Brain Activations During Conscious Self-Monitoring of Speech Production With Delayed Auditory Feedback: An fMRI Study

Yasuki Hashimoto[1] and Kuniyoshi L. Sakai[1,2]*

[1]*Department of Cognitive and Behavioral Science, Graduate School of Arts and Sciences, The University of Tokyo, Komaba, Tokyo, Japan*
[2]*SORST, Japan Science and Technology Corporation, Kawaguchi-shi, Japan*

Abstract: When a speaker's voice returns to one's own ears with a 200-ms delay, the delay causes the speaker to speak less fluently. This phenomenon is called a delayed auditory feedback (DAF) effect. To investigate neural mechanisms of speech processing through the DAF effect, we conducted a functional magnetic resonance imaging (fMRI) experiment, in which we designed a paradigm to explore the conscious overt-speech processing and the automatic overt-speech processing separately, while reducing articulatory motion artifacts. The subjects were instructed to (1) read aloud visually presented sentences under real-time auditory feedback (NORMAL), (2) read aloud rapidly under real-time auditory feedback (FAST), (3) read aloud slowly under real-time auditory feedback (SLOW), and (4) read aloud under DAF (DELAY). In the contrasts of DELAY-NORMAL, DELAY-FAST, and DELAY-SLOW, the bilateral superior temporal gyrus (STG), the supramarginal gyrus (SMG), and the middle temporal gyrus (MTG) showed significant activation. Moreover, we found that the STG activation was correlated with the degree of DAF effect for all subjects. Because the temporo-parietal regions did not show significant activation in the comparisons among NORMAL, FAST, and SLOW conditions, we can exclude the possibility that its activation is due to speech rates or enhanced attention to altered speech sounds. These results suggest that the temporo-parietal regions function as a conscious self-monitoring system to support an automatic speech production system. *Hum. Brain Mapping 20:22–28, 2003.*

Key words: functional imaging; auditory system; speech production; self-monitoring; conscious processing; the superior temporal gyrus

INTRODUCTION

Temporal asynchrony between speech production and its feedback to the auditory system causes disruption of fluent speech [Lee, 1950; Yates, 1963]. This phenomenon, known as the delayed auditory feedback (DAF) effect, is observed as articulatory changes, such as slower speech rates, stuttering, intonation changes, and phoneme exchanges [Chapin et al., 1981]. According to a recent behavioral measurement, maximal disruptions occur at a delay of approximately 200 ms [Stuart et al., 2002]. While speech fluency is maintained without any conscious effort under real-time auditory feedback (RAF), conscious self-monitoring for overt-speech processing is required

Contract grant sponsor: Japan Science and Technology Corporation (JST); Contract grant sponsor: Human Frontier Science Program (HFSP).

*Correspondence to: Dr. Kuniyoshi L. Sakai, Department of Cognitive and Behavioral Science, Graduate School of Arts and Sciences, The University of Tokyo, Komaba, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan. E-mail: sakai@mind.c.u-tokyo.ac.jp

Received for publication 31 October 2002; Accepted 28 April 2003
DOI 10.1002/hbm.10119

under DAF in accordance with the change of auditory feedback. If the difference in the brain activations between the DAF and RAF conditions can be clarified, this may shed light not only on the mechanism of the DAF effect but on neural mechanisms of conscious and automatic overt-speech processing in general.

In spite of the significance of these issues, most of the previous studies on the DAF effect have been behavioral experiments, and there have been few brain imaging studies. A positron emission tomography (PET) study has reported activation increases in the bilateral superior temporal gyri (STG) for reading aloud single words with modified feedback (by pitch elevation or in someone else's voice) [McGuire et al., 1996]. Another PET study has reported that activation in the bilateral STG was observed during the overt-speech processing under DAF compared to the resting condition, but not during the overt-speech processing under RAF compared to the resting condition [Hirano et al., 1997]. These results suggest that the bilateral STG is recruited under speech conditions with abnormal feedback, but this activation may be confounded by either enhanced attention to altered speech sounds or by slower speech rates under DAF. Moreover, a direct comparison between DAF and RAF conditions should be conducted to identify critical regions for eliciting the DAF effect.

In the present study, we conducted a functional magnetic resonance imaging (fMRI) experiment to clarify the mechanism of the DAF effect and thereby infer neural mechanisms that support overt-speech processing. We investigated the brain activations during four conditions: (1) normal speech production under RAF (NORMAL), (2) rapid speech production under RAF (FAST), (3) slow speech production under RAF (SLOW), and (4) speech production under DAF (DELAY). The common regions activated in the comparisons of SLOW-NORMAL and DELAY-NORMAL, if any, would be primarily related to slower speech rates. In contrast, the comparisons among NORMAL, FAST, and SLOW conditions may reveal general effects for effortful production of accelerated or prolonged speech, which are typically associated with the DAF condition, as well as enhanced attention to those altered speech sounds. The fMRI technique has a higher spatial resolution than does PET, but the articulatory movements of the jaw and mouth may produce magnetic susceptibility changes that affect MR signals [Barch et al., 1999; Birn et al., 1999]. However, it has been suggested that the artifacts are reduced in the comparison between two conditions that equally include overt-speech processing [Barch et al., 1999]. In the present study, we adopted this procedure for minimizing the motion artifacts, together with scanning while not speaking.

SUBJECTS AND METHODS

Subjects

Eighteen native Japanese speakers (ages 21–37; 14 men and 4 women) participated in the present study. All subjects showed right-handedness. Twelve subjects participated in both experiment I (NORMAL, FAST, and DELAY) and II (NORMAL, SLOW, and DELAY), while six subjects were tested in either experiment (15 subjects in each experiment). The participant's head was immobilized with padding inside the radio-frequency coil, and the subject was instructed not to speak during inter-trial intervals. Informed consent was obtained from each participant after the nature and possible consequences of the studies had been explained. Approval for these experiments was obtained from the institutional review board of the University of Tokyo, Komaba.

Stimuli

We prepared 27 Japanese sentences, each of which consisted of seven *hiragana* letters: for example, "*reiseini kike*" (*Listen calmly*). We used a fixed number of *hiragana* letters to ensure constant reading time among all sentences, as one *hiragana* letter basically corresponds to one syllable. There are five vowels in Japanese: [a], [i], [u], [e], and [o]; they are classified into two groups according to articulatory positions of the tongue: front vowels ([i] and [e]) and back vowels ([a], [u], and [o]). When the back vowels are vocalized, the posterior part of the tongue is articulated, the front of the mouth cavity becomes narrower, and the back of the mouth cavity broadens out. On the other hand, when the front vowels are vocalized, the opposite changes occur. In our pilot experiments, we observed strong artifacts near the cerebral ventricles when the subjects pronounced the back vowels but not when they pronounced the front vowels. It is possible that the volume changes in the back of the mouth cavity lead to the artifacts. We, therefore, prepared stimuli that contained noise-free vowels of [i] and [e] alone.

We used an eyeglass-like MRI compatible display (resolution: 800 × 600) and a sound delivery system (VisuaStim XGA, Resonance Technology, Inc., Northridge, CA). The sentence stimuli were visually presented against a dark background at the center of the display. When a subject read the sentences into the headset microphone, his or her own voice was heard

through the headphone via the sound effecter (RFX-2000; ZOOM Corp., Tokyo, Japan). By using the sound effecter, we controlled the sound delivery from the microphone to the headphone with or without delay. In addition, the sound delivery was interrupted by using an intermittent timer during scanning periods, so that subjects did not hear the scanning noise through the headphone.

Tasks

Using a block design paradigm, we tested three types of overt-speech processing: NORMAL, FAST, and DELAY for experiment I, and NORMAL, SLOW, and DELAY for experiment II. Under the NORMAL and DELAY conditions, in which letters were presented in green, the subjects were asked to read the presented sentences overtly at normal speed. Under the FAST condition, in which letters were presented in purple, they were asked to read them at rapid speed; under the SLOW condition, in which letters were presented in purple, they were asked to read them at slow speed. The subject's voice was returned to the headphone without delay under the RAF conditions, whereas it was returned with 200 msec delay under the DAF condition.

Each sentence was presented for 2,800 msec, and the voice was fed back for 2,100 msec within this stimulus period. During the inter-stimulus interval of 2,200 msec that exactly coincides with the scanning period, the subjects were instructed to stop speaking until the next stimulus was presented, even if the subjects did not finish reading aloud the stimulus presented in the preceding trial. A red cross for fixation was always shown at the center of the display. We presented six consecutive stimuli in a single block of one session, which was in a sequence of either N-D-N-X-N-X-N-D-N or N-X-N-D-N-D-N-X-N (N, NORMAL; X, FAST or SLOW; D, DELAY). Each sentence was presented twice in one session and arranged not to appear twice in D or X. We tested 16 sessions (eight sessions for each sequence) for each subject, and the order of tasks was counterbalanced within and across subjects.

Behavioral Data Analyses

To assess the degree of the DAF effect on the subjects' speech performance, we measured the change in speech fluency as follows. First, we counted the number of morae pronounced correctly for each sentence and measured its reading time. Next, the mean spoken morae per second (M) under each condition were calculated. Finally, we calculated a DELAY index defined by $(1 - M_{DELAY}/M_{NORMAL}) \times 100$. Its positive value indicates that the subjects showed less fluent speech under the DELAY condition. We also assessed speech fluency during the FAST or SLOW condition by using a fluency index $(1 - M_X/M_{NORMAL}) \times 100$, where X is FAST or SLOW.

fMRI Data Acquisition and Analyses

The fMRI scans were conducted using a 1.5-T MRI scanner (STRATIS II, Premium; HITACHI, Tokyo, Japan). We scanned over 15 horizontal slices, each 7 mm thick, covering from z = −49 to 56 mm, with a gradient echo echo-planner imaging sequence (repetition time = 5 sec; echo time = 50.5 ms; acquisition time, 2,200 msec; flip angle, 90°; field of view, 192 × 192 mm^2; resolution, 3 × 3 mm^2). To eliminate articulatory motion artifacts during scanning, the scanning sounds were confined within the inter-stimulus interval by using a clustered volume acquisition sequence [Edmister et al., 1999].

For analyses of functional data, we used statistical parametric mapping software (SPM99, Wellcome Department of Cognitive Neurology, London, UK). We removed sessions that included data with a translation of more than 2 mm or a rotation of more than 1.2° in one of the three directions. The data were realigned, spatially normalized to the standard brain space, resampled every 3 mm using sinc interpolation, and smoothed with an isotropic Gaussian kernel of 14 mm full width at half maximum. Low-frequency noise and global changes in activity were further removed. Task-specific effects were estimated with a general linear model using a boxcar waveform convolved with the canonical hemodynamic response function. For random effects analyses, a contrast image between tasks was generated for each participant and used for inter-subject comparisons. A statistical threshold was set at $P < 0.05$ for the voxel level, corrected for multiple comparisons. For the anatomical identification of activated regions, we used the Anatomical Automatic Labeling method [Tzourio-Mazoyer et al., 2002].

RESULTS

All subjects showed the DAF effect, and the DELAY index for each subject was always positive (Table I). As instructed, subjects spoke more rapidly under the FAST condition than NORMAL and spoke more slowly under the SLOW condition than NORMAL, as shown by the negative and positive fluency indices, respectively. Generally, speech under DAF is less fluent until speakers begin to disregard the auditory

TABLE I. Speech rates in all conditions*

Experiment	DELAY	FAST	SLOW
I	23.8 ± 13.8 (6.8∼53.9)	−62.0 ± 14.0 (−82.2∼−24.0)	—
II	18.4 ± 12.5 (4.0∼41.1)	—	36.4 ± 10.4 (16.9∼49.7)

* DELAY indices and fluency indices (see Materials and Methods) are shown in mean ± SD (range; n = 15).

feedback of their own speech. However, under the present condition of sufficient feedback of speech, the same individuals did not show significant reduction of the DAF effect, i.e., the change of the DELAY indices, between the experiments I and II (paired *t*-test, $P > 0.4$).

We first identified the activated regions under the DELAY condition. In DELAY–NORMAL of the two experiments, significant activations were observed mostly in the bilateral superior temporal gyrus (STG) and the supramarginal gyrus (SMG) that extended into the middle temporal gyrus (MTG) (Fig. 1A, Table II). In DELAY–FAST, a similar pattern of activation was evident (Fig. 1B), and the right temporo-parietal regions showed significant activation in DELAY-SLOW (Table II). These results demonstrated that the bilateral temporo-parietal regions were more preferentially activated under the DAF condition than the other RAF conditions.

Next, we tested whether the activation of the temporo-parietal regions were predictive of how much each individual subject showed the DAF effect. We found that the signal changes of the left STG in DELAY–NORMAL were significantly correlated with the DELAY index among the subjects (Fig. 2; $r = 0.73$, $P < 0.01$). Moreover, the right STG also showed correlation with the degree of DAF effect ($r = 0.59$, $P < 0.05$). Therefore, we conclude that activation in the bilateral temporo-parietal regions is a good indicator of the degree of the DAF effect.

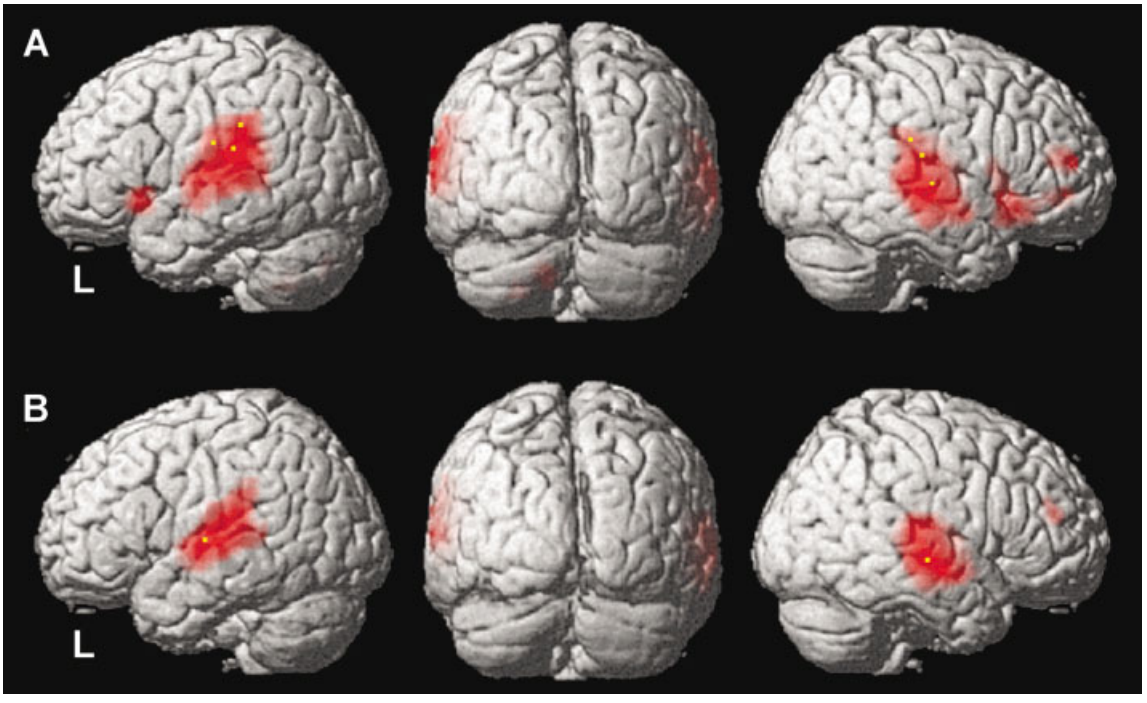

Figure 1.

Activations under speech conditions with delayed auditory feedback. **A:** Regions identified by DELAY–NORMAL. Activated regions are rendered on the surface of a standard brain. Left to right: the left lateral view, the posterior view, and the right lateral view. Note that the bilateral temporo-parietal regions showed significant activation. **B:** Regions identified by DELAY–FAST. Similar activation patterns were observed in the bilateral temporo-parietal regions. For display purposes, the threshold is set at uncorrected $P < 0.0005$ with an extent threshold of 22 voxels, but the local maxima of *t*-values (yellow dots) reached a threshold of $P < 0.05$, corrected for multiple comparisons.

TABLE II. Brain regions selectively associated with the DAF effect*

Brain region	Side	*x*	*y*	*z*	*Z*
DELAY > NORMAL					
Superior temporal g	L	−60	−30	21	5.7
Supramarginal g	L	−57	−33	33	5.5
Postcentral g	L	−63	−18	24	5.3
Superior temporal g	R	60	−18	3	5.2
Supramarginal g	R	60	−30	27	5.5
Supramarginal g	R	66	−21	18	5.4
DELAY > FAST					
Superior temporal g	L	−60	−12	3	5.2
Middle temporal g	R	60	−12	−12	5.2
DELAY > SLOW					
Middle temporal g	R	66	−27	−6	5.0

* Stereotactic coordinates (*x, y, z*) in Montreal Neurological Institute (MNI) space are shown for each voxel with a local maximum of *Z* values in the contrasts indicated ($P < 0.05$, corrected for the voxel-level; shown as yellow dots in Fig. 1). g, gyrus.

Because the same RAF condition was employed under the NORMAL, FAST, and SLOW conditions, the comparisons among these conditions would reveal activations that reflect changes in speech rates when speaking fluently. FAST–NORMAL resulted in significant activation of the left insula [(−30, 18, 6), $Z = 4.5$] and the caudate [(−9, 0, 9), $Z = 4.5$], which may be related to the motor control of accelerated speech production, together with the left cuneus [(−21, 45, 24), $Z = 4.6$]. In contrast, we observed no significant activation in NORMAL–FAST and SLOW–NORMAL, which are the contrasts for slower speech production. Furthermore, FAST–DELAY and SLOW–DELAY did not elicit significant activation. Finally, NORMAL–DELAY (experiment I) resulted in significant activation of the superior prefrontal cortex [(−9, 54, 0), $Z = 5.1$; (−12, 66, 6), $Z = 4.5$], the middle frontal cortex [(−3, 45, −6), $Z = 5.4$; (6, 57, −3), $Z = 4.9$], and the bilateral fusiform gyrus [(−45, −51, −21), $Z = 4.5$; (36, −60, −15), $Z = 4.6$], whereas NORMAL–SLOW resulted in activation of the left middle occipital gyrus [(−42, −72, 33), $Z = 4.8$].

DISCUSSION

In the present study, we observed significant activations in the temporo-parietal regions, when the DAF condition (DELAY) was compared with the RAF conditions (NORMAL, FAST, and SLOW). Furthermore, we found that the bilateral STG activation was correlated with the degree of the DAF effect. By introducing both FAST and SLOW conditions in the present study, we successfully eliminated the possible confounding factors of the slower speech rates occurring under DAF, as well as enhanced attention to those altered speech sounds. Another possible reason for the enhanced activation in the temporo-parietal regions is that the subjects heard their own voice for a longer time under DELAY than NORMAL and FAST, because they spoke less fluently under the DELAY condition. However, this possibility can be excluded by noting that significant activation was not observed in the temporo-parietal regions in NORMAL–FAST and SLOW–NORMAL, which should have shown the same effect. Here we propose that an additional system of the bilateral temporo-parietal regions is recruited when conscious self-monitoring is required under the DAF condition, in addition to the automatic speech production system that is used under the RAF conditions.

It is interesting to note that the auditory and association cortices of the bilateral temporo-parietal regions show selective activation modulation for the DAF effect, but motor and premotor cortices do not. Recent fMRI studies have shown that the bilateral STG is related to auditory attention and conscious awareness of auditory stimuli [Hashimoto et al., 2000; Pugh et al., 1996], which is consistent with the fact that feedback control is conscious processing. However, the absence of significant activation in SLOW–NOR-

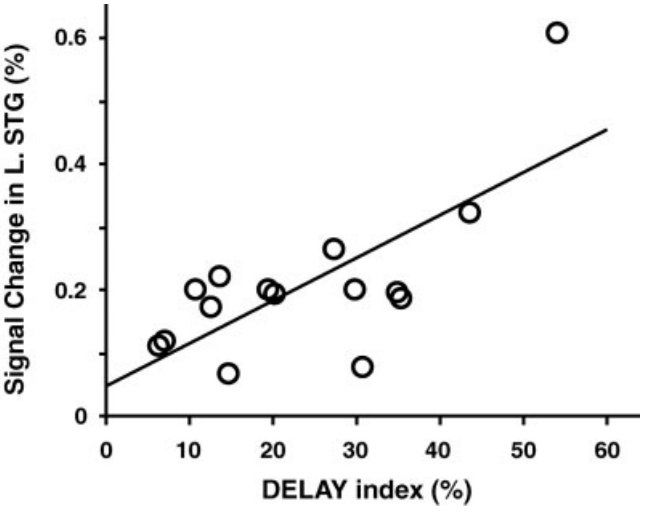

Figure 2.
A correlation between the degree of the DAF effect and the signal change in the left STG [the most ventral local maximum at (−60, −30, 21) in Fig. 1A]. Among the subjects, the DELAY index [$(1 - M_{DELAY}/M_{NORMAL}) \times 100$], where M is the mean spoken morae per second, showed significant correlation with the signal change in the left STG for the comparison DELAY-NORMAL, as fitted with a straight line.

MAL suggests that enhanced attention to prolonged speech sounds in SLOW is not comparable to the DAF effect. Moreover, the correlation between the bilateral STG activation and the DAF effect indicates that the more STG is activated, the more it interferes with controlling one's own overt speech. This possibility is in agreement with the *perceptual* theory of self-monitoring, such that speakers have little or no access to their speech production process, and that self-monitoring is probably based on parsing one's own inner or overt speech [Levelt, 1983]. Furthermore, both behavioral and theoretical studies have shown that articulatory errors tend to be more perseveratory (e.g., "beef needle soup") than anticipatory (e.g., "cuff of coffee") when the error rate is higher [Dell et al., 1997]. Consistent with this serial-order model, the higher error rate accompanied by the DAF effect may be due to perseveratory errors. For example, when "beef" is heard with DAF while trying to speak "noodle," the speech output may result in erroneous "needle." This explains why the mechanism of the DAF effect mainly involves the auditory cortex and it leads to disruptions of self-monitoring.

In NORMAL–DELAY and NORMAL–SLOW, we observed activation in the medial prefrontal cortex and occipital regions, which suggests that these regions were significantly deactivated during DELAY and SLOW than NORMAL. Recent imaging studies have demonstrated that these regions show task-induced deactivation [Gusnard et al., 2001; Raichle et al., 2001; Stark and Squire, 2001]. It has been proposed that these areas are involved in default brain activity, such as day-dreaming, self-reflection, and problem solving. This default activity decreased during a high-load task condition in comparison with a low-load task condition, since subjects must pay more attention to the former condition. It follows from this discussion that the subjects probably paid more attention to the DELAY condition than other conditions, which is consistent with our proposal that conscious self-monitoring is involved in overt-speech processing under DAF.

Speech production experiments have been hampered in fMRI studies, because articulatory movements can potentially produce motion artifacts. In previous fMRI studies of overt-speech processing, the event-related paradigm was preferred to the block-design paradigm [Huang et al., 2001; Palmer et al., 2001], because the event-related paradigm can minimize motion-artifacts by taking advantage of the different temporal characteristics of the hemodynamic response and motion-related signal changes [Birn et al., 1999]. However, a recent fMRI study demonstrated that the block-design paradigm could obtain artifact-free images when one compared two conditions that both used overt-speech processing, because motion-related artifacts were subtracted out [Barch et al., 1999]. Therefore, we used task conditions that all contained overt-speech processing. In addition, we were able to reduce possible motion artifacts in our block-design fMRI experiments by selecting vowels, together with a clustered volume acquisition sequence. The present study thus extends the limit of fMRI experiments for understanding cortical mechanisms of audition and speech production.

ACKNOWLEDGMENTS

We thank Dr. Ryuichiro Hashimoto, Dr. Fumitaka Homae, Mr. Kei Suzuki, and Mr. Yasuki Noguchi for their technical assistance; and Ms. Hiromi Matsuda for her administrative assistance. This work was supported in part by a SORST grant from JST (K.L.S.) and by a Young Investigators' Grant from HFSP (K.L.S.).

REFERENCES

Barch DM, Sabb FW, Carter CS, Braver TS, Noll DC, Cohen JD (1999): Overt verbal responding during fMRI scanning: Empirical investigations of problems and potential solutions. Neuroimage 10:642–657.

Birn RM, Bandettini PA, Cox RW, Shaker R (1999): Event-related fMRI of tasks involving brief motion. Hum Brain Mapp 7:106–114.

Chapin C, Blumstein SE, Meissner B, Boller F (1981): Speech production mechanisms in aphasia: a delayed auditory feedback study. Brain Lang 14:106–113.

Dell GS, Burger LK, Svec WR (1997): Language production and serial order: a functional analysis and a model. Psychol Rev 104:123–147.

Edmister WB, Talavage TM, Ledden PJ, Weisskoff RM (1999): Improved auditory cortex imaging using clustered volume acquisitions. Hum Brain Mapp 7:89–97.

Gusnard DA, Akbudak E, Shulman GL, Raichle ME (2001): Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. Proc Natl Acad Sci U S A 98:4259–4264.

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

Hirano S, Kojima H, Naito Y, Honjo I, Kamoto Y, Okazawa H, Ishizu K, Yonekura Y, Nagahama Y, Fukuyama H, Konishi J (1997): Cortical processing mechanism for vocalization with auditory verbal feedback. Neuroreport 8:2379–2382.

Huang J, Carr TH, Cao Y (2001): Comparing cortical activations for silent and overt speech using event-related fMRI. Hum Brain Mapp 15:39–53.

Lee BS (1950): Some effects of side-tone delay. J Acoust Soc Am 22:639–640.

Levelt WJM (1983): Monitoring and self-repair in speech. Cognition 14:41–104.

McGuire PK, Silbersweig DA, Frith CD (1996): Functional neuroanatomy of verbal self-monitoring. Brain 119:907–917.

Palmer ED, Rosen HJ, Ojemann JG, Buckner RL, Kelley WM, Petersen SE (2001): An event-related fMRI study of overt and covert word stem completion. Neuroimage 14:182–193.

Pugh KR, Shaywitz BA, Shaywitz SE, Fulbright RK, Byrd D, Skudlarski P, Shankweiler DP, Katz L, Constable RT, Fletcher J, Lacadie C, Marchione K, Gore JC (1996): Auditory selective attention: an fMRI investigation. Neuroimage 4:159–173.

Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL (2001): A default mode of brain function. Proc Natl Acad Sci U S A 98:676–682.

Stark CEL, Squire LR (2001): When zero is not zero: the problem of ambiguous baseline conditions in fMRI. Proc Natl Acad Sci U S A 98:12760–12766.

Stuart A, Kalinowski J, Rastatter MP, Lynch K (2002): Effect of delayed auditory feedback on normal speakers at two speech rates. J Acoust Soc Am 111:2237–2241.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002): Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–289.

Yates AJ (1963): Delayed auditory feedback. Psychol Bull 60:213–232.