ACADEMIC PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

NeuroImage

NeuroImage 20 (2003) 1865–1871

www.elsevier.com/locate/ynimg

Automated method for extracting response latencies of subject vocalizations in event-related fMRI experiments☆

J.L. Nelles, H.M. Lugar, R.S. Coalson, F.M. Miezin, S.E. Petersen, and B.L. Schlaggar*

Departments of Neurology and Neurological Surgery, Pediatrics, Radiology, Anatomy and Neurobiology, and Psychology, Washington University School of Medicine, St. Louis, MO 63110, USA

Received 14 May 2003; revised 21 July 2003; accepted 23 July 2003

Abstract

For functional magnetic resonance imaging studies of the neural substrates of language, the ability to have subjects performing overt verbal responses while in the scanner environment is important for several reasons. Most directly, overt responses allow the investigator to measure the accuracy and reaction time of the behavior. One problem, however, is that magnetic resonance gradient noise obscures the audio recordings made of voice responses, making it difficult to discern subject responses and to calculate reaction times. ASSERT (Adaptive Spectral Subtraction for Extracting Response Times), an algorithm for removing MR gradient noise from audio recordings of subject responses, is described here. The signal processing improves intelligibility of the responses and also allows automated extraction of reaction times. The ASSERT-derived response times were comparable to manually measured times with a mean difference of −8.75 ms (standard deviation of difference = 26.2 ms). These results support the use of ASSERT for the purpose of extracting response latencies and scoring overt verbal responses.

Introduction

The use of functional magnetic resonance imaging (fMRI) to study the neural substrates of language has proven to be effective (see e.g., Ojemann et al., 1998; Bookheimer, 2002; Phelps et al., 1997; Turkeltaub et al., 2002; Binder et al., 1997), but many studies have avoided the use of overt verbal responding, despite the common use of this response modality in behavioral and cognitive language studies. Two major problems have been identified in the use of verbal responses.

First, movement and susceptibility artifacts are expected to be produced that affect the statistical quality of images (Barch et al., 1999). This problem has been addressed through the use of event-related designs and special pulse sequences (Barch et al., 1999; Birn et al., 1999; Eden et al., 1999; Palmer et al., 2001; de Zubicaray et al., 2001; Huang et al., 2002).

The second problem is that echo planar imaging bold oxygenation-level dependent (EPI BOLD) sequences generate gradient noise that obscures the subject's spoken responses (Munhall et al., 2001). As a consequence, a simple microphone threshold is often insufficient to detect the response onset, making it difficult to measure the voice reaction time. Furthermore, the noise hinders assessment of the content of the subject's response, which is critical to some event-related experiments (e.g., Schlaggar et al., 2002). Making this problem more difficult, the EPI sequence gradient noise spectrum overlaps the human speech spectrum, so simple filtration is ineffective in removing the noise without also degrading the speech signal. Thus, response times were previously extracted manually by measuring the time from stimulus presentation to voice response onset from the recorded audio signal in a sound editing program (Palmer et al., 2001; Schlaggar et al., 2002). This process is labor-intensive and somewhat subjective, which may introduce some user variability.

☆ Source documentation and code can be obtained by E-mailing a request to assert@nil.wustl.edu.

* Corresponding author. Department of Neurology, Washington University School of Medicine, Campus Box 8111, 660 South Euclid Ave., St. Louis, MO 63110, USA. Fax: +1-314-362-6110.

E-mail address: schlaggar_b@kids.wustl.edu (B.L. Schlaggar).

1053-8119/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S1053-8119(03)00471-3

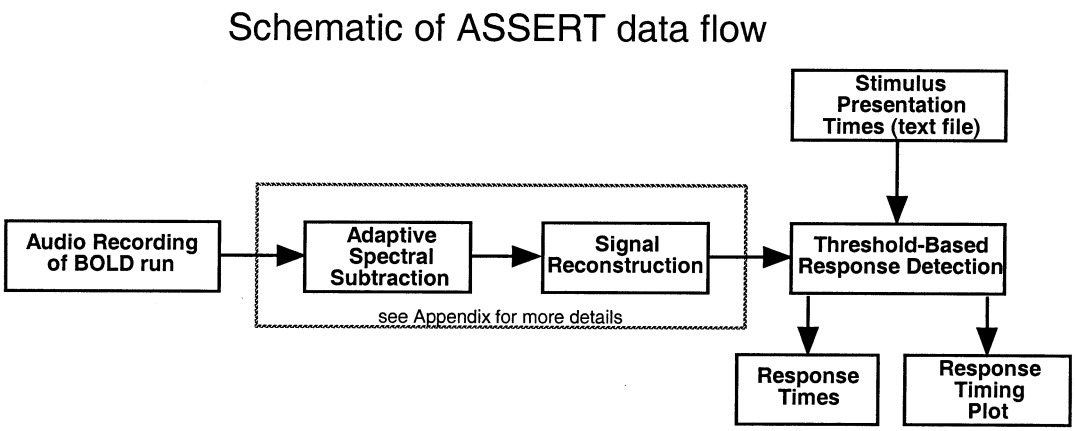


Fig. 1. The major processing steps in ASSERT. The audio recording of subject vocalizations during the BOLD run is initially filtered with an adaptive spectral subtraction algorithm. After the signal is reconstructed, the times at which the signal exceeds a threshold are compared to the inputted stimulus presentation times to extract response times. A graphical plot of the detected responses is also generated.

ASSERT (Adaptive Spectral Subtraction for Extracting Response Times) was developed to address this second problem by removing much of the MR gradient noise from the auditory recordings, both improving the voice response intelligibility and facilitating automated extraction of reaction time latencies based on a simple signal amplitude measurement.

Methods

Recording apparatus

Audio recordings of spoken responses from subjects were made during fMRI image acquisition on a Siemens MAGNETOM Vision 1.5-T scanner (Erlangen, Germany). Functional images were acquired using a nearly continuous acquisition EPI BOLD sequence (TR = 2.5 s, 16 slices, 3.75 × 3.75 × 8-mm voxels), comparable to the EPI sequences used for other fMRI studies (Kelley et al., 1998; Palmer et al., 2001; Schlaggar et al., 2002), with a short delay of 50 ms between whole brain (frame) acquisitions to allow manual timing. Subjects were presented with a visual stimulus and instructed to respond vocally. Stimulus timing was synchronized to the MR scanner and presented using Psyscope (Cohen et al., 1993) with words (all nouns) back-projected onto a screen visible using an MR-compatible mirror. In each run, subjects were instructed to either read the word aloud or generate a verb applicable to the noun presented. The responses were collected using the Resonance Technology Commander XG MRI audio system (Northridge, CA). The signal from the Audio System's microphone was recorded directly into a PC (Micron PC with a Pentium III @ 450 MHz) running CoolEdit 2000 (Syntrillium Software Corp., Phoenix, AZ) to create a digitized sound file for processing using ASSERT. To minimize file sizes, 8-bit recordings were made while sampling at 11025 Hz.

Manual calculation of response times

Manual response time extraction involved viewing and listening to the digital sound file in a sound editing program (SoundEdit 16, Macromedia, San Francisco, CA) and then measuring the time from each stimulus presentation to the first point at which the response became visible and/or audible. Visual clues include the presence of signal peaks above the background noise level and the presence of signal frequencies not found in the response-free sections of the signal. Using image acquisition sequences with a 50-ms delay between whole brain (frame) acquisitions, the gradient noise manifests as bursts of audible "beeping" with intervening short pauses. In SoundEdit16, it was possible to measure the time from the onset of a burst (coincident with a stimulus presentation) to the apparent beginning of a subject response. The 50-ms gap in the imaging sequence is ideal for manual timing to minimize user labor and errors.

Signal processing algorithm

ASSERT is an adaptive spectral subtraction algorithm (Boll, 1979) and a threshold-based response time detection program created using MATLAB (The MathWorks; Natick, MA). Spectral subtraction is a signal processing algorithm that works, in this case, on the principal that the frequency spectrum of the recorded signal is the sum of a vocal response spectrum and a noise spectrum. The assumption is that the frequency spectrum of the noise does not vary significantly from moment to moment, but the amplitude of the noise may vary. The magnitude spectrum of a sample of extracted noise is determined separately and then the amplitude of its component in short time segments of the signal is estimated before being subtracted from the spectrum of those short segments. The resulting "cleaned" signal is then reconstructed using the phase of the original signal. This approach allows the entire spectrum of the noise to be removed while leaving the speech signal essentially intact.

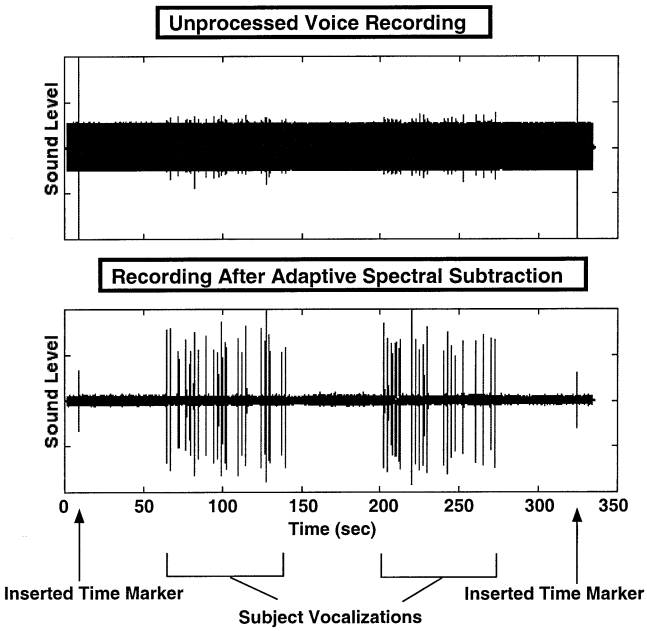


Fig. 2. Digital recording before and after processing using ASSERT. The two spikes at the beginning and end of the signals are inserted markers, and the prominent peaks in both the upper and lower signal traces are the spoken subject responses.

Technique (Fig. 1)

An audio recording of the subject responses during a multiframe BOLD run is collected in ".wav" format. ASSERT requires indicating both the duration of the BOLD run (with BEGIN and END markers inserted in the .wav file (see Fig. 2)) and the stimulus presentation times relative to the start of the BOLD run in a text file. ASSERT then applies the adaptive spectral subtraction algorithm to the .wav file, determines a threshold for voice responses, graphically plots the stimulus presentation times and response events, and returns a text file containing the response latencies.

Four fundamental components of ASSERT (see Appendix for comprehensive flowchart)

(1) Fourier transformation of the vocalization signal

The vocalization signal is sampled for short time windows (~93 ms) for processing. The discrete Fourier transform is calculated for each time window. The product of this transform is then separated into its real and imaginary parts. The magnitude and phase spectra are computed. The signs of each real and imaginary component are preserved for later reconstruction.

(2) Noise sampling and scaling

A sample of gradient noise from the beginning of the recording is extracted at a time when there is no subject vocalization present. This sample can be extracted without user input because experiments were designed to not have responses during the first three or four image (frame) acquisitions. The magnitude spectrum of the gradient noise sample is computed and resampled to the same size as the magnitude spectrum of the time window. With the magnitude spectra computed, several arrays for scaling the noise spectrum are then created. The first array, α_f, contains the ratio of the magnitude spectrum in the original signal to the noise spectrum for each element in the discrete spectrum. A second value, α_p, is computed as the maximum ratio at a number of spectral noise peaks identified as being outside the human voice spectrum. α_p serves as a "cap" for the other ratios so that the entire magnitude spectrum is not subtracted. A final array, α_k, is computed as the minimum of α_p or α_f at each frequency. Because α_p is calculated at a frequency that consists mostly of noise, those peaks with a larger ratio in α_f contain both noise and voice response, while those with a lower ratio in α_f contain only noise. Therefore, only the cap, α_p, is used for scaling where α_f exceeds α_p. α_k is then scaled by another constant that dictates what percentage of the noise to remove. This array α_k is used to scale the gradient noise spectrum before subtracting it from the spectrum of the short signal window.

(3) Reiteration to reconstruct vocalization signal

Using the subtracted magnitude spectrum, the original (noisy) phase spectrum, and the signs of the original real and imaginary parts of the complex spectrum, the time domain representation of the short window is reconstructed with an inverse Fourier transform. The processing window is then shifted one-fourth of a window width and the process repeated, with each window overlapping the previous windows and added to the existing signal to create the final reconstructed signal. Additional filtering removes most of the remaining noise (elliptical design; high-pass filter, f_c = 350 Hz; low-pass filter, f_c = 3200 Hz; and a series of narrowband "notch" filters to remove several noise peaks in the signal spectrum (60, 674, 958, 1977, 3265, and 3283 Hz)). This allows a lower threshold to be used for response detection as described below.

(4) Calculation of a simple response threshold

With the noise removed, a signal amplitude threshold can be applied to detect the location of vocal responses. This threshold is calculated from the maximum amplitude of noise prior to the first response and after the last response. The first point at which the signal crossed this threshold following a stimulus presentation is recorded as the response latency.

Subjects and data collection

Responses were recorded from two healthy adult subjects with two BOLD runs each. The four BOLD runs consisted of 34–40 visually presented single words, us-

Table 1
Manually vs ASSERT measured reaction times

Task:	Subject 1		Subject 2		Total (n = 147)
	Read (n = 40)	Read (n = 39)	Read (n = 34)	Verb gen. (n = 34)	
Manual mean	500.6	517.72	730.12	1216.24	
ASSERT mean	486.96	503.9	735.87	1204.56	
Mean difference	−13.64	−13.81	5.75	−11.67	−8.75
SD of difference	14.28	12.53	34.84	32.51	26.20
SD of responses	50.50	77.66	118.37	502.24	

Note. Mean differences and standard deviation of differences between manual and ASSERT times for the 4 BOLD runs. All numbers are shown in milliseconds. The standard deviation of responses is calculated from the ASSERT measurements for each task. The response standard deviation for the second task in subject 2 is much higher because the task is different.

ing standard simple read and verb generate paradigms, as described above. After one stimulus was discarded due to a subject's failure to respond, the collected data set consisted of a total sample size of $n = 147$ discrete vocalizations.

For validation, the response times derived via ASSERT were compared to those derived manually. To minimize user bias, the manual response times were determined before running ASSERT.

Results

A representative BOLD run is depicted in Fig. 2. Qualitatively, the subject responses are visually barely discernible from the underlying gradient noise in the raw signal. By contrast, in the ASSERT processed signal the responses are readily evident.

Table 1 shows the mean differences between automatic and manual timing for each BOLD run, as well as the mean across BOLD runs and the standard deviation of the differences. In three of four BOLD runs, ASSERT averaged a shorter response time than manual detection yielded. This suggests that ASSERT was able to detect vocalizations in the processed .wav file sooner than what was manually detected in the raw sound file.

A comparison of ASSERT to manual timing is shown in Fig. 3, a plot of automated vs manual response times for all four BOLD runs. Not only is the slope of the fitted line close to the expected value of 1 ($m = 1.0012$), but the correlation coefficient of 0.9953 is also remarkably close to an ideal value of 1.

Discussion

The main finding from this study is that ASSERT consistently yielded response times similar to those derived manually. Both the mean and the standard deviation of differences in the response times were smaller than the standard deviation of response times for the tasks used even for simple word reading. From Table 1, the averaged difference between manual and ASSERT response times for a given BOLD run fell into a range of roughly ±15 ms. The average difference across all four BOLD runs was less than 10 ms (total standard deviation of difference = 26.2 ms), and the correlation coefficient of 0.9953 from the linear regression in Fig. 3 also bears out the validity of those times. The slope of the fit line is almost exactly 1 (1.0012), implying that the accuracy of the automated responses is not dependent upon the length of the response times.

This finding suggests that ASSERT can be used to accurately detect response times without concern of losing statistical differences in response times between tasks or item types. A pilot test of a previous version of the software showed that subtle effects of frequency by regularity remain at least as significant when processing using ASSERT as when manually timing (Nelles et al., 2002). A more rigorous test would be necessary to conclude that ASSERT can distinguish subtle differences as small as 10 ms between any two tasks.

Additionally, the program produced no false positives or false negatives for responses in these BOLD runs. That is, ASSERT never detected noise as a response, and never failed to detect a response that was present. In the second BOLD run, our subject failed to respond to one stimulus, and no reaction time was returned for that stimulus. However, the possibility of a missed response remains for softly spoken responses, and false positives are likely if the subject were to respond multiple times to a given stimulus. In fact, the sensitivity of the response detection is such that responses may be recorded if the subject clears his throat or breathes directly onto the microphone. If multiple responses do occur between a pair of stimulus presentations, the response detection algorithm currently records only the first response after a stimulus. One of the outputs of the algorithm (the Response Timing Plot, see Fig. 1) generates a figure of the timing of the detected responses superimposed on the processed voice recording which can be used to screen for these types of problems.

Another possible source of error is that the response detection is dependent upon the signal amplitude reaching a

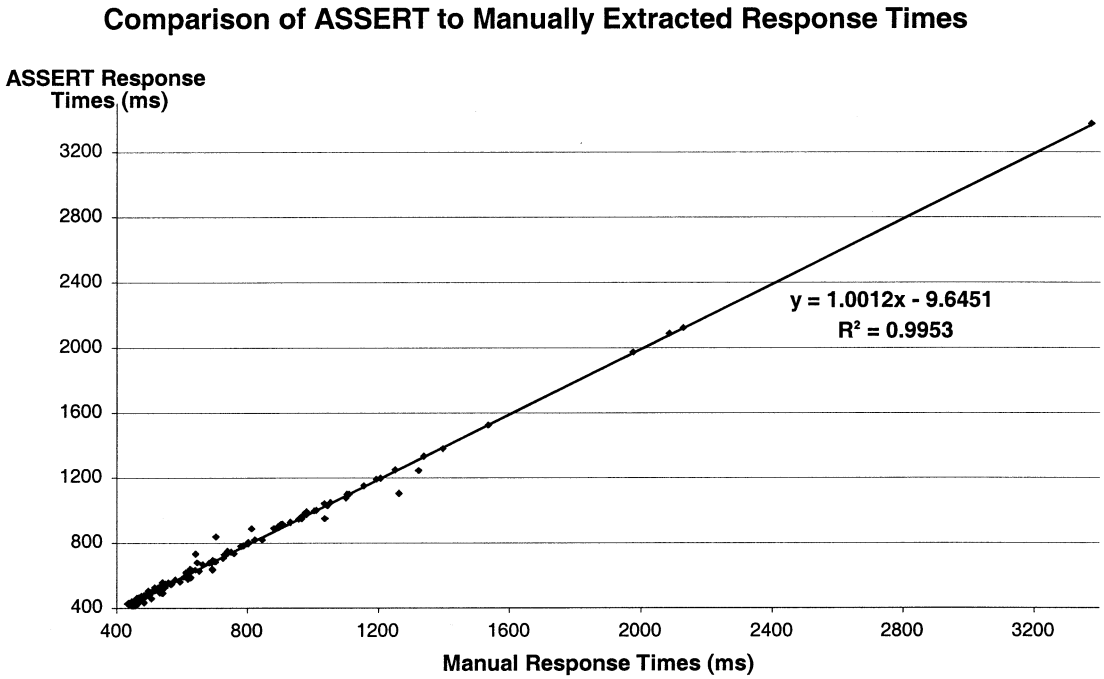


Fig. 3. Comparison of individual ASSERT times to manually extracted time.

threshold. Therefore, the beginning of "soft" sounding words may have a short delay before detection compared to words with more plosive beginnings. If an experiment were designed to compare response times where sibilant and plosive words were highly unbalanced between tasks, it could introduce a small difference in response times, although this has not been directly investigated.

Another important application which has not been thoroughly explored is the use of ASSERT in higher field magnets. An older version of ASSERT has been tested in a Siemens MAGNETOM 3-T Allegra scanner, with results that appeared to be comparable to those from the same version in our 1.5-T machine. There are several reasons to expect that ASSERT is applicable to higher field machines. First, the sensitivity of response detection suggests that a higher level of residual noise is acceptable. Second, because the algorithm uses a noise estimate derived from each recording, the difference in the noise spectra should not effect the performance. Finally, the postfiltering can be optimized for each scanner to remove distracting residual noise.

The results of using our spectral subtraction method could not be compared to other signal processing techniques because our other attempts failed to produce a signal which was acceptable for voice response detection. Because the noise and voice spectra overlap across broad spectral bands, filters that remove the noise also degrade the voice signal considerably. Our attempt to separate the voice signal from the MR gradient noise using available independent component analysis code was unable to separate the signals sufficiently to create a threshold above the noise that did not also exclude most voice responses. The minimum subject response amplitude for ASSERT to detect a vocalization has not been quantified in decibels. However, the entire range of volumes from children is detected in its current use, with responses rarely if ever missed.

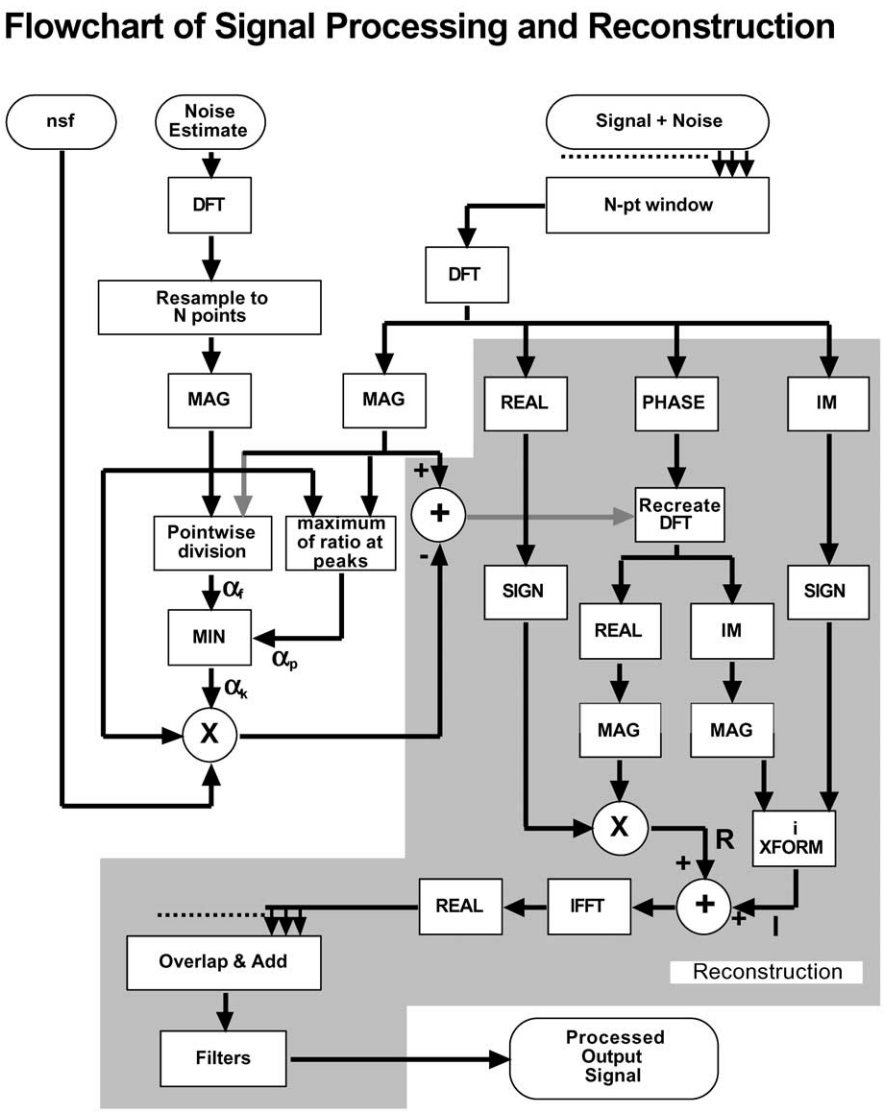
Conclusions

Our adaptive spectral subtraction algorithm, ASSERT, was able to extract response latencies of subject vocalizations quickly, accurately, and reliably. The response times correlated well with manually extracted response times, and with only a small average difference. No responses were missed, and no noise peaks were incorrectly scored as responses. Not only does this software eliminate the need for manual scoring of response times, but the removal of noise from the recording also makes behavioral scoring of the correctness of responses both simpler and more user-friendly.

Acknowledgments

This work was approved by the Washington University Human Studies Committee and was supported by grants from the McDonnell Center for the Study of Higher Brain Function, NIH Grants LM06858 (S.E.P.), NS41255 (S.E.P.), and NS55582 (B.L.S.). B.L.S. is a Scholar of the Child Health Research Center of Excellence in Developmental Biology at Washington University School of Medicine (HD01487).

Appendix



The input signals and parameter are listed on the top; the processed output signal on the bottom. "Signal + Noise" refers to the audio recording of subject vocalizations during the BOLD run. The "Noise Estimate" is a short time sample during the initial part of that recording when no subject vocalizations are present. Abbreviations: DFT, discrete Fourier transform; MAG, the unsigned magnitude of the complex input array; REAL, real part of complex input array; IM, imaginary part of the complex input array; SIGN, sign (+/−) of the complex input array; PHASE, phase of the complex input array; MIN, minimum value; nsf, noise scale factor (constant); N-pt window, a sample with N data points of a segment of the input; Recreate DFT, combine a magnitude and a phase input to create a new DFT; i XFORM, convert to imaginary part of complex number; IFFT, inverse discrete Fourier transform; α_f, calculated scaling array; α_k, applied scaling array; α_p, cap (maximum) scaling array from selected noise only peaks.

References

Barch, D.M., Sabb, F.W., Carter, C.S., Braver, T.S., Noll, D.C., Cohen, J.D., 1999. Overt verbal responding during fMRI scanning: empirical investigations of problems and potential solutions. NeuroImage 10, 642–657.

Binder, J., Frost, J., Hammeke, T., Cox, R., Rao, S., Prieto, T., 1997. Human brain language areas identified by functional magnetic resonance imaging. J. Neurosci. 17 (1), 353–362.

Birn, R.M., Bandettini, R.A., Cox, R.W., Shaker, R., 1999. Event-related fMRI of tasks involving brief motion. Human Brain Mapp. 7 (2), 106–114.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoustics, Speech, Signal Process. 27, 113–120.

Bookheimer, S., 2002. Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. Annu. Rev. Neurosci. 25, 151–188.

Cohen, J.D., MacWhinney, B., Flatt, M., Provost, J., 1993. Psyscope: a new graphic interactive environment for designing psychology experiments. Behav. Res. Methods Instruments Computers 25, 257–271.

Eden, G.F., Joseph, J.E., Brown, H.E., Brown, C.P., Zeffiro, T.A., 1999. Utilizing hemodynamic delay and dispersion to detect fMRI signal change without auditory interference: the behavior interleaved gradients technique. Magn. Reson. Med. 41 (1), 13–20.

Huang, J., Carr, T.H., Cao, Y., 2002. Comparing cortical activations for silent and overt speech using event-related fMRI. Human Brain Mapp. 15 (1), 39–53.

Kelley, W.M., Miezin, F.M., McDermott, K.B., Buckner, R.L., Raichle, M.E., Cohen, N.J., Ollinger, J.M., Akbudak, E., Conturo, T.E., Snyder, A.Z., Petersen, S.E., 1998. Hemispheric specialization in human dorsal frontal cortex and medial temporal lobe for verbal and nonverbal memory encoding. Neuron 20 (5), 927–936.

Munhall, K.G., 2001. Functional imaging during speech production. Acta Psychol.(Amst) 107 (1–3), 95–117.

Nelles, J.L., Lugar, H.M., Coalson, R.S., Miezin, F.M., Petersen, S.E., Schlaggar, B.L., 2002. An automated technique for extracting reaction times from vocalizations recorded during fMRI. Poster presented at Cognitive Neuroscience Society, April 2002, San Francisco, CA.

Ojemann, J.G., Buckner, R.L., Akbudak, E., Snyder, A.Z., Ollinger, J.M., McKinstry, R.C., Rosen, B.R., Petersen, S.E., Raichle, M.E., Conturo, T.E., 1998. Functional MRI studies of word-stem completion: reliability across laboratories and comparison to blood flow imaging with PET. Human Brain Mapp. 6 (4), 203–215.

Palmer, E.D., Rosen, H.J., Ojemann, J.G., Buckner, R.L., Kelley, W., Petersen, S.E., 2001. An event-related fMRI study of overt and covert word stem completion. NeuroImage 14, 182–193.

Phelps, E.A., Hyder, F., Blamire, A.M., Shulman, R.G., 1997. FMRI of the prefrontal cortex during overt verbal fluency. Neuroreport 8 (2), 561–565.

Schlaggar, B.L., Brown, T.T., Lugar, H.M., Visscher, K.M., Miezin, F.M., Petersen, S.E., 2002. Functional neuroanatomical differences between adults and school-age children in the processing of single words. Science 296 (5572), 1476–1479.

Turkeltaub, P.E., Eden, G.F., Jones, K.M., Zeffiro, T.A., 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. NeuroImage 16 (3 Pt 1), 765–780.

de Zubicaray, G.I., Wilson, S.J., McMahon, K.L., Muthiah, S., 2001. The semantic interference effect in the picture-word paradigm: an event-related fMRI study employing overt responses. Human Brain Mapp. 14 (4), 218–227.