Visual speech perception without primary auditory cortex activation

Lynne E. Bernstein,[CA] Edward T. Auer Jr, Jean K. Moore,[1] Curtis W. Ponton,[2] Manual Don[3] and Manbir Singh[4]

Departments of Communication Neuroscience, [1]Neuroanatomy and; [2]Neuroscan Labs, El Paso, TX; [3]Electrophysiology, House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057; [4]Departments of Radiology and Biomedical Engineering, University of Southern California

[CA]Corresponding Author

Received 8 November 2001; accepted 19 December 2001

Speech perception is conventionally thought to be an auditory function, but humans often use their eyes to perceive speech. We investigated whether visual speech perception depends on processing by the primary auditory cortex in hearing adults. In a functional magnetic resonance imaging experiment, a pulse-tone was presented contrasted with gradient noise. During the same session, a silent video of a talker saying isolated words was presented contrasted with a still face. Visual speech activated the superior temporal gyrus anterior, posterior, and lateral to the primary auditory cortex, but not the region of the primary auditory cortex. These results suggest that visual speech perception is not critically dependent on the region of primary auditory cortex. *NeuroReport* 13:311–315 © 2002 Lippincott Williams & Wilkins.

Key words: fMRI; Primary auditory cortex; Visual speech perception

INTRODUCTION

Humans often use their eyes to perceive spoken language. For example, speech perception in a noisy environment is more accurate when the listener can see the talker [1]; watching a talker can affect what speech sounds are heard [2]; and sometimes vision is the sole basis for perceiving spoken language [3]. These observations suggest that there is a visual processing pathway for perceiving the physical, visual speech stimulus (visual phonetics). However, the neural basis for auditory speech perception is thought to involve a processing pathway through the primary and secondary auditory cortex [4–8], and there has even been a suggestion that this is true for visual speech stimuli as well [9].

The region of primary auditory cortex (PACr), located on the superior temporal plane, has been shown in functional brain imaging studies to be activated unfailingly by speech and non-speech sounds [4–8]. Primary auditory cortex [10,11] (koniocortex; Brodmann [12] area 41) has been identified cytoarchitectonically as an area, usually 2–3 cm in diameter, located on the caudomedial aspect of the transverse temporal gyrus [13]. Primary auditory cortex also includes the parakoniocortex (BA 42), which surrounds BA 41 laterally, anteriorly, and posteriorly. BA 42 is a region that usually covers the remainder of the transverse temporal (Heschl's) gyrus and may extend for some distance laterally, or for a variable distance anteriorly onto the planum polare, or posteriorly onto the planum temporale. The superior temporal plane rostral and caudal to BA 41/42 and the lateral surface of the superior temporal gyrus (STG) down to the superior temporal sulcus (STS) was designated area 22 by Brodmann. PACr and the surrounding areas on the superior temporal plane respond to a wide range of auditory stimuli, including noise, tones, and speech. Broadly described, activation due to speech signals appears to spread from the PACr, across the superior temporal plane, to the lateral surface of the STG, to the STS, and the middle temporal gyrus (MTG) [4–8], the latter areas (the lateral STG, STS, and MTG) likely including areas specialized for spoken language processing.

In support of this same route for visual speech perception, Calvert *et al.* [9] reported a fMRI study that indicated activation by visual speech in primary auditory cortex (BA 41/42) with hearing adults. Their conclusions were consistent with the possibility that visual speech information is injected into primary auditory cortex, at which point it could follow the same processing route as auditory speech stimuli, including phonetic analysis and mapping of information into the lexicon. This cortical pathway is plausible in light of a human lesion study that showed direct monosynaptic connections from the visual areas attributed to face processing to the STG, including the superior temporal plane [14]. If this route were the means by which the brain processes the visual phonetic speech stimulus, then visual speech perception would be essentially an auditory function.

However, the Calvert *et al.* report left some room to question whether the activation attributed to primary auditory cortex indeed occurred there. If the coordinates for activation peaks reported in Calvert *et al.* corresponding

to BA 41/42 for each of the experiments are referred to the Penhune *et al*. probabilistic maps of primary auditory cortex [15], the highest levels of probability of having activated primary auditory cortex appear to be in the 25–50% range (experiment 1 auditory, x = −49, y = −19, z = 13; experiment 2 visual, x = −52, y = −22, z = 8). In a replication by Calvert *et al*. of their visual Experiment 2 reported in the same paper [9], the activation peak (x = −56, y = −26, z = 9.5) fell outside the probable range for PAC.

That visual stimuli activate primary auditory cortex is also not consistent with Penfield's studies of electrical stimulation [16]. Stimulation to the superior temporal cortex produced simple sensations such as buzzing or complex sensations such as voices and music, but visual sensations were not reported. Visual sensations were obtained only with stimulation of more lateral regions of the temporal lobe.

Given variability in the cortical location of primary auditory cortex in terms of stereotaxic space [15], establishing a visual route through the PACr requires examination of results on a per-participant basis. Using fMRI, we imaged participants who performed in a silent lipreading task. A pulse-tone stimulus rather than speech was used to localize PACr, because auditory speech stimuli would result in activation beyond the PACr [6]. Activation by visual speech in secondary or association areas was not the focus but rather, whether visual speech is processed by a network involving the PACr. Our results showed cortical activation for visual speech stimuli that was distinct from cortical activation to pulse-tone stimuli in the PACr.

MATERIALS AND METHODS

Participants: Participants were seven young (age 19–31 years) right-handed adults with normal hearing, with English as a first language, and average or better lipreading relative to their normative group. Testing was approved by an Institutional Review Board. Participants gave informed consent and were paid $75.

Stimuli: The stimulus used to localize the temporal plane and the PACr was a sequence of 100 ms, 1000 Hz tones repeating at a rate of 5/s for 30 s. This stimulus was contrasted with a 30 s no stimulus interval, during which only gradient noise was present. At high amplitudes, a single 1000 Hz pulse-tone should activate the cochlea at most of the frequencies related to speech. The amplitude of the tone stimuli was adjusted to be easily audible in the fMRI scanner.

In lipreading runs, a sequence of spoken monosyllabic words was presented contrasted with a sequence of colored-shapes overlayed on a still frame of the same talker's face. A total of 240 stimuli was presented, 120 word stimuli in the lipreading condition, and 120 colored-shape stimuli in the control condition. The monosyllabic words were spoken in isolation by a male talker whose face filled the video frame. Word stimuli consisted of a sequence that included immediately repeated words (27%), perceptually similar words (24%), and perceptually dissimilar words (49%) [17]. The still-face control stimuli consisted of one still frame of the talker's face with one of five colored shapes overlayed on the bridge of the talker's nose. The colored-shapes consisted of a green triangle, a yellow circle, a red star, a blue square, and a purple pentagon. The duration of the presentation of the colored-shapes was matched to the durations of the words. Over the entire experiment, the pattern and number of immediately repeated colored-shape stimuli matched the sequence of the word stimuli. All stimuli were dubbed onto video tape (at 30 frames/s) such that a stimulus was presented every 2 s with four blank frames of video in the inter-stimulus interval. After the blank frames, 13 video frames of a still face were presented before the spoken word or colored-shape stimulus began. The word stimuli and the colored-shape stimuli were each presented in 1 min blocks. Stimuli were projected from a screen onto a mirror mounted above the head of the subject. Images of the face subtended visual angles of $8 \times 4^{\circ}$.

Procedure: Participants listened passively during the periods of pulse-tone sequences and the intervening no-stimulus periods of the pulse-tone run. During the separate lipreading runs, they actively responded by indicating via a bulb squeeze whenever two consecutive stimuli were identical. The visual speech control condition, for which the colored-shape stimuli were presented, involved the same motor response, the same type of discrimination judgment, the same duration of stimulus tokens, and the same face. Each imaging session comprised a pulse-tone run and at least one lipreading run. All but two participants received two lipreading runs.

Imaging: Imaged tissue was four contiguous 10 mm transaxial sections, with the imaging centered on the approximate location of Heschl's (transverse temporal) gyrus, where PAC is typically located. The imaged tissue thus included the posterior half of the STG and most of the superior temporal plane. It also included most of the angular and supramarginal gyri. Four 10 mm contiguous coronal sections were also imaged for one participant, with sections centered on the posterior end of the Sylvian fissure.

A GE 1.5 T Signa Horizon MRI system equipped with echo-planar imaging (EPI) was used with a quadrature head-coil to acquire a time-series of images using an EPI sequence with the following parameters: TR (repetition time) = 4 s, effective TE (echo time) = 45 ms, 90° flip angle, 64×128 acquisition matrix, $20 \times 40\,cm^2$ field of view and NEX (no. of excitations) = 1. A total of 500 images was acquired from four 10 mm contiguous transaxial or coronal sections (125 images per section). Prior to the EPI time-series acquisition, spin-echo anatomical images of the same four sections were also acquired using TR = 400 ms and TE = 14 ms to obtain good gray/white-matter demarcation.

Analyses: For each run, the first five images per section were ignored to establish equilibrium, and starting at image 6, the experimental stimulus and the corresponding control condition were presented in an alternating sequence of 30 s on for the pulse-tone *vs* 30 s off (8 min total), and an alternating sequence of 60 s on for the lipreading *vs* 60 s control (8 min total). The time-series of images were co-registered to a reference image within the time-series using the method described in detail in Singh *et al*. [18]. Activated voxels were identified using multiple linear regression

implemented in SPM99 [19]. A box-car reference function was used, matching the task and control presentation sequence and delayed by one image (i.e. 4 s), to account for the hemodynamic delay.

In order to determine whether lipreading activated the PACr, the z-score of each voxel was computed, and voxels whose z-scores (uncorrected) were above a threshold set by $p < 0.001$ were identified for the lipreading and pulse-tone runs separately. Then, the above-threshold activations in clusters of ≥ 5 voxels in the lipreading and tone runs, within each participant, were color-coded in terms of the respective run type and registered on the participant's own anatomical images. These maps were visually inspected for potential common activation due to tones and visible speech in areas that anatomically qualified as the PACr.

Independently, in SPM analyses, the tone activations were used as inclusive masks for the lipreading activations on a per-participant basis. When there were two runs of the lipreading, each run was submitted to the inclusive mask analysis separately. Within each participant, the coordinates of voxels ($p < 0.001$, uncorrected) in clusters of ≥ 5 became candidates for PAC activation, if they were within a bounding box in SPM coordinates containing all of the Penhune *et al*. [15] probable areas of PACr. Voxel size was $2 \times 2 \times 2$ mm, and inter-peak distance was set at > 8 mm. Candidate peaks were transformed into the Talairach-Tournoux [20,21] space, and the probability of their being in PAC was estimated directly using the maps in Penhune *et al.* [15]. Finally, additional areas outside of PACr were identified for the lipreading condition by visual inspection and database lookup [22].

RESULTS

Figure 1a shows the individual participant results for a representative lipreading run and the single tone run projected on the participant's own anatomical MR images. The tone stimuli resulted in bilateral activation of the superior temporal plane, including areas that visual inspection identified as the PACr, the temporal plane, and the lateral surface of the STG. The activated PACr appears generally as a pattern of activation angled inward from the lateral surface of the temporal lobe (see Fig. 1b for PACr location). Figure 1a shows that tone activation was fairly extensive for several of the participants. Some areas of common activation were obtained (coded blue in Fig. 1a). Activation in response to lipreading was more widespread than in the tone experiment and included some frontal areas (coded red in Fig. 1a).

When the pulse-tone activation was used to mask inclusively the lipreading activation for each participant and lipreading run, several significant clusters were obtained. However, when each peak of activation was located on the Penhune *et al*. [15] probabilistic maps of the PACr, following transformation to Talairach-Tournoux stereotaxic space [20,21], none of the peaks was found to be within a probable location for the PACr.

Locations that were activated by lipreading across individual participants are shown in Fig. 1b. This figure shows the highest peak for each significant cluster in the lipreading run, across all participants and runs, after normalization in SPM99. The activation peaks in the temporal lobe were obtained in the STP, the STS, and the MTG. There were additional peaks in the inferior frontal gyrus, middle frontal gyrus, and the superior frontal gyrus. The figure also shows in red the location of the highest probability regions for PAC on the Penhune *et al.* [15] probabilistic maps.

One participant, NH6, was imaged in the coronal plane to investigate further the possibility of a common activation in the PACr. In this participant, the pulse-tone stimuli resulted in bilateral activation mainly of cortex on the superior temporal plane. Lipreading resulted in activation of the superior temporal sulcus predominantly on the right, with some activation in parietal cortex. Activation common to pulse-tones and visual speech was obtained along the lateral surface of the STG.

DISCUSSION

Results of this study did not support the existence of a pathway for processing visual speech that passes through the PACr. The current study suggests that dissociation of activity due to auditory stimuli versus activity due to visible speech stimuli can be shown on an individual-participant basis using a non-speech auditory stimulus versus a lipreading stimulus.

The pulse-tone stimulus effectively activated the functionally defined area of PACr (including BA 41/42), but visual speech did not. The visual speech activated the lateral STG, the STS, and the MTG, areas substantially similar to ones outside of the PACr reported by Calvert *et al.* [9] for lipreading. Common activation for auditory and visual speech are predicted for processes involving the lexicon [5–8]. Neither our study nor that of Calvert *et al.* dissociated phonetic from lexical processing.

Had our study confirmed that visual speech activates the PACr, then the problem of explaining visual speech perception would be partly solved: visual speech is injected into the PAC and is likely then processed as though it were auditory in origin. Given our results, we must seek elsewhere the cortical locations responsible for processing the visual forms of speech stimuli. Auer *et al*. [24] recently compared lipreading and fingerspelling in deaf and hearing adults in an fMRI study in which coronal images were obtained. They observed activation to lipreading in both groups in the STS, the fusiform gyrus (BA 19/37), the MTG, and the inferior temporal gyrus. Presently, it is not known whether any of these areas are specialized for processing the visual forms (the phonetic information) of speech stimuli. Further studies must dissociate phonetic from lexical and visual from auditory speech processing to determine whether there is/are cortical locations specialized for the visual forms of speech.

CONCLUSION

A pathway through the PACr appears to be obligatory for auditory speech perception [4–8,15] but not so for visual speech perception. Vision can influence heard speech [1,2]. Vision can also function alone for speech perception, particularly in some prelingually deaf adults who lipread with high levels of accuracy (e.g., 80% of words correct in isolated sentences) [3]. Taken together, the fMRI and

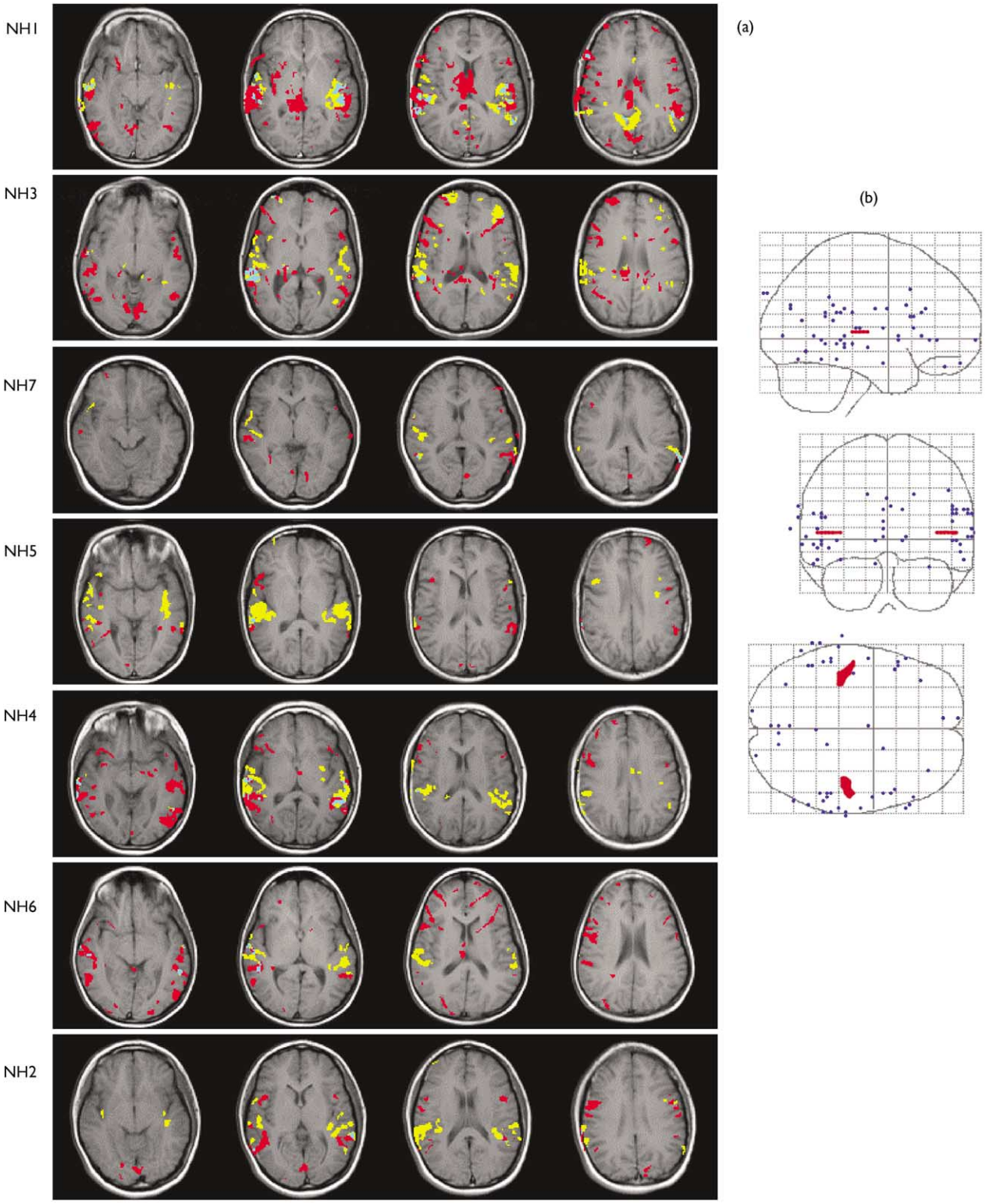

Fig. 1. (**a**) Brain activation maps superimposed on MRI images. Each participant's results are shown from inferior to superior (left to right). Orientation of images is with the left hemisphere shown on the right. Transaxial sections from individual participants show the areas (displayed in yellow) that were activated during the pulse-tone run, the areas (displayed in red) that were activated during a representative lipreading run for the participant, and the areas (displayed in blue) that were activated during both runs. (SPM threshold, $p < 0.001$, uncorrected). (**b**) Locations of peak activations, normalized in SPM, for every cluster obtained ≥ 5 voxels at a threshold of $p < 0.001$, corrected, plotted on a glass brain [23]. The largest, highest probability location for the PACr in Penhune *et al.* [15] is approximated in red. The majority of peaks are in the STG, the MTG, the inferior frontal gyrus, middle frontal gyrus, and the superior frontal gyrus.

behavioral evidence point to the likelihood that there is a modality-specific pathway for the processing of the visual phonetic forms of speech stimuli. Future studies will be needed to characterize and dissociate further this pathway from that serving auditory speech perception.

REFERENCES

1. Sumby WH and Pollack I. *J Acoust Soc Am* **26**, 212–215 (1954).
2. McGurk H and MacDonald J. *Nature* **264**, 746–748 (1976).
3. Bernstein LE, Demorest ME and Tucker PE. *Percept Psychophys* **62**, 233–252 (2000).
4. Benson RR, Whalen DH, Richardson M *et al*. *Brain Lang* **78**, 364–396 (2001).
5. Wise JS, Scott SK, Blank C *et al*. *Brain* **124**, 83–95 (2001).
6. Binder JR, Frost JA, Hammeke RA *et al*. *Cerebr Cortex* **10**, 512–528 (2000).
7. Zatorre RJ, Meyer E, Gjedde A *et al*. *Cerebr Cortex* **6**, 21–30 (1996).
8. Scott SK, Blank CC, Rosen S *et al*. *Brain* **123**, 2400–2406 (2000).
9. Calvert GA, Bullmore ET, Brammer MJ *et al*. *Science* **276**, 593–596 (1997).
10. Galaburda A and Sanides F. *J Comp Neurol* **190**, 597–610 (1980).
11. Lauter JL, Herscovitch P and Formby C. *Hear Res* **20**, 199–205 (1985).
12. Brodmann KJ. *Psych Neurol* **10**, 231–246 (1908).
13. von Economo C. *The Cytoarchitectonics of the Human Cerebral Cortex*. London: Oxford University Press; 1929.
14. Di Virgilio G and Clarke S. *Hum Brain Mapp* **7**, 29–37 (1997).
15. Penhune VB, Zatorre RJ, MacDonald JD *et al*. *Cerebr Cortex* **6**, 661–672 (1996).
16. Penfield W and Perot P. *Brain* **86**, 595–596 (1963).
17. Auer ET and Bernstein LE. *J Acoust Soc Am* **102**, 3704–3710 (1997).
18. Singh M, Al-Dayeh L, Patel P *et al*. *IEEE Trans. Nucl Sci* **45**, 2162–2167 (1998).
19. Friston K, Ashburner J, Poline J-B *et al*. *Hum Brain Mapp* **2**, 189–210 (1995).
20. Talairach J and Tournoux PA. *Co-planar Stereotactic Atlas of the Human Brain*. Stuttgart: Thieme; 1988.
21. Brett M. MatLab Code. (1999). http://www.mrc-cbu.cam.ac.uk/Imaging/mnispace.html.
22. Gitelman D. MIP_MAKER (MATLAB code). ftp://ftp.cnadc.nwu.edu/pub/spm/mip_maker.
23. Lancaster JL, Woldorff MG, Parsons LM *et al*. *Hum Brain Mapp* **10**, 120–131 (2000).
24. Auer ET Jr, Bernstein LE and Singh M. Comparing cortical activity during the perception of two forms of biological motion for language communication. In Massaro DW, Light J and Geraci K (eds). *Proceedings of AVSP 2001*. University of CA, Santa Cruz: Perceptual Science Lab, 2001: pp. 40–41.

Acknowledgements: We thank Paula Tucker, Sheri Hithe, and Betty Kwong for testing participants; Jeong-Won Jeong, Witaya Sungkarat, and Angelika Dimoka (fMRI acquisition and processing assistance); Linda Needham (operating the MRI); Patrick Colletti (access to the MRI at LA County, USC Imaging Science Center); and Geraint Rees (advice on analyses). Supported by House Ear Institute, NIH/NIDCD (DC02107), and NIH/NIA (AG05142).