Does Semantic Context Benefit Speech Understanding through "Top–Down" Processes? Evidence from Time-resolved Sparse fMRI

Matthew H. Davis[1], Michael A. Ford[1,2], Ferath Kherif[3], and Ingrid S. Johnsrude[4]

Abstract

■ When speech is degraded, word report is higher for semantically coherent sentences (e.g., *her new skirt was made of denim*) than for anomalous sentences (e.g., *her good slope was done in carrot*). Such increased intelligibility is often described as resulting from "top–down" processes, reflecting an assumption that higher-level (semantic) neural processes support lower-level (perceptual) mechanisms. We used time-resolved sparse fMRI to test for top–down neural mechanisms, measuring activity while participants heard coherent and anomalous sentences presented in speech envelope/spectrum noise at varying signal-to-noise ratios (SNR). The timing of BOLD responses to more intelligible speech provides evidence of hierarchical organization, with earlier responses in peri-auditory regions of the posterior superior temporal gyrus than in more distant temporal and frontal regions. Despite Sentence content × SNR interactions in the superior temporal gyrus, prefrontal regions respond after auditory/perceptual regions. Although we cannot rule out top–down effects, this pattern is more compatible with a purely feedforward or bottom–up account, in which the results of lower-level perceptual processing are passed to inferior frontal regions. Behavioral and neural evidence that sentence content influences perception of degraded speech does not necessarily imply "top–down" neural processes. ■

INTRODUCTION

Comprehending spoken language requires a complex sequence of perceptual and cognitive processes to convert the acoustic signal into a representation of the intended meaning. Spectrally complex, rapidly changing speech sounds are analyzed by peripheral and cortical auditory perceptual processes before being mapped onto higher-level linguistic representations, which are combined to derive the meaning of the utterance. This hierarchical description accords with anatomical studies of the macaque auditory system (Hackett, 2008; Scott & Johnsrude, 2003; Rauschecker, 1998), with which we share a number of neuroanatomical homologies (Petrides & Pandya, 1999, 2009). At least four cortical processing levels radiate outward from primary auditory cortex (Kaas & Hackett, 2000; Kaas, Hackett, & Tramo, 1999; Hackett, Stepniewska, & Kaas, 1998; Rauschecker, 1998; Pandya, 1995) around the transverse temporal Heschl's gyrus (HG; Rademacher et al., 2001). A cortical hierarchy for speech processing has also been supported by human neuroimaging: Responses in the vicinity of primary auditory cortex are sensitive to the acoustic form of speech (Okada et al., 2010; Davis & Johnsrude, 2003), whereas higher-level semantic and syntactic integration processes are associated with activation of temporal and frontal regions that are more distant from primary auditory cortex (Peelle, Johnsrude, & Davis, 2010; Price, 2010; Saur et al., 2008; Hagoort, 2005; Rodd, Davis, & Johnsrude, 2005; Friederici, Ruschemeyer, Hahne, & Fiebach, 2003; Humphries, Willard, Buchsbaum, & Hickok, 2001; Mazoyer et al., 1993). Here, we explore the functional organization of human speech processing, asking whether information flow is strictly feedforward ("bottom–up") or whether higher-level semantic and syntactic computations interact directly with perceptual regions to change their activity, guiding lower-level perceptual processing "top–down."

One possible indication of top–down influences is that successful comprehension of degraded speech depends on speech content as well as perceptual clarity (Miller & Isard, 1963; Miller, Heise, & Lichten, 1951). For example, word report for speech in noise is more accurate for normal sentences than for syntactically malformed sentences (Miller & Isard, 1963), word lists (Miller et al., 1951), or syntactically correct sentences without coherent meaning (Boothroyd & Nittrouer, 1988; Kalikow, Stevens, & Elliott, 1977; Miller & Isard, 1963). Despite elegant mathematical methods for quantifying contextual benefit (Boothroyd & Nittrouer, 1988), disagreements remain concerning the neurocomputational mechanisms that are responsible. In short, this effect is often colloquially termed

[1]Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK, [2]University of East Anglia, [3]University of Lausanne, [4]Queen's University, Kingston, Canada

"top–down" without in fact requiring the direct interaction between regions supporting semantic processing (the "top") and those supporting perceptual processing (the "down"). There are in fact two distinct classes of account. One proposes that, indeed, contextual benefit arises through top–down processes that allow higher-level content to influence peripheral perceptual mechanisms for word or phoneme identification (McClelland & Elman, 1986; Marslen-Wilson & Tyler, 1980). The other class of account is not top–down. In feedforward accounts, processing is exclusively bottom–up, and context influences integration of perceptual hypotheses in higher-level lexical or semantic regions without need for interaction between regions supporting higher cognitive and lower perceptual processes (e.g., Norris, McQueen, & Cutler, 2000; Massaro, 1989).

On the basis of behavioral evidence that higher-level content influences perception, we predict that critical neural mechanisms will be revealed when demands on contextual integration are high. This occurs both during perception of coherent yet degraded sentences and when the semantic context of a sentence is weak or anomalous (see Figure 1A). Neuroimaging findings from participants listening to degraded coherent sentences have sometimes been interpreted as providing evidence for top–down mechanisms (e.g., Obleser, Wise, Dresner, & Scott, 2007; Zekveld, Heslenfeld, Festen, & Schoonhoven, 2006), but the existing data cannot distinguish between the top–down and bottom–up explanations discussed above. Here, we assess the magnitude and timing of fMRI responses to spoken sentences that vary in semantic content and signal quality to assess these two contrasting neural accounts by testing the predictions illustrated in Figure 1B and C and explained below.

A first test for top–down effects is to assess whether lower-level perceptual or lexical processes are influenced not only by speech clarity (reflecting the acoustics of the signal and thus compatible with lower-level processing) but also by semantic content, which is presumed to depend on the involvement of higher-level, cognitive regions. Previous studies have shown that sentence content modulates neural responses to clear speech (Friederici et al., 2003; Kuperberg et al., 2000) and degraded speech (Obleser et al., 2007) in the superior temporal gyrus (STG). This same lower-level area also shows effects of speech clarity (Zekveld et al., 2006; Davis & Johnsrude, 2003). However, simple effects of sentence content or speech clarity need not imply involvement in compensation for distortion rather than more general processes (e.g., changes in attention). Those previous studies that have simultaneously manipulated speech content and intelligibility (e.g., Obleser et al., 2007) did not localize the critical interaction between these two factors, hence, cannot rule out purely bottom–up accounts. As shown in the center panel of Figure 1B and C, the critical difference between top–down and bottom–up accounts is whether the interaction between sentence content and speech clarity extends to lower-level lexical and perceptual processes. We will assess this in the present study by testing for interactions between sentence content and signal quality during the perception of coherent sentences (e.g., "the *recipe* for the *cake* was easy to *follow*") and anomalous sentences created by substitution of matched content words ("the *idea* for the *soap* was easy to *listen*"). Sentences of this sort were presented without repetition in speech envelope and spectrum noise at varying signal-to-noise ratios (SNRs) including clear speech (Figure 2A–C), ensuring that all parts of the sentence are equally masked (Schroeder, 1968). Top–down mechanisms can thus only improve intelligibility through contextual support for word identification rather than through glimpsing (Cooke, 2006) or other mechanisms (e.g., perceptual learning; Samuel & Kraljic, 2009).

A second test for top–down neural mechanisms concerns the relative timing of higher-level (contextual) and lower-level (perceptual) processes. In top–down accounts, activity for anomalous compared with coherent materials will diverge at an earlier time point in brain regions supporting higher-level processes (the source of top–down feedback) than in regions subserving lower-level processes (the recipients of top–down feedback). According to bottom–up accounts when speech is degraded, increased activity in higher-level integrative processes can only follow, rather than lead, changes in regions supporting lower-level perceptual processing. Thus, the timing of neural interactions between sentence content and intelligibility may provide a second test of top–down accounts. Although previous fMRI studies have assessed the timing of neural responses to manipulations of sentence content (Humphries, Binder, Medler, & Liebenthal, 2007; Dehaene-Lambertz et al., 2006), these studies presented speech against a background of continuous scanner noise preventing comparison of activity during natural, effortless comprehension of connected speech (Peelle, Eason, Schmitter, Schwarzbauer, & Davis, 2010). Here, we combine the quiet listening conditions provided by sparse imaging (Hall et al., 1999) with rapid acquisition of multiple images by using a hybrid sparse-continuous scanning protocol: interleaved silent steady-state (ISSS) imaging (Schwarzbauer, Davis, Rodd, & Johnsrude, 2006; see Figure 2D). In this way, we can measure both the magnitude and timing of BOLD responses to sentences varying in speech content and signal clarity.

METHODS

Participants

Twenty volunteers participated in the sentence report test, and thirteen right-handed volunteers participated in an fMRI study approved by the Cambridgeshire Regional Research Ethics Committee. All were aged between 18 and 45 years (mean age of fMRI volunteers = 26 years, 10 women), native speakers of English, without neurological illness, head injury, or hearing impairment.

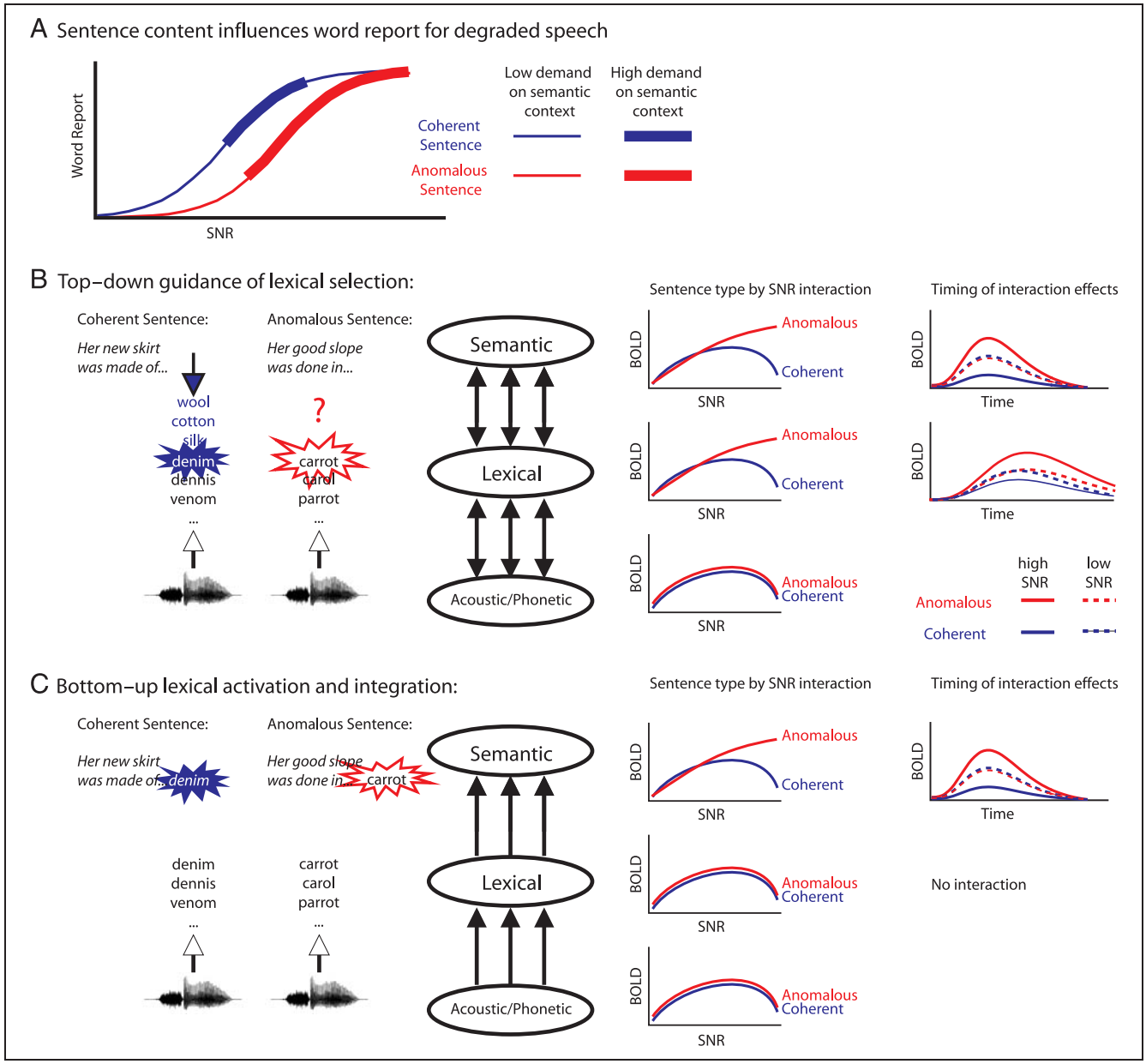


Figure 1. Behavioral and neural predictions for the influence of meaningful semantic context and SNR on perception of degraded speech. (A) Word report scores are higher for coherent than anomalous sentences at intermediate SNRs, reflecting an influence of semantic context on speech perception. Additional demands are placed on meaning-based integration processes when speech is coherent and moderately degraded or when speech is comprehensible and anomalous (thicker lines). However, two different neural explanations are possible depending on whether lower-level processes are modulated by changes to sentence content (top–down accounts (B) or whether lower-level perceptual processing is unaffected by sentence content (bottom–up accounts; C). These accounts can be distinguished by the location and timing of neural interactions between sentence type and signal clarity. Note that other interaction profiles may also occur in brain regions that reflect the outcome of speech comprehension—for instance, systems involved in rehearsal-based STM will show a response profile that is correlated with word report, hence elevated for coherent sentences in intermediate SNR conditions. (B) Neural predictions for top–down accounts of speech comprehension. When speech is coherent and clearly perceived, semantically compatible lexical candidates receive additional activation through top–down mechanisms and are more easily recognized. This leads to differential lexical activation for more clearly recognized coherent sentences compared with degraded speech (for which recognition is challenged by distorted bottom–up input) and compared with clearly recognizable anomalous sentences (for which recognition is challenged by the lack of top–down support). This top–down account, therefore, predicts an interaction between speech intelligibility (i.e., SNR) and sentence type (anomalous vs. coherent) at both semantic and lexical levels (center graphs) that should arise earlier in higher-level semantic regions than in lower-level lexical or perceptual processes (rightmost graphs). (C) Neural predictions of bottom–up accounts in which lexical activation is based only on the perceptual input. Only later, integration processes are differentially engaged for anomalous or degraded sentences. For intelligible coherent sentences, constraints from previous words can be used to guide interpretation of upcoming material and semantic integration is therefore easier. Recognition of degraded sentences can produce additional uncertainty, hence increase processing load in higher-level semantic integration processes that are also challenged by anomalous sentences. However, changes in higher-level semantic integration are independent of lower-level lexical and perceptual processes. These lower levels are therefore only modulated by acoustic distortion and not by sentence type.

Stimulus Preparation and Pilot Behavioral Experiments

One hundred declarative sentences between 6 and 13 words in length were selected from sentences generated for the "low ambiguity condition" of a previous study (Rodd et al., 2005). For each "coherent" sentence, a matched anomalous sentence was created by randomly substituting content words matched for syntactic class, frequency of occurrence and length in syllables (cf. Marslen-Wilson & Tyler, 1980). The anomalous sentences thus have identical phonological, lexical, and syntactic properties but lack coherent meaning. Five pairs of sample sentences are listed in Appendix A.

The resulting 200 sentences (1.2–3.5 sec in duration, speech rate = 238 words/min) were recorded by a male speaker of British English and digitized at a sampling rate of 44.1 KHz. To assess the timing of anomalous content in these sentences, a group of 27 participants were presented with clearly spoken coherent and anomalous sentences over headphones (Sennheiser HD250) and required to press one of two buttons to indicate as quickly and as accurately as possible whether each sentence made sense or not. Responses showed that the majority of the anomalous sentences (74%) were judged as anomalous before sentence offset (average response latency was at 90.8% of the sentence duration, range = 57.7–144.9%). Thus, our sentences are established as coherent or anomalous on the basis of multiple, mutually constraining content words, not just the final word.

These sentences were degraded by adding speech spectrum, signal-correlated noise (SCN; cf. Schroeder, 1968) at a range of SNRs using a custom script and Praat software (www.praat.org). This preserves the duration, amplitude, and average spectral composition of the original sentences at all SNRs, although fine structure becomes progressively more degraded at low SNRs. As intended, the perceptual clarity, hence intelligibility, of the sentences changes dramatically with increasing noise. At extreme SNRs (−10 dB), the stimulus is indistinguishable from pure SCN, a rhythmically modulated noise stimulus that provides an acoustically matched nonspeech baseline (Rodd et al., 2005). At positive SNRs, sentences are highly intelligible, although still somewhat masked. Spectrograms illustrating the acoustic properties of clear speech, SCN, and speech degraded by the addition of SCN are shown in Figure 2A–C.

A pilot study outside the scanner was conducted in a separate group of participants to assess: (a) the intelligibility of

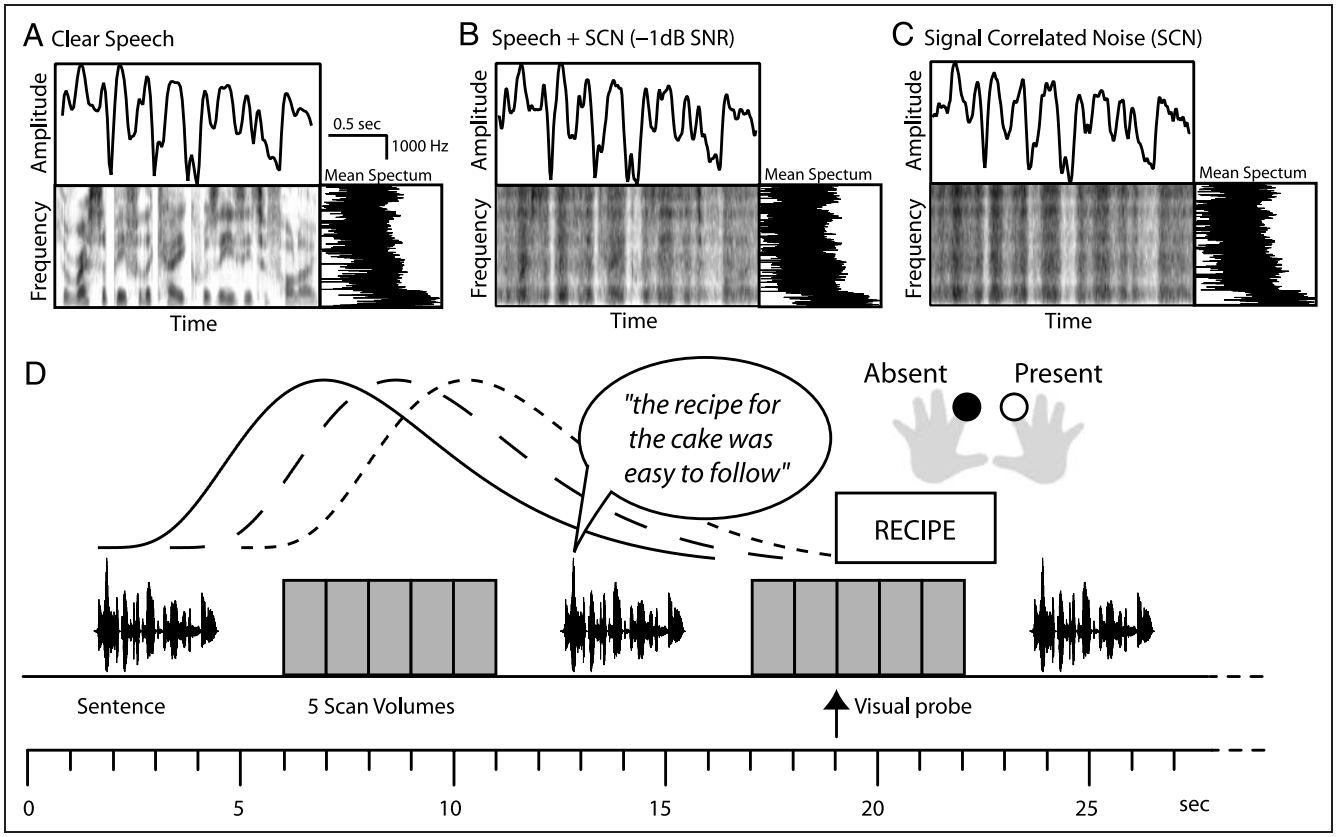


Figure 2. (A–C) Spectrogram (center), amplitude envelope (top), and mean spectrum (right) for (A) a single sentence of clear speech, (B) speech mixed with SCN at a SNR of −1 dB, and (C) SCN matched for amplitude envelope and long-term spectrum. (D) Timeline of stimulus presentation and EPI acquisition using the ISSS sequence (Schwarzbauer et al., 2006). Scans are timed to capture the peak of the hemodynamic response to early, intermediate, and late events during the preceding sentence. Following 50% of sentences a probe word was visually presented for a present/absent decision.

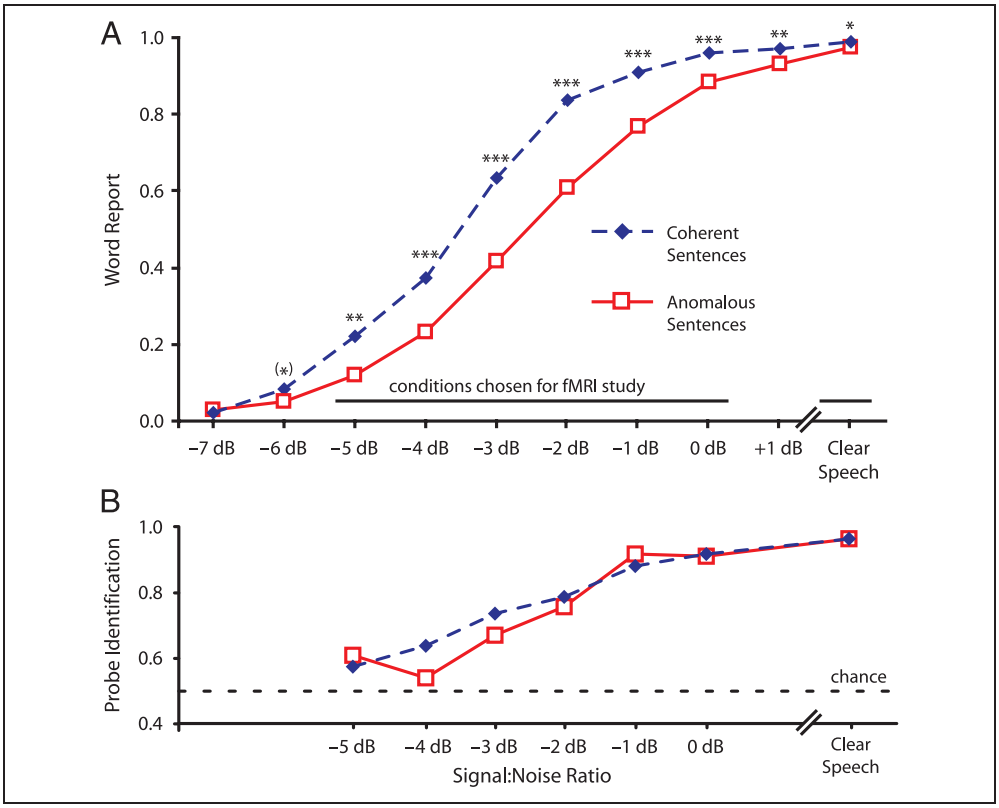


Figure 3. Measured intelligibility of coherent and anomalous sentences at different SNR values. (A) Percentage correct word report data from 20 pilot participants. Increased report scores for coherent sentences: *$p < .1$, *$p < .05$, **$p < .01$, ***$p < .001$. (B) Probe identification decisions for fMRI participants compared with chance performance (50% correct).

sentences degraded through the addition of SCN at varying SNRs and (b) the impact of sentential content on intelligibility. Participants heard single sentences over headphones and were required to type as many words as they could understand immediately after a single presentation of each sentence. Ten coherent and 10 anomalous sentences were presented at each of nine SNRs (in 1 dB steps from −7 to +1 dB) and as clear speech. Sentences were pseudorandomly assigned to a level of degradation for each subject and counterbalanced such that each sentence was heard only once by each participant but was presented at each of the 10 SNRs across subjects. ANOVA on mean word report scores for coherent and anomalous sentences at 10 SNRs (including clear speech, shown in Figure 3A) showed significant main effects of sentence type ($F(1, 19) = 185.05$, $p < .001$) and SNR ($F(9, 171) = 1135.84$, $p < .001$) and the expected interaction between these two factors ($F(9, 171) = 16.659$, $p < .001$) because of greatest contextual facilitation at intermediate SNRs.

fMRI Procedure

Coherent and anomalous spoken sentences were presented to participants at six SNRs between −5 and 0 dB in 1-dB steps; these values were chosen as showing a robust report score benefit for coherent compared with anomalous sentences. In addition, we included trials containing clear speech (both coherent and anomalous sentences), pure SCN, and a silent resting baseline. This made a total of 16 conditions from which 14 conditions formed a factorial crossing of 7 SNR conditions (including clear speech) × 2 Sentence types. Participants were told that they would be listening to sentences, distorted with different amounts of background noise, and instructed to listen attentively to each sentence. To ensure attention, after 50% of sentences, a word was visually presented and participants responded with a button press to indicate if this word was present in (right index figure) or absent from (left index figure) the preceding sentence. Following 50% of silent intervals or SCN presentations (neither of which contained intelligible speech), the words "right" or "left" appeared on the screen and participants were instructed to press the corresponding button on the response box.

We acquired imaging data with a 3-T MR system (Bruker Biospin GmbH, Ettlingen, Germany), using a head-gradient insert and a quadrature birdcage head coil. We used an ISSS sequence (Schwarzbauer et al., 2006) in which a 6-sec silent period was followed by a train of five 1-sec EPI acquisitions. To avoid T1-related signal delay, the ISSS sequence maintains steady-state longitudinal magnetization with a train of silent slice-selective excitation pulses between acquisitions. This sequence provides an optimal compromise between the need to present auditory stimuli during silence and the desire to collect multiple, rapid EPI acquisitions to track the time course of the BOLD response to spoken sentences. A schematic of the experimental procedure is shown in Figure 2D.

We used rapid, near-whole-brain EPI acquisitions with a repetition time of 1 sec during which time 18 × 4 mm thick slices were acquired with a 1-mm interslice gap. Slices were acquired in ascending interleaved order with a 20 × 20 cm field of view, 64 × 64 matrix size, and in-plane spatial resolution of 3.1 × 3.1 mm; acquisition bandwidth was 101 kHz, and echo time was 27.5 msec. Acquisition was transverse-oblique, angled away from the eyes, and covered most of the brain except the top of the superior parietal lobule. In addition to EPI acquisitions, field maps to facilitate geometric undistortion during preprocessing were acquired (Cusack, Brett, & Osswald, 2003), and a high-resolution spoiled gradient echo T1-weighted structural image was acquired for spatial normalization (1-mm isotropic resolution).

Imaging data were acquired in four runs of 10.5 min, each run consisting of 285 acquisitions (57 sets of five EPI volumes, with five initial volumes discarded in each run). Each run included three or four trials of each of the 16 conditions; over four runs, each condition was tested 14 times. Because of problems with stimulus presentation equipment, data from seven scanning runs were discarded (eight participants supplied four scanning runs, three participants supplied three scanning runs, and two participants supplied two scanning runs for analysis). A single sentence was presented during the 6 sec of silence before each set of five EPI acquisitions. Sentence onset was timed so that sentence midpoint coincided with the midpoint of the silent period and EPI acquisitions occurred between 3 and 8 sec after the middle of each sentence, capturing the peak of the evoked hemodynamic response (Hall et al., 1999). After 50% of sentence trials, a probe word was visually presented, 2 sec into the 5-sec acquisition period. Participants were instructed to respond to the written word with a button press to indicate if it was present or absent from the preceding sentence. On 50% of SCN or rest trials, the words "left" or "right" were presented to cue a button press response. Presentation of sentences, probe words, and scans were timed to ensure our fMRI data were optimally sensitive to sentence presentation and insensitive to the probe word task (see Figure 2D).

Coherent and anomalous stimulus items were pseudorandomly assigned to different SNR conditions (including clear speech) for each participant with 14 sentences presented in each condition. This ensures that all sentences occurred equally in all SNR conditions over the group of participants tested. Sentences presented as SCN were chosen equally from coherent and anomalous stimuli. Aside from these unintelligible SCN presentations, no sentence was presented more than once to each participant. Auditory stimuli were presented diotically over high-quality headphones (Resonance Technology, Commander XG System, Northride, CA). To further attenuate scanner noise, participants wore insert earplugs (E.A.R. Supersoft, Aearo Company, www.aearo.com), rated to attenuate by approximately 30 dB. When wearing earplugs and ear defenders, participants reported that the scanner noise was unobtrusive and sentences were presented at a comfortable listening level. Visual stimuli were back-projected using an LCD projector and viewed using an angled mirror inside the scanner head coil. Stimulus presentation and response measurement were controlled using DMDX software (Forster & Forster, 2003) running on a Windows PC.

Preprocessing and Analysis of fMRI Data

Data processing and analysis were accomplished using SPM2 (Wellcome Department of Cognitive Neurology, London, UK). Preprocessing steps included realignment and unwarping to correct for subject motion and interactions between movement and field inhomogeneities (Andersson, Hutton, Ashburner, Turner, & Friston, 2001). The phase and magnitude of the field maps was unwrapped, scaled, and used to remove geometric distortions from the EPI data (Cusack et al., 2003). Individual structural images were coregistered to the functional data and then normalized to the ICBM152 template. These normalization parameters were then applied to preprocessed EPI data, followed by smoothing with a 10-mm FWHM Gaussian filter.

Statistical analysis was conducted using the general linear model in SPM2 (www.fil.ion.ucl.ac.uk/spm/) supplemented by custom Matlab scripts. We used an finite impulse response basis set such that the five scans after sentences in each condition were separately averaged. Head movement parameters and dummy variables coding the scanning sessions were included as covariates of no interest. No high-pass filter or correction for serial autocorrelation was used in estimating the least-mean-squares fit because of the discontinuous nature of ISSS data. The mean activity level in each condition over time can nonetheless be computed because unmodeled temporal correlation has no impact on estimates of the mean effect, only on estimates of scan-to-scan variation (Schwarzbauer et al., 2006). Our analysis procedure focuses on the significance of activation estimates across the group of participants with intersubject variation as a random effect. Hence, within-subject temporal autocorrelation is irrelevant for our statistical analysis.

To identify brain areas in which BOLD signal was correlated with intelligibility, irrespective of sentence content and time, we used word report scores from pilot testing as the predictor of the BOLD response averaged over all five scans after each sentence. We used pilot behavioral data (Figure 3A) rather than data collected in the scanner because (a) word report scores are a more direct measure of intelligibility, (b) ceiling effects in forced-choice identification reduce sensitivity at high SNRs, and (c) good correspondence between word report and in-scanner behavioral data has been shown in previous sparse imaging studies (Davis & Johnsrude, 2003). To assess the effect of intelligibility on activation, we computed the average report score over both sentence types for each

SNR (including clear speech). These values were zero-mean corrected and used as contrast weights for neural responses measured at different SNRs. Thus, we compute in each voxel the slope of the line relating BOLD signal magnitude and speech intelligibility. Group analysis with a one-sample *t* test assesses whether this slope estimate is consistently greater than zero (i.e., a positive correlation between BOLD signal and report score). Intelligibility-responsive regions identified in this analysis were then used as a search volume for analyses to assess the main effect of sentence type and the sentence type by SNR interaction. Because intelligibility-responsive regions were determined on the basis of responses averaged over the two sentence types (main effect of intelligibility), these follow-up analyses contrasting the two sentence types (main effect of sentence type) are orthogonal to the contrast used to define the search volume (Friston, Rotshtein, Geng, Sterzer, & Henson, 2006). Hence, the search volume and follow-up contrasts are independent under the null hypothesis (as recommended by Kriegeskorte, Simmons, Bellgowan, & Baker, 2009).

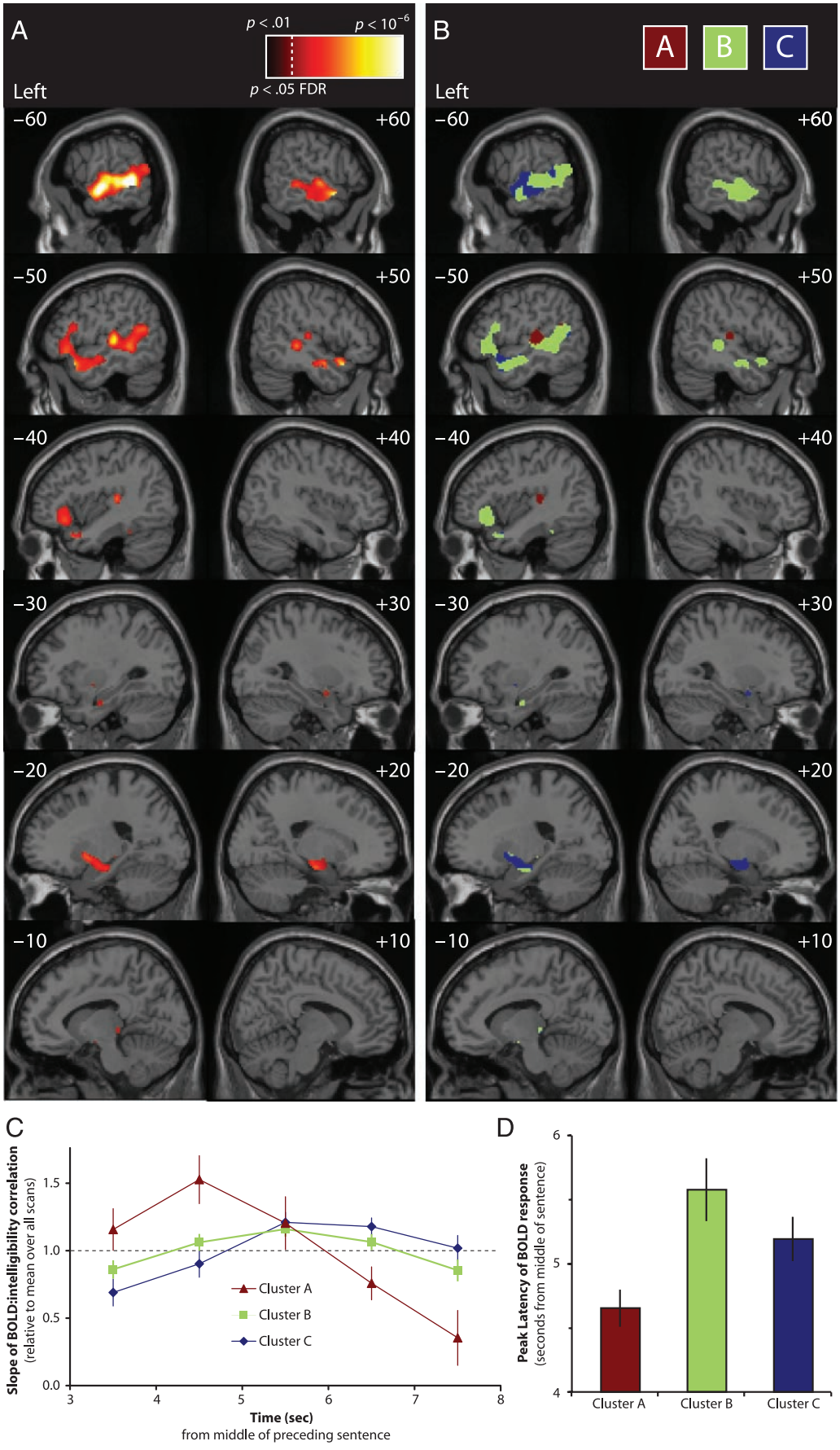
The main effect of sentence type was assessed using a one-sample *t* test with both positive and negative contrasts testing for additional activity in response to coherent compared with anomalous sentences (and vice versa), averaged over all SNR conditions. To assess sentence type by SNR interactions, we computed seven contrast images (coherent vs. anomalous) at each SNR (six degraded and one clear speech condition) in each subject averaged over time. These images were entered into a one-way repeated measures ANOVA in SPM2 and the "effects of interest" *F* test used to test the sentence type by SNR interaction (Penny & Henson, 2006). This method tests for a differential response to the sentence type manipulation at different SNRs, including both the interaction shown in behavioral report scores and that predicted in Figure 1B-C. For all these analyses, we report results after correction for multiple comparisons using the false discovery rate procedure (Genovese, Lazar, & Nichols, 2002). We applied a voxel threshold of $p < .05$ corrected for the whole brain or search volumes derived from orthogonal contrasts.

Clustering Analysis

Because a number of different response profiles can give rise to a sentence type by SNR interaction, we used a data-driven approach to characterize the activation profiles over conditions within brain regions showing significant sentence type by SNR interactions. We used a *k*-means clustering algorithm implemented in Matlab v6.5 (MathWorks, Natick, MA) to identify a number of mutually exclusive subsets of voxels that show consistently different interaction profiles (see Simon et al., 2004). This method starts by randomly setting seven values for each of a number (*k*) of cluster centroids. These sets of seven values reflect the average effect of sentence type at each of the seven SNR values in the interaction analysis for each of *k* clusters. Voxels are assigned to the most similar centroid, and once assigned, the centroids are updated to be the mean of the assigned voxels. These phases are iterated until no voxels are assigned to different clusters in consecutive runs. Because this procedure is sensitive to starting conditions, it was repeated 50 times using different random seeds and the solution that maximized the between-cluster (explained) variance divided by the within-cluster (unexplained) variance was selected.

We used two iterative procedures: first, to determine the number of clusters (*k*) that best explains the observed data without redundancy and, second, to assess the statistical reliability of response differences between clusters. First, we ran cluster analyses with increasing values of *k* at each step. We tested for statistically significant differences among the response profiles of different clusters by extracting the mean response profile for each cluster and subject using MarsBar (Brett, Anton, Valabregue, & Jean-Baptiste, 2002) and comparing these profiles using repeated measures ANOVAs. The absence of a significant Cluster × Condition interaction provides evidence that two redundant clusters have been discovered by the clustering routine and the previous value of *k* is assumed to be the best description of the response profiles established by *k*-means clustering. If all pairwise comparisons between different clusters are statistically reliable, the number of clusters (*k*) is increased by one and the clustering procedure is repeated. This allows us to determine the maximum number of different response profiles within the observed activation map. However, assessment of cluster differences might be biased by the use of the same data in generating the clusters and in subsequent statistical analysis. We therefore use a second iterative procedure (a leave-one-out analysis) to ensure that statistical analysis of cluster response profiles could be conducted on data that were independent of the original clustering procedure.

Figure 4. Spatial and temporal profile of brain responses to speech intelligibility for coherent and anomalous sentences combined. (A) Brain regions in which the magnitude of the BOLD signal at each SNR (averaged over scans and over the two sentence types) is correlated with intelligibility (word report). Statistical maps displayed at $p < .05$ FDR corrected on sagittal slices of the Montreal Neurological Institute (MNI) canonical brain. (B) Intelligibility-responsive regions divided into clusters with three distinct temporal profiles discovered by *k*-means analysis. (C) Time course of correlation between BOLD signal and intelligibility between 3 and 8 sec after the middle of each sentence. Results show the ratio of slope estimates for each time point relative to the mean slope averaged over scans (dotted line) for each of three clusters depicted in B. (D) Estimated peak latency for BOLD responses from a leave-one-out clustering procedure that provides independent estimates of the temporal profile suitable for group analysis. Error bars show the standard error after between-subjects variance has been removed, suitable for repeated measures comparison (cf. Loftus & Masson, 1994).


A
p < .01
p < 10^-6
p < .05 FDR
Left
−60
+60
−50
+50
−40
+40
−30
+30
−20
+20
−10
+10
B
A
B
C
Left
C
Slope of BOLD:intelligibility correlation
(relative to mean over all scans)
1.5
1.0
0.5
0.0
3
4
5
6
7
8
Time (sec)
from middle of preceding sentence
Cluster A
Cluster B
Cluster C
D
Peak Latency of BOLD response
(seconds from middle of sentence)
6
5
4
Cluster A
Cluster B
Cluster C

Clustering is performed on mean data generated from 12 participants, resulting in a set of clusters with good spatial correspondence to the clusters determined from the mean of all participants. Critically, however, cluster locations determined in this way are independent of the data from the thirteenth, left-out participant. Hence, data can be extracted from the remaining (left-out) participant from clusters that correspond with the clusters generated from analysis of the entire group but generated without bias or circularity. This "leave-one-out" procedure is repeated until we have independent cluster data from all 13 participants. These data can then be entered into group statistical analyses to assess the significance of Cluster × Condition interactions indicative of statistically distinct response profiles for the group of participants tested (Henson, 2006).

fMRI Timing Analysis

To test for differences in the timing of neural responses in intelligibility responsive brain regions, we again used *k*-means clustering to separate regions sensitive to intelligibility or Sentence type × SNR interactions with different temporal profiles. To ensure clusters were only distinguished by the timing of the BOLD signal, response magnitude at each time point was normalized by divided by the average response magnitude for that contrast in all five scans. As before, the number of distinct clusters and statistical significance in group analysis was decided using the group clustering procedure, followed by a leave-one-out clustering procedure. However, given the nonsphericity that is expected of time-course data, we used multivariate ANOVAs to confirm the presence of statistically significant differences among the temporal profiles of different clusters. In a complementary analysis, we measured the scan at which the peak of the BOLD response was observed in each cluster (Bellgowan, Saad, & Bandettini, 2003). These latency (i.e., time to peak) values were compared among clusters using repeated measures ANOVAs.

RESULTS

Probe Word Detection

Because of the small number of observations in each condition, signal detection analysis was inappropriately skewed by responses that were 0 or 100% correct in specific conditions. We, therefore, conducted analysis on the proportion of correct responses (hits and correct rejections). This revealed a significant main effect of SNR ($F(6, 72) = 33.551$, $p < .001$), although the effect of Sentence Type ($F(1, 12) = 1.192$, $p > .05$) and the Sentence Type × SNR interaction ($F(6, 72) = 1.109$, $p > .05$) did not reach significance (Figure 3B). These data are much less fine-grained than the word report data from the pilot study (see Figure 3A), yet results are largely consistent, lending further support to our decision to assess intelligibility effects in the imaging data using word report scores from the pilot study.

Effects of SNR, Type of Sentence, and Their Interaction

Analysis of the magnitude of activity, averaged over all five scans following each sentence, shows a network of frontal and temporal lobe regions in which BOLD signal correlated with the intelligibility of speech as quantified by mean word report (from the pilot study) across sentence types (see Figure 4A and Table 1). These included extensive, bilateral, temporal lobe regions extending from posterior middle temporal gyrus (MTG) and STG to the temporal pole. In the left hemisphere, this activation extends into the left inferior frontal gyrus (LIFG), including peaks in partes opercularis, triangularis, and orbitalis. Correlations between intelligibility and BOLD signal were also observed in bilateral medial-temporal regions, the left putamen, and left inferior temporal and fusiform gyri. This extensive fronto-temporal and subcortical system of speech-responsive regions provides a search volume within which to assess the magnitude and timing of neural responses reflecting type of sentence (coherent vs. anomalous). However, the contrast between coherent and anomalous sentences (averaged over intelligibility and SNR levels) failed to reach FDR-corrected significance (in either direction) within intelligibility-responsive regions or elsewhere in the brain. Several regions exhibited robust interactions between sentence type and SNR, and this may have eliminated a main effect of sentence type.

We assessed the Sentence type × SNR interaction using a repeated measures ANOVA in a search volume of intelligibility-responsive regions as before. As shown in Figure 5A and Table 2, this analysis revealed a number of cortical and subcortical regions in which the differential response to coherent and anomalous sentences depended on SNR, including several regions of the STG and MTG. A large cluster crossing all three anatomical divisions of the LIFG (partes opercularis, triangularis, and orbitalis) was also observed. In addition to these cortical regions, bilateral clusters were evident in medial-temporal regions and in the left lentiform nucleus. We conducted a whole-brain analysis to determine whether significant interactions were observed outside the brain regions that respond to speech intelligibility, but this did not reveal any effects at a corrected level of significance. Because our effects of interest were predicted to occur within regions contributing to speech comprehension (see Figure 1B and C), we focused on findings in this intelligibility-responsive search volume in subsequent statistical analyses.

Several different response profiles give rise to an interaction between SNR and sentence type. We used a *k*-means clustering procedure to identify subregions showing statistically distinct response profiles as confirmed by three-way Brain region × Sentence type × SNR interactions. We obtained evidence for three regionally specific response

Table 1. Intelligibility Correlation for Coherent and Anomalous Prose Combined

Location	*Voxels (n)*	*p (FDR)*	*Z*	*MNI Coordinates*		
				x	*y*	*z*
L lateral temporal/frontal lobe	3658					
Posterior MTG[B]		.001	5.49	−60	−42	0
Anterior MTG[B]		.001	5.15	−58	−10	−8
Mid STG[C]		.002	4.55	−62	−18	4
Temporal pole (middle)[C]		.002	4.41	−46	18	−26
Posterior MTG[B]		.004	4.12	−54	−56	14
IFG (orbitalis)[B]		.004	4.06	−46	26	−6
IFG (triangularis/opercularis)[B]		.007	3.74	−48	18	16
Anterior MTG[B]		.007	3.74	−50	0	−24
R lateral temporal lobe	1294					
Temporal pole (superior)[B]		.001	5.09	58	10	−14
Anterior MTG[B]		.002	4.38	54	−2	−22
Temporal pole (superior)[B]		.002	4.33	50	16	−20
Posterior MTG[B]		.003	4.18	54	−34	−2
Mid STG/MTG[B]		.003	4.16	66	−18	−8
Mid STG[B]		.004	4.06	60	−8	0
Heschl's gyrus[a]		.007	3.79	46	−22	10
Post STG[B]		.007	3.73	66	−32	−4
Mid MTG[B]		.009	3.63	58	−22	−16
L medial temporal/lentiform nucleus	439					
Putamen[C]		.004	4.10	−24	4	−6
Hippocampus (head)[B]		.009	3.63	−20	−10	−14
R medial temporal	286					
Hippocampus (head)[C]		.006	3.85	20	−14	−16
Amygdala		.010	3.56	26	0	−12
L inferior colliculus[B]	19	.014	3.36	−8	−30	−2
L inferior temporal lobe	26					
Fusiform gyrus[b]		.015	3.35	−38	−38	−22
Inferior Temporal Gyrus[C]		.017	3.30	−44	−48	−20
L anterior fusiform[B]	13	.017	3.27	−32	−10	−24
L hippocampus (body)[B]	6	.021	3.18	−22	−26	−6

Thresholded at $p < .001$ uncorrected (equivalent to $p < .05$ whole-brain FDR corrected). The table shows MNI coordinates and anatomical location of all peak voxels separated by more than 8 mm in clusters larger than five voxels. Superscripts A, B, and C indicate which of three temporal profiles illustrated in Figure 4B–D is shown by the voxels at each peak. L = left; R = right; FDR = false discovery rate.

profiles, shown in Figure 5B–E. Profile A (blue in Figure 5B) is observed in inferior frontal regions and in the left anterior STG. In these regions, the two sentence types yield similar activity at low SNRs. However, at high SNRs (when both sentence types are highly intelligible), activity for anomalous sentences continues to increase whereas it declines for coherent sentences (Figure 5C). The second response profile (B, red in Figure 5B and D) is reminiscent of the behavioral advantage seen for reporting words in coherent compared with anomalous sentences and is seen in

medial-temporal regions (hippocampus, parahippocampus, and amygdala) and in left BG and inferior colliculus. BOLD responses for coherent sentences are maximal even in intermediate SNR conditions (−2 and −1 dB SNR), whereas for anomalous sentences, BOLD responses gradually increase as SNR goes up. Statistical comparisons on the basis of data from leave-one-out clustering confirmed that this response profile differs from Profile A ($F(6, 72) = 7.125$, $p < .001$ for the three-way interaction among clusters, sentence type, and SNR). The third response profile was exhibited in bilateral regions of the posterior STG and MTG and in right anterior STG (C, green in Figure 5B and E). This profile resembles a combination of the other two profiles, with both an increased response to coherent sentences at

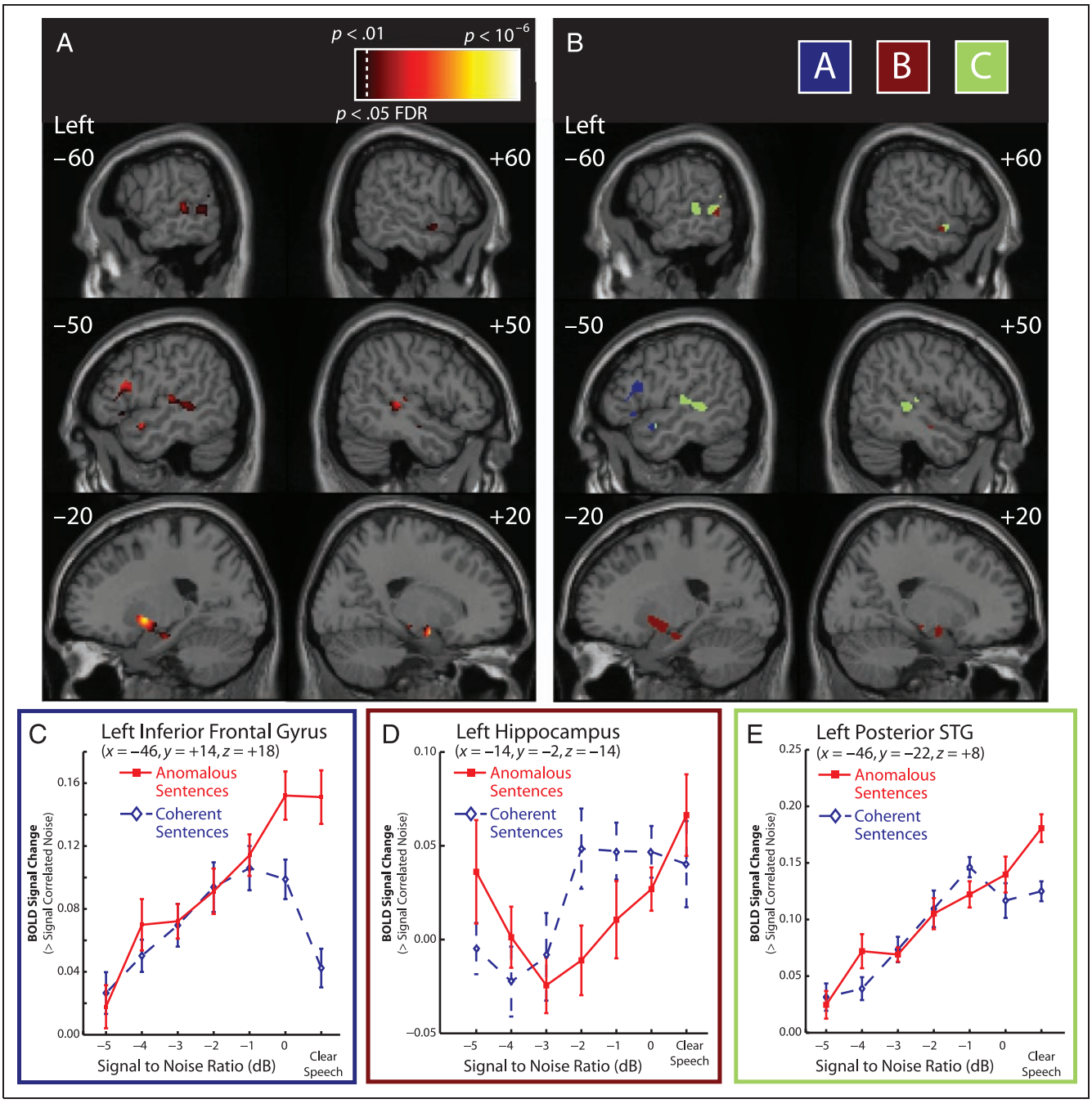


Figure 5. Brain regions that show an interaction between sentence condition (coherent/anomalous) and SNR (−5 to 0 dB, clear speech). (A) Interaction map displayed at a threshold of $p < .05$ FDR corrected within a search volume of intelligibility-responsive regions (see Figure 3A). (B) The interaction map divided into three clusters that show differential response profiles determined by *k*-means analysis. (C) BOLD responses compared with SCN baseline for all sentence conditions/SNRs for a voxel in the inferior frontal gyrus cluster (Cluster A, blue in Figure 4B). Error bars show standard error without between-subjects variance (Loftus & Masson, 1994). (D) BOLD response of a voxel in the left hippocampus (Cluster B, red in Figure 4B). (E) BOLD response of left posterior MTG (Cluster C, green in Figure 4B).

Table 2. Sentence Type × SNR Interaction

Location	*Voxels (n)*	*p (FDR)*	*Z*	*MNI Coordinates*		
				x	*y*	*z*
L medial temporal/basal ganglia[B]						
Lentiform nucleus	269	.006	4.35	−18	2	−6
Hippocampus		.009	3.58	−14	−2	−14
Hippocampus/Parahippocampus		.013	3.28	−16	−16	−18
R medial temporal[B]						
Amygdala/Hippocampus	111	.006	4.28	24	0	−14
R parahippocampal gyrus[B]	25	.006	4.14	14	−22	−12
*L anterior STG[A]	37	.006	4.1	−48	4	−16
L Inferior colliculus[B]	19	.007	3.88	−8	−30	−4
*L inferior frontal gyrus[A]						
IFG (opercularis)	135	.009	3.68	−46	14	18
IFG (triangularis)		.031	2.76	−46	28	2
IFG (orbitalis)		.046	2.44	−50	22	−6
L superior/middle temporal[C]						
L posterior STG	319	.009	3.54	−46	−22	8
L posterior MTG		.016	3.19	−56	−34	4
L posterior MTG		.02	3.05	−64	−48	4
R STG[C]	64	.012	3.37	52	−28	2
R STG[B]	31	.024	2.94	54	−10	−14
R STG[C]	16	.031	2.77	62	0	−10
L STG[C]	6	.044	2.49	−66	−20	4

Thresholded at $p < .05$ FDR corrected within region that respond to intelligibility (see Figure 5A). The table shows MNI coordinates and anatomical location of all peak voxels separated by more than 8 mm in clusters larger than five voxels. Superscripts A, B, and C indicate which of three Prose type × SNR interaction profiles illustrated in Figure 5B–E is shown by the majority of voxels in each cluster. The temporal profile of the clusters marked * are plotted in Figure 6.

intermediate SNRs and to anomalous clear sentences. Statistical comparison confirmed that this profile significantly differed from the response of Profile A (Clusters × Sentence type × SNR interaction, $F(6, 72) = 5.678, p < .001$) and from Profile B ($F(6, 72) = 3.019, p < .05$).

The Timing of Neural Responses to Intelligibility and Sentence Type

We applied the *k*-means procedure to segregate clusters that exhibit differential timing of BOLD responses to intelligibility (i.e., showing an interaction between intelligibility and time). Incremental clustering of the mean response profile suggested three distinct temporal profiles (Figure 4B–D and Table 1), although timing differences between clusters were only confirmed by significant Cluster × Time interactions in leave-one-out analysis for two of the three clusters. The earliest response correlated with intelligibility is observed in bilateral regions of posterior Heschl's gyrus and planum temporale (red in Figure 4B and C). This response, close to primary auditory regions, peaks less than 5 sec after the middle of the sentence (Figure 4D). Consistent with hierarchical organization, intelligibility-sensitive anterior and posterior portions of the MTG and IFG (Cluster B) and medial-temporal regions (Cluster C) both show a later response that peaks over 5 sec after the middle of the preceding sentence. Pairwise comparison of Clusters A and B showed a significant multivariate Cluster × Time interaction $F(4, 9) = 3.914, p < .05$ and a significant difference in the time of the maximum response ($t(12) = 2.52, p < .05$), these differences were also reliable for comparison of Clusters A and C ($F(4, 9) = 3.860, p < .05, t(12) = 2.50, p < .05$). However, comparison of Clusters B and C in leave-one-out analysis shows no significant Cluster × Time interaction, ($F(4, 9) = 1.239, ns$) nor any significant difference in peak

latency ($t(12) = 0.959$, *ns*), suggesting that the timing differences apparent in Figure 4C may be artifacts of the clustering procedure. Although we must necessarily be cautious in drawing conclusions from differences in timing between regions, we note that equivalent differences in temporal profile are absent when we assess a low-level baseline contrast (SCN versus rest), despite all three clusters showing a reliable response (the average response over time and voxels is significant at $p < .001$ in all three clusters). Hence, differences in the temporal profile of the three clusters in response to intelligible speech seem unlikely to be explained by hemodynamic variables and rather by changes in the timing of neural activity over the course of sentences lasting approximately 3 sec.

Fronto-temporal regions that show an interaction between Sentence type × SNR (Figure 5A) show some overlap with regions that show different temporal profiles in responding to speech intelligibility (Figure 4B). To assess differential timing of the Sentence type × SNR interaction, we condensed the three interaction profiles in Figure 5 into a single contrast that tests for additional activity evoked by anomalous versus coherent sentences at high SNRs (clear speech and 0 dB SNR, cluster A in Figure 4B), and the reverse difference for coherent versus anomalous sentences at moderate SNRs (−1 and −2 dB, cluster B[1]). This contrast captures most of the critical interactions between SNR and sentence type shown in Figure 5. *k*-means analysis of the time-course of this Sentence type × SNR interaction was then applied to each of the interaction clusters shown in Figure 5B. However, leave-one-out analysis failed to confirm significant differences in timing, as shown either by Cluster × Time interactions (Cluster A: $F(4, 9) = 1.042$, *ns*, Cluster B: $F < 1$, Cluster C, $F(4, 9) = 2.062$, *ns*) or by differences in the timing of peak responses (A: $t(12) = 1.379$, *ns*, B: $t(12) = 1.032$, *ns*, C: $t(12) = 1.620$, *ns*). Given the importance of the relative timing of frontal and temporal lobe activity in distinguishing top–down and bottom–up accounts of speech comprehension, we averaged the response of the inferior frontal regions that show the predicted interaction between sentence type and speech clarity, and did the same for anterior temporal regions (both parts of cluster A in Figure 5B, two regions marked * in Table 2). The interaction in these two regions averaged over time (Figure 6A) is essentially equivalent, whereas the temporal evolution shown in Figure 6B-D suggests some differentiation because the peak of the interaction in the anterior STG occurs earlier than in the IFG (Figure 6B, $t(12) = 2.347$, $p < .05$), reflecting an interaction that is present at all time-bins in the anterior STG (Figure 6C) but builds up over time in the IFG (Figure 6D). Whilst we hesitate to draw strong conclusions from small differences in the timing of the hemodynamic response that were insufficiently consistent to appear in the leave-one-out analysis, we note that this is the reverse of the temporal profile predicted for the top–down account in Figure 1B. To the extent that these timing differences are reliable they are opposite to the predictions of a top–down account.

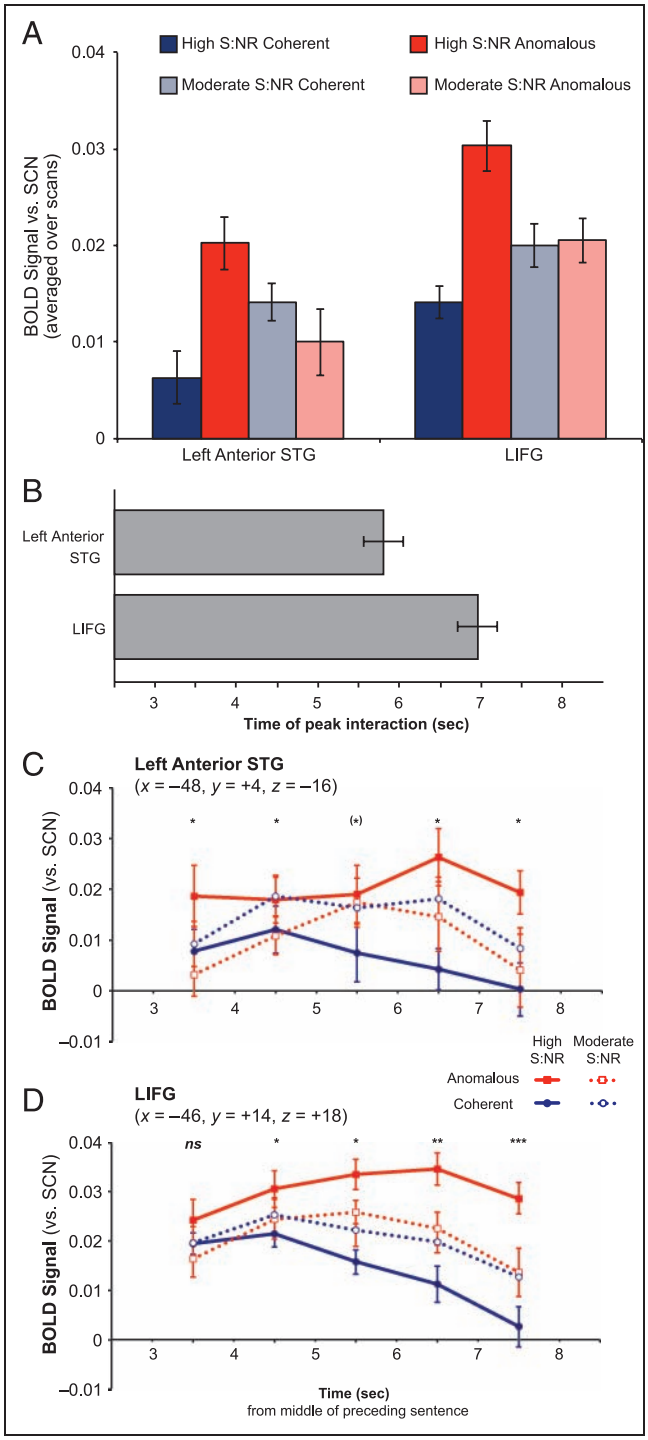


Figure 6. Interaction profiles indicative of additional load on semantic processing during degraded and clear speech comprehension. (A) The interaction between sentence type and speech clarity averaged over scans in left anterior STG (cluster peak: −48, +4, −16) and LIFG (−46, +14, +18) marked * in Table 2. (B) Time of peak interaction in seconds, measured from the midpoint of the preceding sentence in these two clusters. Time course of interaction in (C) left anterior STG and (D) LIFG.

DISCUSSION

Semantic content is a ubiquitous and powerful aid to speech comprehension in noisy environments. However, the neural mechanisms responsible remain underspecified. Here, we test the proposal that top–down neural processes, driven by coherent sentence-level meaning, contribute to the perception of speech under challenging listening situations. Specifically, we examine (a) whether Sentence type × SNR interactions because of increased difficulty of contextual integration are observed in low-level, perceptual areas as well as in higher-level semantic areas, and (b) whether the BOLD signal in areas supporting higher-level, linguistic processes is modulated by sentence content before areas supporting lower-level perceptual processes. Before discussing these two findings, we first discuss results concerning the location and timing of BOLD responses correlated with speech intelligibility. These findings provide methodological validation for our use of the location and timing of interactions between sentence type and speech clarity as evidence for top–down neural processes in the comprehension of degraded speech.

Timing of Responses to Intelligible Speech

Consistent with previous studies (Okada et al., 2010; Awad, Warren, Scott, Turkheimer, & Wise, 2007; Obleser et al., 2007; Davis & Johnsrude, 2003; Scott, Blank, Rosen, & Wise, 2000), activity in a fronto-temporal network (including left frontal cortex, bilateral temporal cortex, hippocampus, and subcortical structures) correlated with the intelligibility of spoken sentences (Figure 4A). Going beyond previous studies, two different clusters of activity can be identified on the basis of their different temporal profiles (Figure 4B–D). The cluster with the shortest peak latency, included regions posterior to Heschl's gyrus in the STG bilaterally. The rest of the intelligibility-responsive regions showed a significantly longer latency response, including anterior STG/MTG and posterior MTG regions, the left lentiform nucleus and hippocampal formation, as well as the LIFG. Previous work has shown longer temporal receptive fields for speech responsive regions further from auditory cortex (Lerner, Honey, Silbert, & Hasson, 2011), differences in the phase lag of the BOLD signal (Dehaene-Lambertz et al., 2006), and directional influences using dynamic causal modeling (Leff et al., 2008), all of which are consistent with the earlier responses to speech intelligibility we observed in the posterior STG. These temporal profiles are also consistent with hierarchical models of the auditory system (Price, 2010; Hackett, 2008; Davis & Johnsrude, 2007; Kaas et al., 1999), with lower-level perceptual processes occurring in or near Heschl's gyrus (HG) and higher-level, linguistic processes supported by more distant regions along the lateral temporal STG, STS, and MTG, anterior and posterior to HG, and LIFG.

It could be argued that BOLD fMRI is ill-suited to detecting what may be subtle differences in the timing of neural activity during sentence comprehension. Interregional variation in the timing of the hemodynamic response (e.g., because of differences in vasculature) will confound attempts to compare the timing of responses to the same contrast in different regions. However, between-condition comparisons in the same region are interpretable (Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000; Menon & Kim, 1999). With appropriate comparison conditions, then, it is possible to make some, tentative spatio-temporal inferences concerning the neural systems involved in sentence comprehension (cf. Sabatinelli, Lang, Bradley, Costa, & Keil, 2009, for faces). One reliable finding from our temporal clustering analysis was that neural responses that correlate with sentence intelligibility peak earlier in posterior regions of the STG than in more inferior and anterior temporal regions or in more distant frontal and medial-temporal regions. This finding seems unlikely to be explained on purely hemodynamic grounds because we see differences in the timing of the response to intelligibility in spatially contiguous regions (e.g., left posterior STG/MTG) that share the same blood supply and vasculature. Furthermore, the clusters identified in temporal clustering of intelligibility responses do not show similar hemodynamic timing differences in their response to nonspeech stimuli (such as for SCN vs. silence).

Our observation of earlier responses to intelligible speech in regions close to auditory cortex is also consistent with the results of magnetoencephalography studies of single-word perception (see Pulvermuller, Shtyrov, & Hauk, 2009; Marinkovic et al., 2003) and of EEG studies that compare the timing of mismatch responses for unexpected phonetic and semantic elements (Uusvuori, Parviainen, Inkinen, & Salmelin, 2008; van den Brink, Brown, & Hagoort, 2001; Connolly & Phillips, 1994). These electrophysiological measures have many advantages in determining the timing of neural responses on the msec scale. However, combined analyses of responses to speech content and auditory form in source space are required to infer the direction of information flow in neural systems (see Gow, Segawa, Ahlfors, & Lin, 2008, for illustrative data from phoneme and word perception). In the absence of similar data for effects of sentence type on responses to degraded speech, our results from time-resolved fMRI provide a novel source of evidence concerning top–down and bottom–up neural mechanisms responsible for behavioral effects of semantically constraining context on the comprehension of degraded speech.

Interactions between Sentence Content and Speech Clarity

The data presented here highlight neural systems that contribute to the perception and comprehension of spoken sentences in suboptimal listening conditions similar to those found in everyday life. A network of frontal and temporal lobe regions (Saur et al., 2008) respond to these challenging listening situations with computations that appear

to combine information in the speech signal with responses that differ as a function of sentence type. A novel contribution of the leave-one-out clustering analyses that we apply in the present study is the demonstration that antero-lateral, postero-lateral, and medial regions of the temporal lobe display functionally distinct interactions between sentence type and speech clarity.

The response profile most reflective of effortful comprehension of degraded speech is observed in anterior temporal and inferior frontal regions. These show a Sentence type × Clarity interaction, in which activity is high for degraded speech and clear anomalous sentences, but low for clear coherent sentences (Figure 5C), as predicted (Figure 1). For both coherent and anomalous sentences, this profile is consistent with effortful semantic integration. For coherent sentences, effortful processing would be most manifest at intermediate intelligibility levels, whereas for anomalous sentences, as speech clarity increases, the load placed on regions attempting to derive a coherent sentence-level meaning would only increase. Thus, our work supports other studies in which perceptual challenges (Zekveld et al., 2006; Davis & Johnsrude, 2003) and semantic disruption (e.g., Friederici et al., 2003; Kuperberg et al., 2000) lead to additional activity in both inferior frontal and superior temporal regions. These interactions provide initial evidence for neural processes that are influenced both by speech clarity and linguistic content.

The profile in the posterior STG/MTG is similar to that observed in anterior STG and LIFG, except that activity for minimally degraded and clear coherent sentences appears to plateau, instead of dropping (Figure 5E). This suggests that posterior STG/MTG is rather less weighted toward semantic integration than are anterior temporal and frontal regions, but the similarity in profiles suggests that these areas may work as an integrated network. This region of the posterior STG, adjacent to HG, probably supports lower-level perceptual processes, and so this provides evidence in support of our first test for "top–down" effects; namely that relatively low-level perceptual areas show activity modulated by both sentence type and speech clarity. In contrast, the response profile of the hippocampus and putamen mirrors behavioral report scores (Figure 3A), suggesting that these regions may operate on the product of sentence comprehension, rather than contribute to comprehension themselves.

Another piece of evidence for top–down influences would be if compensatory activity at higher levels of the processing hierarchy led rather than lagged activity in lower-level areas supporting perceptual processes. However, none of the three clusters that show interactions between sentence type and speech clarity can be divided using leave-one-out-clustering in such a way as to show significant differential timing of neural responses. Indeed, in assessing the timing of responses most indicative of contextually driven compensation for distortion (anterior STG and LIFG regions in Cluster A of Figure 5B), it appears that the temporal lobe response precedes the inferior frontal response (Figure 6). This pattern is opposite to the predictions of top–down accounts. Thus, our data are more consistent with a hierarchically organized, "outward", bottom–up flow of information, rather than with a highly top–down flow of information.

Implications for the Neural Basis of Word and Sentence Comprehension

The idea that semantic content alters intelligibility via top–down mechanisms is not fully supported by our data. Although sensitivity to sentence content was observed in lower-level intelligibility-sensitive areas, the timing of such content sensitivity is earlier in superior temporal than in inferior frontal regions, opposite to the prediction of top–down neural processes (Figure 1). On the basis of this finding, sentence content may not "enhance" lower-level perceptual processing "top–down", but rather the comprehension system may delay making bottom–up commitments to particular perceptual interpretations until lower-level and higher-level representations can be combined. Enhanced comprehension of degraded speech on the basis of sentence content may arise at a later stage of processing that optimally combines multiple sources of uncertainty in the speech signal. If probabilistic representations of phonological, lexical, semantic and syntactic content in the speech signal are to be simultaneously integrated then these computations necessarily involve integration of information over extended stretches of the speech signal—drawing on an extended temporal hierarchy of speech representations that may recruit more anterior regions of the STG and MTG (Lerner et al., 2011) as well as frontal processes involved in higher-level integration (Hagoort, 2005; Rodd et al., 2005; Friederici et al., 2003; Humphries et al., 2001).

We therefore propose that the early response to listening challenges in the anterior STG may reflect storage of partly analyzed speech information, providing a "neural workspace" for later lexical selection and semantic integration. This workspace can be triggered by the presence of perceptual and/or semantic uncertainty in the speech signal before higher-level information about how the bottom–up signal can best be interpreted. A computational illustration of this workspace comes in the internal memory (recurrent hidden units) of the distributed cohort model (DCM; Gaskell & Marslen-Wilson, 1997). The temporal duration of unanalyzed speech input that contributes to current interpretations in DCM depends on the quality of the bottom–up input. Whereas this distributed model of speech perception has not been extended to simulate sentence-level computations, we note that similar mechanisms for incremental processing are to be found in recurrent network models of sentence comprehension (e.g., St. John & McClelland, 1990). Within such a recurrent network account, initial uncertainty concerning the meaning or syntactic function of incoming words (e.g., in an anomalous or ungrammatical sentence) would also lead

to increased load on initial storage of unanalyzed input for later integration.

In combination, then, we propose that perception of both degraded speech and sentences lacking strong semantic context place an additional demand on internal representations of unanalyzed speech. Echoic representations of speech have previously been linked with anterior regions of the STG (Davis & Johnsrude, 2007; Buchsbaum, Olsen, Koch, & Berman, 2005) that show Sentence type × Clarity interactions in the present study. Furthermore, echoic storage of unanalyzed material will, as a downstream consequence, also increase the load on later processes of lexical/semantic selection which have been associated with the LIFG (Bozic, Tyler, Ives, Randall, & Marslen-Wilson, 2010; Righi, Blumstein, Mertus, & Worden, 2010; Rodd et al., 2005). These two computational processes may adequately explain both the timing and location of Sentence type × Speech clarity interactions observed in the present study, without necessary recourse to top–down mechanisms.

Evidence for Top–Down Mechanisms in Other Aspects of Speech Perception

Our data therefore do not yet provide sufficient evidence to support top–down neural mechanisms that use sentence content to support comprehension of degraded speech. However, we do not intend that these results be taken as contradicting the proposal that top–down mechanisms are involved in other aspects of speech perception. We and others and others have provided evidence for top–down mechanisms that after recognition guide perceptual, lexical, and semantic retuning (McClelland, Mirman, & Holt, 2006; Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Norris, McQueen, & Cutler, 2003). These postrecognition perceptual learning processes lead to significantly improved comprehension of similar speech that is presented subsequently and are enhanced in the presence of higher-level lexical feedback (see Samuel & Kraljic, 2009, for a review). Functional imaging evidence has linked these adaptation processes to activity in posterior inferior frontal and premotor regions (Adank & Devlin, 2010; Eisner, McGettigan, Faulkner, Rosen, & Scott, 2010), which are well placed to drive adaptation of lower-level processes in superior–temporal regions (Davis & Johnsrude, 2007). We anticipate that functional imaging investigations of perceptual retuning will provide better evidence for top–down neural processes.

Another top–down effect often proposed in language comprehension is the prediction of upcoming words on the basis of preceding sentence context (Kutas & Hillyard, 1984). This is apparent in EEG studies in which phonological mismatch responses are elicited when spoken words are different from what the listener expected on the basis of a preceding sentence or picture (Desroches, Newman, & Joanisse, 2009; Connolly & Phillips, 1994). A visual EEG study provides strong evidence for lexical prediction (DeLong, Urbach, & Kutas, 2005) because readers registered a mismatch response to the indefinite article "an" when a consonant-initial word (hence, "a") is expected (e.g., "the day was breezy so the boy went out to fly a kite" vs. "…fly an airplane"; DeLong, Urbach, & Kutas, 2005). This cannot reflect lexical integration because the evoked response arises before the critical lexical item (airplane/kite) is presented. What lexical prediction may have in common with perceptual learning is that in both cases top–down effects depend on knowledge or expectations regarding the form of incoming input. The present study, and previous work (e.g., Miller & Isard, 1963) demonstrate contextual support for word recognition without prediction of specific lexical items. Hence, at least in principle effects of sentence content on degraded speech perception can be dissociated from lexical prediction.

In summary, the present data suggest that top–down mechanisms have not yet been shown to contribute to comprehension of speech in noise. Further neural evidence will be required if we are to conclude that the impact of sentence content on report scores for degraded speech can be accurately described as arising from "top–down" neural mechanisms.

APPENDIX A: FIVE EXAMPLE PAIRS OF COHERENT AND ANOMALOUS SENTENCES

A full list of the sentence stimuli can be requested from the authors.

	Coherent	*Anomalous*
1	It was the women that complained when the old bingo hall was closed.	It was the money that exclaimed when the last eagle wall was turned.
2	The furniture in the dining room was removed when the room was decorated.	The corridor in the fishing word was survived when the word was penetrated.
3	The fireman climbed down into the bottom of the tunnel.	The warhead trained down into the sister of the barrel.
4	The new computer was sent back after the first month.	The great election was bought down between the first form.
5	The child left all of his lunch at home.	The thing felt all of his speech at line.

Acknowledgments

We would like to thank Jenni Rodd for help with stimulus preparation, Christian Schwarzbauer for programming the ISSS sequence, radiographers and staff at the Wolfson Brain Imaging Centre, University of Cambridge, for help with data acquisition, Rik Henson and Ian Nimmo-Smith for advice on image processing and statistical analysis, Dennis Norris, Jonas Obleser, Jonathan Peelle, and an anonymous reviewer for helpful comments on a previous drafts. This work was supported by the United Kingdom Medical Research Council (M. H. D.; grant MC_US_A060_0038). I. S. J. is funded by the Canada Research Chairs Program, the Canadian Foundation for Innovation, the Ontario Innovation Trust, and the Canadian Institutes of Health Research.

Reprint requests should be sent to Matthew H. Davis, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, UK, or via e-mail: matt.davis@mrc-cbu.cam.ac.uk.

Note

1. Contrast: ((anomalous$_{clear}$ + anomalous$_{0db}$) − (coherent$_{clear}$ + coherent$_{0db}$)) − ((anomalous$_{-1dB}$ + anomalous$_{-2db}$) − (coherent$_{-1dB}$ + coherent$_{-2db}$)).

REFERENCES

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *Neuroimage, 49,* 1124–1132.

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *Neuroimage, 13,* 903–919.

Awad, M., Warren, J. E., Scott, S. K., Turkheimer, F. E., & Wise, R. J. (2007). A common system for the comprehension and production of narrative speech. *Journal of Neuroscience, 27,* 11455–11464.

Bellgowan, P. S., Saad, Z. S., & Bandettini, P. A. (2003). Understanding neural system dynamics through task modulation and measurement of functional MRI amplitude, latency, and width. *Proceedings of the National Academy of Sciences, U.S.A., 100,* 1415–1419.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America, 84,* 101–114.

Bozic, M., Tyler, L. K., Ives, D. T., Randall, B., & Marslen-Wilson, W. D. (2010). Bi-hemispheric foundations for human speech comprehension. *Proceedings of the National Academy of Sciences, U.S.A., 107,* 17439–17444.

Brett, M., Anton, J.-L., Valabregue, R., & Jean-Baptiste, P. (2002). Region of interest analysis using an SPM toolbox. Paper presented at the 8th International Conference on Functional Mapping of the Human Brain, 2–6 June 2002, Sendai, Japan. Available on CD-ROM in *Neuroimage, 16*.

Buchsbaum, B. R., Olsen, R. K., Koch, P., & Berman, K. F. (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron, 48,* 687–697.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience, 6,* 256–266.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America, 119,* 1562–1573.

Cusack, R., Brett, M., & Osswald, K. (2003). An evaluation of the use of magnetic field maps to undistort echo-planar images. *Neuroimage, 18,* 127–142.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23,* 3423–3431.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top–down influences on the interface between audition and speech perception. *Hearing Research, 229,* 132–147.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. M. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134,* 222–241.

Dehaene-Lambertz, G., Dehaene, S., Anton, J. L., Campagne, A., Ciuciu, P., Dehaene, G. P., et al. (2006). Functional segregation of cortical language areas by sentence repetition. *Human Brain Mapping, 27,* 360–371.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8,* 1117–1121.

Desroches, A. S., Newman, R. L., & Joanisse, M. F. (2009). Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity. *Journal of Cognitive Neuroscience, 21,* 1893–1906.

Eisner, F., McGettigan, C., Faulkner, A., Rosen, S., & Scott, S. K. (2010). Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. *Journal of Neuroscience, 30,* 7179–7186.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavioral Research Methods, Instruments and Computers, 35,* 116–124.

Friederici, A. D., Ruschemeyer, S. A., Hahne, A., & Fiebach, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cerebral Cortex, 13,* 170–177.

Friston, K. J., Rotshtein, P., Geng, J. J., Sterzer, P., & Henson, R. N. A. (2006). A critique of functional localizers. *Neuroimage, 30,* 1077–1087.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage, 15,* 870–878.

Gow, D. W., Jr., Segawa, J. A., Ahlfors, S. P., & Lin, F. H. (2008). Lexical influences on speech perception: A Granger causality analysis of MEG and EEG source estimates. *Neuroimage, 43,* 614–623.

Hackett, T. A. (2008). Anatomical organization of the auditory cortex. *Journal of the American Academy of Audiology, 19,* 774–779.

Hackett, T. A., Stepniewska, I., & Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comparative Neurology, 394,* 475–495.

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences, 9,* 416–423.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping, 7,* 213–223.

Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in Cognitive Sciences, 10,* 64–69.

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2007). Time course of semantic processes during sentence comprehension: An fMRI study. *Neuroimage, 36,* 924–932.

Humphries, C., Willard, K., Buchsbaum, B., & Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: An fMRI study. *NeuroReport, 12,* 1749–1752.

Kaas, J. H., & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences, U.S.A., 97,* 11793–11799.

Kaas, J. H., Hackett, T. A., & Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology, 9,* 164–170.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337–1351.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience – the dangers of double dipping. *Nature Neuroscience, 12,* 535–540.

Kuperberg, G. R., McGuire, P. K., Bullmore, E. T., Brammer, M. J., Rabe-Hesketh, S., Wright, I. C., et al. (2000). Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: An fMRI study. *Journal of Cognitive Neuroscience, 12,* 321–341.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials reflect word expectancy and semantic association during reading. *Nature, 307,* 161–163.

Leff, A. P., Schofield, T. M., Stephan, K. E., Crinion, J. T., Friston, K. J., & Price, C. J. (2008). The cortical dynamics of intelligible speech. *Journal of Neuroscience, 28,* 13209–13215.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience, 31,* 2906–2915.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence-intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Marinkovic, K., Dhond, R. P., Dale, A. M., Glessner, M., Carr, V., & Halgren, E. (2003). Spatiotemporal dynamics of modality-specific and supramodal word processing. *Neuron, 38,* 487–497.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1–71.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21,* 398–421.

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., et al. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience, 5,* 467–479.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10,* 363–369.

Menon, R. S., & Kim, S. (1999). Spatial and temporal limits in cognitive neuroimaging with fMRI. *Trends in Cognitive Sciences, 3,* 207–216.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage, 11,* 735–759.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329–335.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behaviour, 2,* 217–228.

Norris, D., McQueen, J., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23,* 299–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204–238.

Obleser, J., Wise, R. J., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience, 27,* 2283–2289.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., et al. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex, 20,* 2486–2495.

Pandya, D. N. (1995). Anatomy of the auditory cortex. *Review* of *Neurology (Paris), 151,* 486–494.

Peelle, J. E., Eason, R. J., Schmitter, S., Schwarzbauer, C., & Davis, M. H. (2010). Evaluating an acoustically quiet EPI sequence for use in fMRI studies of speech and auditory processing. *Neuroimage, 52,* 1410–1419.

Peelle, J. E., Johnsrude, I. S., & Davis, M. H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in Human Neuroscience, 4,* 51.

Penny, W., & Henson, R. N. (2006). Analysis of variance. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (pp. 166–177). London: Elsevier.

Petrides, M., & Pandya, D. N. (1999). Dorsolateral prefrontal cortex: Comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *European Journal of Neuroscience, 11,* 1011–1036.

Petrides, M., & Pandya, D. N. (2009). Distinct parietal and temporal pathways to the homologues of Broca's area in the monkey. *PLoS Biology, 7,* e1000170.

Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences, 1191,* 62–88.

Pulvermuller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language, 110,* 81–94.

Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., Freund, H. J., et al. (2001). Probabilistic mapping and volume measurement of human primary auditory cortex. *Neuroimage, 13,* 669–683.

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

Righi, G., Blumstein, S. E., Mertus, J., & Worden, M. S. (2010). Neural systems underlying lexical competition: An eye tracking and fMRI study. *Journal of Cognitive Neuroscience, 22,* 213–224.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex, 15,* 1261–1269.

Sabatinelli, D., Lang, P. J., Bradley, M. M., Costa, V. D., & Keil, A. (2009). The timing of emotional discrimination in human amygdala and ventral visual cortex. *Journal of Neuroscience, 29,* 14864–14868.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics, 71,* 1207–1218.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., et al. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences, U.S.A., 105,* 18035–18040.

Schroeder, M. R. (1968). Reference signal for signal quality studies. *Journal of the Acoustical Society of America, 44,* 1735–1736.

Schwarzbauer, C., Davis, M. H., Rodd, J. M., & Johnsrude, I. (2006). Interleaved silent steady state (ISSS) imaging: A new sparse imaging method applied to auditory fMRI. *Neuroimage, 29,* 774–782.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences, 26,* 100–107.

Simon, O., Kherif, F., Flandin, G., Poline, J. B., Riviere, D., Mangin, J. F., et al. (2004). Automatized clustering and functional geometry of human parietofrontal networks for language, space, and number. *Neuroimage, 23,* 1192–1202.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217–257.

Uusvuori, J., Parviainen, T., Inkinen, M., & Salmelin, R. (2008). Spatiotemporal interaction between sound form and meaning during spoken word perception. *Cerebral Cortex, 18,* 456–466.

van den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience, 13,* 967–985.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *Neuroimage, 32,* 1826–1836.