrude, Zatorre, Milner, & Evans, 1997; Schwartz & Tallal, 1980), the temporal processing hypothesis is unlikely to account for the lateralization we observe during speech detection. The sine wave nonspeech counterparts in this study maintain the peak frequency changes of the three formants and the fundamental frequency of speech across time, thus preserving the main rapid temporal changes present in our speech stimuli. As a result, rapidly changing temporal information was preserved in both speech and nonspeech stimuli, yet a left hemisphere lateralization was only observed during speech detection. The left hemisphere advantage for speech processing in this detection task is not related to the temporal attributes that characterize speech, but rather to the fact that the stimulus is speech by nature.

CONCLUSIONS

This study corroborates evidence from previous neuroimaging and neuropsychological studies indicating that human language is processed by unique neural substrates. Moreover, our results suggest that this specialization for speech is present from the early stages of processing. Our study is unique in contributing to this growing body of work in five important ways. (1) We compare speech to carefully controlled complex nonspeech stimuli (sine wave analogues) that track key spectral and temporal aspects of speech. (2) We use an event-related fMRI design that allows us to model the hemodynamic response to individual auditory events. This, in comparison with block designs, allows us to isolate more precisely the neural events associated with speech detection. (3) Our use of a simple oddball detection paradigm embeds speech and complex nonspeech within a uniform background, and equates attentional demands for processing the stimuli of interest. (4) Our testing of a larger sample than most neuroimaging studies, and the corroboration of our results using an exploratory random-effects analysis, strengthen the generalizability of the findings. (5) The differential activation we observe cannot be fully accounted for by the differential recruitment of attention during speech processing, by the acoustic-level characteristics of the speech stimuli, such as complexity or rapid transitions, or by higher-order linguistic processing (e.g., semantic or syntactic) of the stimuli. Instead, these neural substrates appear to be specifically activated by properties intrinsic to speech.

Clearly, detected speech input generally requires further linguistic analysis, and that analysis, be it semantic, syntactic, or phonological, will likely activate additional brain mechanisms. However, the pattern of activation we observe suggests that some distinct neural mechanisms are involved in the initial processing of speech, and elucidates how, in the cacophony of sounds in the environment, spoken language stands out as an exceptionally salient signal for the human brain.

METHODS

Participants

Fifteen healthy right-handed adult volunteers (six women, nine men, mean age 25.5) participated in the study (handedness assessed as per Annett, 1967). Participants provided written informed consent and were screened for MRI compatibility before entry into the scanning room. All experimental procedures met with university ethical approval.

Procedure

Sounds were presented through insert earphones embedded within 30-dB sound-attenuating MR compatible headphones using custom presentation software (http://nilab.psychiatry.ubc.ca/vapp). Because of the difficulty in accurately measuring absolute intensity values at the exit point from insert earphones, all sounds, speech, complex nonspeech, and tones were equated for intensity relative to each other. Moreover, sounds were clearly audible above the noise of the scanner, as evidenced by listeners' near-perfect performance. Participants heard two stimulus runs, an experimental run and a tone run. The background nontarget stimulus in both runs was a 1000-Hz tone, occurring with a probability of .8. The stimuli of interest were presented in a pseudorandom oddball design (separated by three to five nontarget stimuli). In the ''experimental'' run, the infrequent sounds consisted of speech (.1) and complex nonspeech (.1). In the ''tone'' run, 1500-Hz high tones (.1) and 500-Hz low tones (.1) were the infrequent sounds. Each run was 12.5 min, with a 2-sec stimulus onset asynchrony (SOA) for a total of 380 total stimuli per run. Order of presentation of the stimulus runs was counterbalanced across participants. Participants made a motor response on an MRI-compatible fiber optic response device (Lightwave Medical, Vancouver, BC) using their left index finger for every infrequent sound they heard. Reaction times were monitored on-line.

Stimuli

The stimuli of interest were of four different types: (a) speech, (b) complex nonspeech, (c) high tones, and (d) low tones. (A) Speech stimuli consisted of six tokens of a monosyllabic nonsense word ''lif''[2] spoken by a native female English speaker; tokens varied in intonational contour (average minimum and maximum pitch: 202 and 350 Hz, respectively) and in duration (525–711 msec). (B) Complex nonspeech stimuli consisted of time varying sine wave analogues of the speech tokens in which all regions of significant energy were tracked (namely the fundamental frequency and the first three formants; see Figure 1). Sinusoidal waves tracking these energy peaks were created individually in Mathcad 3.1 (Mathsoft, Cambridge, MA). Fundamental frequency (corresponding to pitch) was tracked individually for each of the six speech tokens. Because the first three formants were virtually identical across the multiple natural repetitions, one set of formants from a representative word token was tracked. This set was composed of the first formant of the initial consonant segment (''l''), and the first three formants of the vocalic segment (''i''). The sine analogue to (''f'') was created using a white-noise generator and filtered. This representative set was then added onto the sine wave analogue of the pitch contour of each segment using Signalyze 3.12 (Agora Language Marketplace, Charlestown, MA) to create six different stimuli. Analogues thus retained the duration, pitch contour, amplitude envelope, relative formant amplitude, and relative intensity of their speech counterparts (see Figure 1). (C) Low tones were six pure sinusoidal waves of 500 Hz generated using Sound Edit Pro, version 2 (Macromedia, San Francisco, CA) matched in duration to the speech stimuli. (D) High tones were six pure sinusoidal waves of 1500 Hz generated using Sound Edit Pro, version 2 (Macromedia) matched in duration to the speech stimuli.

Imaging Parameters

Echo-planar images (EPI) were collected on a standard clinical GE 1.5-T system fitted with a Horizon Echospeed upgrade. Conventional spin-echo T_1-weighted sagittal localizers were used to view the positioning of the participant's head and to graphically prescribe the functional image volumes. Functional image volumes were collected with a gradient-echo (GRE) sequence (TR/TE 3000/40 msec, 90° flip angle, FOV 24 × 24 cm, 64 × 64 matrix, 62.5 kHz bandwidth, 3.75 × 3.75 mm in plane resolution, 5.00 mm slice thickness, 29 slices, 145 mm total brain coverage). This sequence is sensitive to the blood oxygen level-dependent (BOLD) contrast (Ogawa, Lee, Kay, & Tank, 1990). Each stimulus run consisted of 246 BOLD scans (full-brain scans). The first 12 sec collected at the beginning of each run were discarded from the analyses, to avoid the T_1 saturation effects that occur in the early scans.

Image Processing

Functional images were reconstructed off-line. Statistical parametric mapping software (SPM99, Wellcome Department of Cognitive Neurology, London, UK) was used for image realignment and normalization into modified Talairach stereotaxic anatomical space (using affine and nonlinear components, as implemented in SPM99). Images were smoothed using a Gaussian kernel (8 mm FWHM) to compensate for intersubject anatomical differences, and to optimize the signal-to-noise ratio. Event-related responses to the stimuli of interest were modeled using a synthetic hemodynamic response composed of two gamma functions and their temporal derivatives (for a discussion of the relative advantages and disadvantages of this modeling method, see Kiehl, Laurens, Duty, Forster, & Liddle, 2001). The peak of the response was modeled at 6 sec poststimulus time, consistent with the results of other event-related fMRI studies (Hickok et al., 1997; but see Belin et al., 1999, for a shorter peaking time in a different acoustic setting). A high-pass filter (cutoff period 89 sec) was incorporated into the model to remove noise associated with low frequency confounds. A low-pass filter (at the Nyquist frequency, with a period of 6 sec) was also applied to remove noise associated with alternations of the applied radio frequency field. Three contrasts were used to create SPM$\{t\}$ maps, later transformed into SPM$\{Z\}$ maps, for three comparisons of interest: (a) activation for speech sounds relative to the complex nonspeech stimuli, (b) activation for the speech sounds relative to simple tones, and (c) activation for the complex nonspeech sounds relative to the simple tones.

Statistical Analyses

Statistical analyses were performed in SPM99 using a fixed-effects model. Because multiple voxels were examined, a correction for multiple comparisons based on the theory of Gaussian fields was employed. The areas of activation reported are significant at the voxel level, with z scores greater than 4.63 corresponding to a corrected significance level of $p < .05$. We further explored hemispheric differences in activation by comparing suprathreshold voxels in each hemisphere for individual listeners. Within the SPM program, we imposed a mask of the MTG on each listener's SPM$\{t\}$ map for the main comparison of interest (see (a) above). A custom script was used to extract suprathreshold voxels ($z = 2.63, p <$ .05 uncorrected) in the left and right hemispheres of every listener. A paired t test was conducted on these hemispheric voxel counts to obtain an index of hemispheric asymmetry.

The standard fixed effects model of analysis was used to analyze patterns of activation within subjects because of the greater power it affords us in detecting details of the activation patterns. In order to demonstrate that our main findings can be generalized to the population, we performed an exploratory analysis using a random-effects model in SPM99 on the main comparison of interest, that of speech sounds relative to complex nonspeech stimuli. We consider this analysis exploratory because the sample size in this study ($n = 15$), though larger than that of most neuroimaging studies, does not allow an adequate level of power to perform a full-fledged random-effects analysis. Images analyzed using this model were smoothed with a 14-mm FWHM Gaussian filter. The areas of activation reported using this analysis are significant at the voxel level, with z scores greater than 4.63, corresponding to a corrected significance level of $p < .05$.

Acknowledgments

This work was supported by a Killam Predoctoral Fellowship to A.V., and grants from the Natural Sciences and Engineering Council (NSERC) of Canada (PGSA/B Fellowships to A.V. and Grant RG1103 to J.F.W.), and the Medical Research Council (MRC) of Canada (Michael Smith Graduate Scholarship to K.A.K., and Grant MOP-36423 to P.F.L.). We thank the participants who graciously volunteered their time for this study. We are very grateful to Kristin Laurens, and Adrien Desjardins for their indispensable help throughout this project. We thank MR technicians Jennifer McCord, Trudy Shaw, and Sylvia Renneberg for their excellent technical assistance. We would also like to thank Tracey Burns, Kristin Laurens, and two anonymous reviewers for their insightful comments on an earlier version of this manuscript. The sine wave stimuli used in this study were generated by Sonya Bird and Dr. Guy Carden.

Reprint request should be addressed to Athena Vouloumanos, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, Canada V6T 1Z4, or via e-mail: athenavo@interchange.ubc.ca.

The data reported in this experiment have been deposited in the National fMRI Data Center (http://www.fmridc.org). The accession number is 2-2001-111TW.

Notes

1. Perceptual tests in our laboratory with eight monolingual English speakers confirmed that all eight listeners identified the nonsense speech sounds as human vocalizations, while none identified the sine wave analogues as such
2. For ease of readability, we chose to describe the speech stimuli using "gloss." The equivalent in IPA symbols is /lɪf/.

REFERENCES

Annett, M. (1967). The binomial distribution of right, mixed, and left handedness. *Quarterly Journal of Experimental Psychology, 19,* 327–333.

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1987). Functional subdivisions of the temporal lobe neocortex. *Journal of Neuroscience, 7,* 330–342.

Belin, P., Zatorre, R. J., Hoge, R., Evans, A. C., & Pike, B. (1999). Event-related fMRI of the auditory cortex. *Neuroimage, 10,* 417–429.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403,* 309–312.

Belin, P., Zilbovicius, M., Crozier, S., Thivard, L., Fontaine, A., Masure, M. C., & Samson, Y. (1998). Lateralization of speech and auditory temporal processing. *Journal of Cognitive Neuroscience, 10,* 536–540.

Benson, R., Whalen, D., Clark, V., & Liberman, A. (2000). *fMRI of phonetic processing.* Paper presented at the Cognitive Neuroscience Meeting, San Francisco, CA, USA.

Binder, J. R., Frost, J. A. Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and non-speech 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.

Binder, J. R., Rao, S. M., Hammeke, T. A., Yetkin, F. Z., Jesmanowicz, A., Bandettini, P. A., Wong, E. C., Estkowski, L. D., Goldstein, M. D., Haughton, V. M., et al. (1994). Functional magnetic resonance imaging of human auditory cortex. *Annals of Neurology, 35,* 662–672.

Celsis, P., Boulanouar, K., Doyon, B., Ranjeva, J. P., Berry, I., Nespoulous, J. L., & Chollet, F. (1999). Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *Neuroimage, 9,* 135–144.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin and use.* New York: Praeger.

Chomsky, N. (2000). *New horizons in the study of language and mind.* Cambridge, UK: Cambridge University Press.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–163). Hillsdale, NJ: Erlbaum.

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping, 8,* 109–114.

Damasio, A. R., & Geschwind, N. (1984). The neural basis of language. *Annual Review of Neuroscience, 7,* 127–147.

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

Démonet, J. F., Fiez, J. A., Paulesu, E., Petersen, S. E., & Zatorre, R. J. (1996). PET studies of phonological processing: A critical reply to Poeppel. *Brain and Language, 55,* 352–379.

D'Esposito, M., Zarahn, E., & Aguirre, G. K. (1999). Event-related functional MRI: Implications for cognitive psychology. *Psychological Bulletin, 125,* 155–164.

Diehl, R. L., & Kluender, K. R. (1986). On the objects of speech perception. *Ecological Psychology, 1,* 121.

Fiez, J. A., & McCandliss, B. D. (2000). *Dishabituation of bold responses to phonetic oddballs: An event-related fMRI study of magnitude of acoustic change and native language history.* Paper presented at the Society for Neuroscience Meeting, New Orleans, LA, USA.

Fiez, J. A., Raichle, M. E., Miezin, F. M., Petersen, S. E., Tallal, P., & Katz, W. F. (1995). PET studies of auditory and phonological processing: Effects of stimulus characteristics and task demands. *Journal of Cognitive Neuroscience, 7,* 357–375.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge: MIT Press.

Galaburda, A., & Sanides, F. (1980). Cytoarchitectonic organization of the human auditory cortex. *Journal of Comparative Neurology, 190,* 597–610.

Galaburda, A. M., Sanides, F., & Geschwind, N. (1978). Human brain. Cytoarchitectonic left–right asymmetries in the temporal speech region. *Archives of Neurology, 35,* 812–817.

Geschwind, N. (1965). Disconnection syndromes in animals and man: I. *Brain, 88,* 237–294.

Geschwind, N., & Levitsky, W. (1968). Human brain: Left–right asymmetries in temporal speech region. *Science, 161,* 186–187.

Grady, C. L., Van Meter, J. W., Maisong, J. Ma., Pietrini, P., Krasuski, J., & Rauschecker, J. P. (1997). Attention-related modulation of activity in primary and secondary auditory cortex. *NeuroReport, 8,* 2511–2516.

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

Hickok, G., Love, T., Swinney, D., Wong, E. C., & Buxton, R. B. (1997). Functional MR imaging during auditory word perception: A single-trial presentation paradigm. *Brain and Language, 58,* 197–201.

Johnsrude, I. S., Zatorre, R. J., Milner, B. A., & Evans, A. C. (1997). Left-hemisphere specialization for the processing of acoustic transients. *NeuroReport, 8,* 1761–1765.

Kiehl, K. A., Laurens, K. R., Duty, T. L., Forster, B. B., & Liddle, P. F. (2001). Neural sources involved in auditory target detection and novelty processing: An event-related fMRI study. *Psychophysiology, 38,* 133–142.

Liberman, A. M. (1996). *Speech: A special code.* Cambridge: MIT Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Mesulam, M. M. (2000). Behavioral neuroanatomy: Large-scale networks, association cortex, frontal syndromes, the limbic system, and hemispheric specializations. In M. M. Mesulam (Ed.), *Principles of behavioral and cognitive neurology* (pp. 1–120). New York: Oxford University Press.

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.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Science, U.S.A., 87,* 9868–9872.

Poeppel, D. (1996). A critical review of PET studies of phonological processing. *Brain and Language, 55,* 317–351.

Price, C. J., Wise, R. J., Warburton, E. A., Moore, C. J., Howard, D., Patterson, K., Frackowiak, R. S., & Friston, K. J. (1996). Hearing and saying. The functional neuro-anatomy of auditory word processing. *Brain, 119,* 919–931.

Pugh, K .R., Shaywitz, B. A., Shaywitz, S. E., Fulbright, R. K., Byrd, D., Skudlarski, P., Shankweiler, D. P., Katz, L., Constable, R. T., Fletcher, J., Lacadie, C., Marchione, K., & Gore, J. C. (1996). Auditory selective attention: An fMRI investigation. *Neuroimage, 4,* 159–173.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology, 8,* 516–521.

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. (1983). Coding of the speech spectrum in three time-varying sinusoids. *Annals of the New York Academy of Sciences, 405,* 485–489.

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

Rosen, B. R., Buckner, R. L., & Dale, A. M. (1998). Event-related functional MRI: Past, present, and future. *Proceedings of the National Academy of Sciences, U.S.A., 95,* 773–780.

Schwartz, J., & Tallal, P. (1980). Rate of acoustic change may underlie hemispheric specialization for speech perception. *Science, 207,* 1380–1381.

Stevens, A. A., & Schwartzreich, E. (2000). *Dissociating message and messenger: fMRI of word and voice recognition.* Paper presented at the Cognitive Neuroscience Meeting, San Francisco, CA, USA.

Tallal, P., Miller, S., & Fitch, R. H. (1993). Neurobiological basis of speech: A case for the preeminence of temporal processing. *Annals of the New York Academy of Science, 682,* 27–47.

Tzourio, N., Massioui, F. E., Crivello, F., Joliot, M., Renault, B., & Mazoyer, B. (1997). Functional anatomy of human auditory attention studied with PET. *Neuroimage, 5,* 63–77.

Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science, 237,* 169–171.

Wong, D., Miyamoto, R. T., Pisoni, D. B., Sehgal, M., & Hutchins, G. D. (1999). PET imaging of cochlear-implant and normal-hearing subjects listening to speech and nonspeech. *Hearing Research, 132,* 34–42.

Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience, 14,* 1908–1919.

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science, 256,* 846–849.