Independence of Early Speech Processing from Word Meaning

Katherine E. Travis¹,², Matthew K. Leonard³,⁴, Alexander M. Chan⁵,⁶, Christina Torres⁸,³, Marisa L. Sizemore⁸, Zhe Qu⁶,¹⁰, Emad Eskandar⁷, Anders M. Dale¹,⁸,³, Jeffrey L. Elman²,⁴, Sydney S. Cash⁹,⁶ and Eric Halgren¹,⁸,³,⁴

¹Department of Neurosciences, ²Department of Cognitive Science, ³Department of Radiology, ⁴Kavli Institute for Brain and Mind and ⁵Joint Doctoral Program in Language and Communication Disorders, San Diego State University/University of California, San Diego, CA 92093, USA ⁶Department of Neurology and ⁷Department of Neurosurgery, Massachusetts General Hospital, Boston, MA 02114, USA ⁸Multimodal Imaging Laboratory, University of California, San Diego, La Jolla, CA 92037, USA ⁹Harvard-MIT Division of Health Science and Technology, Medical Engineering and Medical Physics, Cambridge, MA 02139, USA ¹⁰Department of Psychology, Sun Yat-sen University, Guangzhou 510275, China

Address correspondence to Eric Halgren, PhD, Multimodal Imaging Laboratory, University of California, San Diego 8950 Villa La Jolla Dr., San Diego, CA 92037-0841, USA. Email: ehalgren@ucsd.edu and Matthew K. Leonard, PhD, Multimodal Imaging Laboratory, University of California, San Diego, 8950 Villa La Jolla Dr., San Diego, CA 92037-0841, USA. Email: mkleonard@ucsd.edu

K.E.T., M.K.L., and A.M.C. contributed equally to this research.

We combined magnetoencephalography (MEG) with magnetic resonance imaging and electrocorticography to separate in anatomy and latency 2 fundamental stages underlying speech comprehension. The first acoustic-phonetic stage is selective for words relative to control stimuli individually matched on acoustic properties. It begins ~60 ms after stimulus onset and is localized to middle superior temporal cortex. It was replicated in another experiment, but is strongly dissociated from the response to tones in the same subjects. Within the same task, semantic priming of the same words by a related picture modulates cortical processing in a broader network, but this does not begin until ~217 ms. The earlier onset of acoustic-phonetic processing compared with lexico-semantic modulation was significant in each individual subject. The MEG source estimates were confirmed with intracranial local field potential and high gamma power responses acquired in 2 additional subjects performing the same task. These recordings further identified sites within superior temporal cortex that responded only to the acoustic-phonetic contrast at short latencies, or the lexico-semantic at long. The independence of the early acoustic-phonetic response from semantic context suggests a limited role for lexical feedback in early speech perception.

Keywords: ECoG, MEG, N400, speech processing

Introduction

Speech perception can logically be divided into successive stages that convert the acoustic input into a meaningful word. Traditional accounts distinguish several stages: initial acoustic (nonlinguistic), phonetic (linguistic featural), phonemic (language-specific segments) and, finally, word recognition (Frauenfelder and Tyler 1987; Indefrey and Levelt 2004; Samuel 2011). The translation of an acoustic stimulus from a sensory-based, nonlinguistic signal into a linguistically relevant code presumably requires a neural mechanism that selects and encodes word-like features from the acoustic input. Once the stimulus is in this pre-lexical form, it can be sent to higher level brain areas for word recognition and meaning integration.

While these stages are generally acknowledged in neurocognitive models of speech perception, there is much disagreement regarding the role of higher level lexico-semantic information during early word-form encoding stages. Inspired by behavioral evidence for effects of the lexico-semantic context on phoneme identification (Samuel 2011), some neurocognitive theories of speech perception posit lexico-semantic feedback to at least the phonemic stage (McClelland and Elman 1986). However, others can account for these phenomena with a flow of information that is exclusively bottom-up (Marslen-Wilson 1987; Norris et al. 2000). These models (and the behavioral data supporting them) provide important testable hypotheses to determine whether top-down effects occur during early word identification or late post-lexical processing. To date, neural evidence for or against feedback processes in speech perception has been lacking (Fig. 1B), partly because hemodynamic measures such as positron emission tomography and functional magnetic resonance imaging do not have the resolution to separate them temporally, and furthermore they find that all of these processes activate overlapping (but not identical) cortical locations (Price 2010). Temporal resolution combined with sufficient spatial localization thus provides essential additional information for untangling the dynamic interaction of the different processes contributing to speech understanding, as well as defining the road of feedback from later to earlier stages (Fig. 1B).

We sought to disambiguate various stages involved in speech processing by using the temporal precision afforded by electromagnetic recording techniques. An analogy may be drawn with the visual modality, where some evidence supports an area specialized for word-form encoding in the left posterior fusiform gyrus, peaking at ~170 ms (McCandliss et al. 2003; Dehaene and Cohen 2011). This activity reflects how closely the letter string resembles words (Binder et al. 2006), and is followed by distributed activation underlying lexico-semantic associations peaking at ~400 ms termed the N400 (Kutas and Federmeier 2011), or N400m when recorded with magnetoencephalography (MEG) (Marinkovic et al. 2003). Intracranial recordings find N400 generators in the left temporal and posteroverentral prefrontal cortex (Halgren et al. 1994a, 1994b). These classical language areas also exhibit hemodynamic activation during a variety of lexico-semantic tasks (Hickok and Poeppel 2007; Price 2010). It is also well established that auditory words evoke N400m activity (Van Petten et al. 1999; Marinkovic et al. 2003; Uusvuori et al. 2008), which begins at ~200 ms after word onset and peaks within similar distributed left fronto-temporal networks at
digms. We also performed this task using electrocorticography (Marinkovic et al., 2003) indexed as the N400m response observed using similar picture was expected to reveal lexico-semantic activity preceded by a picture stimulus that, in some cases, provided a structural MRI. During MEG recordings, adult subjects listened to single-syllable auditory words randomly intermixed with identical time-varying spectral acoustics in multiple frequency bands (Shannon et al. 1995). The contrast of words versus noise stimuli was generated using a noise-vocoding procedure that matches each word in their initial phonemes (e.g. ‘ball’ presented after a picture of a dog) so that information necessary for recognition of the mismatch would be present from word onset. Words were single-syllable nouns recorded by a female native English speaker. Sensory control stimuli were generated using a noise-vocoding procedure that matches each individual word’s time-varying spectral content (Shannon et al. 1995). Specifically, white noise was band passed and amplitude modulated to match the acoustic structure of a corresponding word in total power in each of 20 equal bands from 50 to 5000 Hz, and the exact time versus power waveform for 50–247, 248–495, and 496–5000 Hz. This procedure smears across frequencies within the bands mentioned above, rendering the stimuli unintelligible without significant training (Davis et al. 2005). These types of control stimuli have been used extensively to provide comparisons that isolate acoustic-phonetic processing (Scott et al. 2000; Scott and Johnsrude 2003; Davis et al. 2005). Following the scanning session, subjects were presented a sample of the noise stimuli without the picture context and reported

Methods

Subjects

For MEG and MRI, 8 healthy right-handed, monolingual English-speaking adults (3 males; 21–29 years) gave informed, written consent, approved by the UCSD Institutional Review Board. For intracranial recordings (ECoG), 2 patients undergoing clinical evaluation for medically intractable epilepsy with intracranial electrodes (Patient A, 29-year-old female, and Patient B, a 32-year-old male) participated in this study. Intellectual and language functions were in the average or low average range (Patient A: FSIQ 93, VIQ 85; Patient B: FSIQ 86, VIQ 91). Written informed consent was approved by the Massachusetts General Hospital IRB.

Tasks

Picture–Word Matching with Noise Control Sounds

In the primary task, an object picture (<5° visual angle) appeared for the entire 1300 ms trial duration (600–700 ms interval trial). Five hundred milliseconds after picture onset, either a congruously or incongruously paired word or noise stimulus was presented binaurally (Fig. 1). Sounds (mean duration = 445 ± 63 ms; range = 304–637 ms; 44.1 kHz; normalized to 65 dB average intensity) were presented binaurally through plastic tubes fitted with earplugs. Four conditions (250 trials each) were presented in a random order: picture-matched words, picture-mismatched noise, picture-mismatched words, and picture-mismatched noise. Behavioral responses were recorded primarily to ensure that subjects maintained attention during the experiment. Participants were instructed to key press when the sound they were hearing matched the picture being presented. The response hand alternated between 100 trial blocks. Subjects were not instructed as to the type of sound stimuli (words, noise) that would be played during the experiment. Incongruous words differed from the correct word in their initial phonemes (e.g. “ball” presented after a picture of a dog) so that information necessary for recognition of the mismatch would be present from word onset. Words were single-syllable nouns recorded by a female native English speaker. Sensory control stimuli were generated using a noise-vocoding procedure that matches each individual word’s time-varying spectral content (Shannon et al. 1995). Specifically, white noise was band passed and amplitude modulated to match the acoustic structure of a corresponding word in total power in each of 20 equal bands from 50 to 5000 Hz, and the exact time versus power waveform for 50–247, 248–495, and 496–5000 Hz. This procedure smears across frequencies within the bands mentioned above, rendering the stimuli unintelligible without significant training (Davis et al. 2005). These types of control stimuli have been used extensively to provide comparisons that isolate acoustic-phonetic processing (Scott et al. 2000; Scott and Johnsrude 2003; Davis et al. 2005). Following the scanning session, subjects were presented a sample of the noise stimuli without the picture context and reported

Figure 1. Experimental design. (A) Trials present words (preceded by a congruous or incongruous picture visible for the duration of the trial) or matched noise control sound. In the auditory modality, N400 activity is typically preceded by evoked activity in the posterior superior temporal region termed the N100 (or M100 in MEG). The N100/M100 is a composite of different responses (Nätänen and Picton 1987), and several studies have found that basic phonetic/phonemic parameters such as voice onset time can produce effects in this latency range (Gage et al. 1998, 2002; Frye et al. 2007; Uusvuo et al. 2008). However, it is not known when, during auditory word recognition, the phonetic and phonemic processing which eventually leads to word recognition diverges from nonspecific acoustic processing. Furthermore, it is unknown whether such processing is influenced by the lexico-semantic contextual manipulations which strongly influence the N400m.

In order to investigate the spatiotemporal characteristics of neural responses representing successive stages during speech perception, we employed a well-established and validated neuroimaging technique known as dynamic statistical parametric mapping (dSPM; Dale et al. 2000) that combines the temporal sensitivity of MEG with the spatial resolution of structural MRI. During MEG recordings, adult subjects listened to single-syllable auditory words randomly intermixed with unintelligible matched noise-vocoded control sounds with identical time-varying spectral acoustics in multiple frequency bands (Shannon et al. 1995). The contrast of words versus noise stimuli was expected to include the neural processes underlying acoustic-phonetic processing (Davis et al. 2005; Davis and Johnsrude 2007). These stimuli were immediately preceded by a picture stimulus that, in some cases, provided a semantic prime. The contrast between words that were preceded by a semantically congruous versus incongruous picture was expected to reveal lexico-semantic activity indexed as the N400m response observed using similar (Marinkovic et al. 2003) or identical (Travis et al. 2011) paradigms. We also performed this task using electrocorticography (ECoG) in 2 patients with semi-chronic subdural electrodes, allowing us to validate the timing and spatial localization inferred from MEG, and to discern distinct sub-centimeter cortical organization in posterior superior temporal regions for acoustic-phonetic and lexico-semantic processing. These results were also replicated with MEG in a passive listening task using single-syllable words spoken by a different speaker. Subjects were also presented with a series of tones at the end of the MEG recording session, in order to distinguish the word-evoked responses from the well-studied M100 component. While, as described above, the acoustic-phonetic and lexico-semantic stages of speech processing have been examined extensively in separate tasks, to our knowledge, this study is the first to isolate both stages and compare their onset and interaction within the same task and using the same stimuli.
being unable to name the underlying words. Intelligibility of noise stimuli were examined in an additional pilot experiment in which adult subjects (n = 22) were asked to listen and rate on a scale (1 =
low confidence to 7 = high confidence) how well they could understand a random series of noise voiced sounds (n = 200; constructed in the same manner as the main experiment) interspersed with a sub-
sample of corresponding words (n = 50). Subjects were instructed that all stimuli were English words, some of which had been made noisy. Despite these instructions, subjects rated the noise stimuli as minimally intelligible; confidence ratings to noise were 2.17 ± 0.54, versus
6.72 ± 0.23 for words.

Tones
Following the picture–word task, subjects listened to 180 1000 Hz binaural tones at 1 Hz while maintaining fixation. This task was used to evoke nonlinguistic acoustic processing for comparison with the hypothesized acoustic-phonetic response in the word–noise compari-
son. Data from 1 subject were lost due to an equipment malfunction.

Neuroimaging

MEG and MRI
The procedures involved for both the recording and the post-
processing of MEG and MRI data have been described previously (Leonard et al. 2010; Travis et al. 2011) and are only briefly described here. Two hundred and four planar gradiometer channels and 102 magnetometer channels distributed over the scalp were recorded at 1000 Hz with minimal filtering (0.1–200 Hz) using an Elekta Neuro-

mag Vectorview system. Due to lower signal-to-noise ratio and more prominent artifacts, magnetometer data were not used; previous studies have found similar source localizations when gradiometer and magnetometer data are analyzed with the current methods in cogni-
tive tasks using auditory stimuli (Halgren et al. 2011). The MEG data were epoched from −200 to 800 ms relative to the onset of the audi-
tory stimuli, low-pass filtered (50 Hz), and inspected for bad channels (channels with excessive noise, no signal, or unexplained artifacts), which were excluded from all further analyses. Blink artifacts were removed using independent component analysis (Delorme and Makeig 2004) by pairing each MEG channel with the electrooculo-
gram (EOG) channel and by rejecting the independent component that contained the blink. If the MEG evoked by a particular word was rejected due to an artifact from the word average, then the corre-
sponding noise stimulus was removed from the noise average and vice versa. Cortical sources of MEG activity were estimated using a linear minimum-norm approach, noise normalized to a pre-stimulus period (Dale et al. 2000; Liu et al. 2002). Candidate cortical dipoles and the boundary element forward solution surfaces were located in each subject from 3D T1-weighted MRI. Regional time courses were extracted from regions of interest (ROIs) on the resulting dSPM maps, and were tested for between-condition differences. The average head movement over the session was 5.3 ± 3.6 mm (2.9 ± 1.1 mm for the passive listening experiment, Supplementary Fig. S1).

ROI Analysis
Specific ROI locations were determined by visual inspection of group average dSPM maps without regard to condition and then automatically projected to individual brains by aligning the sulcal–gyral patterns of their cortical surfaces (Fischl et al. 1999). For the early acoustic-phonetic response, ROIs were selected in 2 bilateral regions exhibiting the largest peak in group average responses to all words and noise (90–110 ms; Supplementary Fig. S5). Six additional ROIs were selected in bilateral fronto-temporal regions exhibiting the largest group average responses to all words during a period when N400m activity to auditory words is known to occur (200–400 ms; Supplementary Fig. S5). A 20 ms time window surrounding the largest peak in group average activity to all words (250–300 ms) was selected to display and test for the later semantic priming effects (Fig. 2, Supplementary Fig. S2).

Intracranial Methods
While MEG provides whole-brain coverage and allows for better cross-subject averaging, MEG and EEC source estimation methods (e.g. dSPM) are inherently uncertain because the inverse problem is ill-posed. We thus confirmed dSPM results by recording ECoG in 2 patients who performed the identical word–noise task as the healthy subjects in the MEG. ECoG was recorded from platinum subdural surface electrodes spaced 1 cm apart (Adtech Medical, Racine, WI) over left posterior–superior temporal gyrus semi-chronically placed to localize the seizure origin and eloquent cortex prior to surgical treat-
ment. In one of these subjects, an additional 2 by 16 contact micro-
grid (50 µm platinum–iridium wires embedded in, and cut flush with the silastic sheet, Adtech Medical, Racine, WI) was implanted over the same area (Fig. 4, Supplementary Fig. S3). Electrodes were localized by registering the reconstructed cortical surface from preoperative MRI to the computed tomography performed with the electrodes in situ, resulting in an error of <3 mm (Dykstra et al. 2012). High gamma power (HGP) from 70 to 190 Hz was estimated using wavelets on individual trials and weighted by frequency (Chan et al. 2011). ECoG recordings, especially HGP, are primarily sensitive to the tissue immediately below the recording surface (0.1 mm diameter for micro-

grids, 2.2 mm for macrogrids), and the distance between electrode centers is also small (1 mm for microgrids, 10 mm for macrogrids). In contrast, analysis of the point-spread function of dSPM (Dale et al. 2000; Liu et al. 2002) and the lead-fields of MEG planar gradiometers on the cortical surface (Halgren et al. 2011) suggests that activity in a ~3 cm diameter cortical patch can contribute to an MEG response. Thus, the certainty of spatial localization, as well as the ability to resolve different closely spaced responses, is both greater for ECoG than MEG. Although patients suffered from long-standing epilepsy, the recorded contacts were many centimeters away from the seizure focus, and spontaneous EEG from the local cortex appeared normal.

Sensor-Level Statistics
Significance of responses at single MEG or ECoG sensors in planned comparisons between task conditions was tested using random resampling (Maris and Oostenveld 2007). Individual trials were ran-
donically assigned to different conditions, and a t-test was performed across trials of sensor values (potential in ECoG, or flux gradient in MEG) between the different pseudo-conditions at each latency. For each randomization (performed 500 times), the duration of the longest continuous string of successive latencies with P<0.05 was saved to create the distribution under the null hypothesis. The same t-test was then applied to the actual trial assignment, and all signifi-
cant strings longer than the 5th longest string from the randomization was considered to be significant at P<0.01 (because 0.01 = 5/500). This statistic does not require correction for multiple comparisons at different latencies (Maris and Oostenveld 2007).

Results
Early differential activity presumably reflecting acoustic-

phonetic processes was found by contrasting MEG responses to words and their matched noise controls. We then isolated activity related to lexico-semantic processing by contrasting MEG responses to the same words and task when they were preceded by a congruous versus incongruous picture. The time course and inferred source localization of these responses were confirmed and refined with intracranial recordings in the same task. Comparison of these contrasts revealed that acoustic-phonetic processes occur prior to top-down lexico-semantic effects, and in partially distinct cortical locations.

The word>noise contrast revealed an MEG response that peaked in a left posterosuperior temporal sensor at ~100 ms
responses that vary with phonemic characteristics of sublexical context and constructed in the same manner as the main stimuli to the subjects after the MEG recording session (see Methods; Supplementary Fig. S1). Again, comparison between words and incongruous words with a key press (97% ± 6.21) and omitting responses for incongruous words (99.6% ± 0.52). In contrast, the accuracy for noise stimuli was more variable. Participants key pressed correctly on 67.5% ± 30.87 of the trials when the noise matched the picture, and witheld responding on 100% of the mismatched trials. Overall, subjects responded significantly slower to matched noise (676.26 ms ± 102.9) than to matched words (558.95 ms ± 96.13; t(7) = 11.03, P < 0.00001), and were more accurate to words, t(7) = 3.18, P < 0.01. This suggests that noise stimuli contained sufficient sensory information to guess above chance when a noise sound was derived from a word that matched a picture context. However, it is unlikely that the noise controls contained adequate phonemic information necessary for lexical identification, and indeed this level of performance did not require that the words be uniquely identified. We tested this explicitly in both a pilot experiment with noise stimuli presented out of context and constructed in the same manner as the main experiment, and also by presenting a sample of the noise stimuli to the subjects after the MEG recording session (see Methods). During the task, the subjects had a specific word in mind from seeing its picture. They were able to discern at above chance levels if the amplitude envelope or other low-level, nonlexical, or semantic characteristic of the noise was consistent with the target word, and if so, they responded with a key press. The fact that the subjects were able to guess above chance indicates that, first, the picture did activate the presence of the fundamental frequency (Parviainen et al. 2005).

Thus, we refer to the word>noise response as the M100p, an acoustic-phonetic selective component of the M100.

Subjects were highly accurate at correctly identifying congruous words with a key press (97% ± 6.21) and omitting responses for incongruous words (99.6% ± 0.52). In contrast, the accuracy for noise stimuli was more variable. Participants key pressed correctly on 67.5% ± 30.87 of the trials when the noise matched the picture, and withheld responding on 100% of the mismatched trials. Overall, subjects responded significantly slower to matched noise (676.26 ms ± 102.9) than to matched words (558.95 ms ± 96.13; t(7) = 11.03, P < 0.00001), and were more accurate to words, t(7) = 3.18, P < 0.01. This suggests that noise stimuli contained sufficient sensory information to guess above chance when a noise sound was derived from a word that matched a picture context. However, it is unlikely that the noise controls contained adequate phonemic information necessary for lexical identification, and indeed this level of performance did not require that the words be uniquely identified. We tested this explicitly in both a pilot experiment with noise stimuli presented out of context and constructed in the same manner as the main experiment, and also by presenting a sample of the noise stimuli to the subjects after the MEG recording session (see Methods). During the task, the subjects had a specific word in mind from seeing its picture. They were able to discern at above chance levels if the amplitude envelope or other low-level, nonlexical, or semantic characteristic of the noise was consistent with the target word, and if so, they responded with a key press. The fact that the subjects were able to guess above chance indicates that, first, the picture did activate the

(Fig. 2A, C, H). When examined in each subject separately, this sensor showed a similar significant early difference between individual word and noise trials using a nonparametric randomization test with temporal clustering to correct for multiple comparisons (Maris and Oosterveld 2007; Fig. 2A, H). These effects were replicated in an additional experiment in which 9 subjects (5 repeat subjects) passively listened to a separate set of single syllable words (recorded by a different speaker), and noise control stimuli (used in a pilot experiment to examine noise intelligibility (see Methods; Supplementary Fig. S1). Again, comparison between words and their individually matched controls revealed a significant response in the left postero-temporal area in this latency range, thus confirming the generality of the response across stimuli, speakers, and tasks.

Although each noise stimulus matched its corresponding word in its acoustic characteristics, sufficient differences had to be present to render the noise unintelligible. These differences entail the possibility that the word>noise response may reflect acoustic modulation of the generic M100. However, direct comparison of the word>noise response to the M100 evoked by tones shows that they are lateralized to opposite hemispheres in both individual-subject sensors (Fig. 3A) and group-based estimated localization (Fig. 3B). A further indication that the word>noise effect does not result from nonspecific sensory differences is the lack of differences in any channel at early latencies (e.g. Fig. 2A, H), and at any latency in surrounding channels. Conversely, the word>noise response occurs at about the same time and location as MEG responses that vary with phonemic characteristics of sublexical stimuli such as voice onset time (Frye et al. 2007) or the

Figure 2. Acoustic-phonetic processing indexed by M100p is distinct from later lexico-semantic processing indexed by N400m. (A) Single subject left temporal gradiometer shows early word>noise response beginning ∼52 ms after stimulus onset. Shaded gray areas represent the first cluster of significance in the event-related time-course for each subject, as determined by Monte Carlo between-condition statistics (P < 0.01) (Maris and Oosterveld 2007). (B) The same left temporal gradiometer channel shows an incongruous>congruous word difference at ∼244 ms (192 ms after M100p onset). No semantic difference is observed during M100p response. (C) Plot of all 204 gradiometers indicating the location of the left temporal channel shown in (A, B and H). (D–G). Estimated cortical localization of group average activity using dSPM (8 subjects)). The earliest significant words>noise response peaks in superior temporal regions between 90 and 110 ms (D) and becomes more distributed by later time windows (E). Significant incongruous>congruous semantic effects are absent at ∼100 ms (F), occurring later in both hemispheres, especially left (G). Color bars represent square-root of F-values, which are a measure of signal-to-noise. (H) Single-subject left temporal gradiometer responses showing M100p onset in each additional subject. (I) N400 responses in the same channel for each subject, with all positive onset differences indicating temporal separation between early and late stages.
desired lexical element, and secondly that the noise stimuli were sufficiently well matched on their acoustic characteristics that they permitted accurate guessing. The fact that the noise stimuli could not be recognized when presented in isolation shows that they do not adequately activate acoustic-phonetic elements sufficient for word recognition.

Figure 4. Confirmation of spatiotemporally distinct acoustic-phonetic and lexico-semantic processing stages with macro and micro-electrode recordings from the superior temporal gyrus. HGP and LFP were recorded from macrogrid contacts on the external aspects of Heschl’s gyrus (1) and transverse temporal sulcus (2, white arrow), and from microgrid contacts on the planum temporale (3) and angular gyrus (4). Words evoked greater HGP (filled orange arrows) and LFP (filled brown arrow) than matched noise at \( \sim 100 \text{ ms} \); at the same latency and locations, no difference was observed between congruous (cong) and incongruous (incong) words (open orange and brown arrows). In other sites, at \( \sim 400 \text{ ms} \), incongruous words evoke greater HGP (filled magenta arrows) and LFP (filled cyan arrows) than do congruous words; at the same latency, some of these locations may show differential responses to words versus noise (open magenta and cyan arrows). Gray boxes indicated periods with significant differences between conditions at \( P < 0.01 \) using resampling statistics. Additional channels are shown in Supplementary Figure S3, documenting a complex mosaic of sites with different levels of activity and task correlates for this patient (Pt. A). An additional subject (Pt. B) with similar results is shown in Supplementary Figure S4.
In order to investigate whether top-down lexico-semantic information can modulate this initial acoustic-phonetic processing, we compared the early MEG response to words when its meaning had been preactivated with a congruous picture versus the response to the same word when it was preceded by an incongruous (control) picture. As expected, this contrast revealed a distributed left fronto-temporal incongruous>congruous difference peaking at ~400 ms, i.e. a typical N400m associated with lexical access and semantic integration (Fig 2B, F, G, I). We examined the response in the left postero-temporal MEG sensor where the maximal M100p to the same words was recorded in the same task using a Monte Carlo random effects resampling statistic to identify the onset of the difference between the 2 conditions (Maris and Oostenveld 2007). Critically, this difference did not begin until ~150 ms after the beginning of the word>noise difference. Across the group, we found that word>noise differences onset significantly earlier (average onset 61 ± 22 ms) than incongruous-congruous semantic priming effects (average onset 217 ± 130 ms; t(7) = −3.51, P < 0.01). Examination of individual subject responses in this same left temporal sensor further confirmed that M100p activity consistently occurred prior to the onset of semantic priming effects for all participants, despite relatively large variability in the onset of the later response (Fig 2A, B, H, I). Post hoc analyses of the M100p time window revealed that the early semantic responses (<120 ms; Fig. 2D) observed in 2 subjects (Supplementary Figs S6 and 8) were likely driven by strategic differences in how these subjects performed the experimental task.

Additional post hoc power analyses were performed to determine whether a semantic effect might be present immediately after the onset of the M100p response, but was not detected due to the lack of power. We found that the effect size (comparing congruous versus incongruous words) was extremely small (Cohen’s d = 0.05) from this same channel which showed a strong word versus noise effect, measured during the first 20 ms after the onset of the M100p response in each subject (i.e. on average from 61 to 81 ms after stimulus onset; see Fig. 2A, B, H, I). In contrast, a much larger effect size was obtained for M100p responses during this same time (Cohen’s d = 0.91). Conversely, a large semantic effect size was clearly observed both at 50 ms following the onset of semantic effects in each subject (i.e. on average from 217 to 267 ms after stimulus onset; Cohen’s d = 0.93), as well as during the 250–300 ms time window when semantic effects were a priori predicted to occur (Cohen’s d = 1.33). Thus, the time surrounding the onset of the M100p (~60–80 ms) is not affected by lexico-semantic context, which only begins to exert its influences later. Together, both group and individual subject analyses suggest that the acoustic-phonetic processes indexed by the M100p are initially independent of the semantic processes indexed by the N400m.

The cortical sources of these responses were estimated with dSPM in each subject, and then averaged across subjects on the cortical surface (Dale et al. 2000). The cortical distribution for words versus noise during the time of the M100p (90–110 ms) concentrated mainly on superior temporal regions, especially on the left (Fig. 2D, E; Supplementary Figs. S1 and 2). No significant differences to incongruous versus congruous words were observed at this time, but were present during later windows (200–400 ms; Fig. 2F, G) in the left inferior frontal, insular, ventral temporal, and posterior superior temporal regions. Right hemispheric activity was concentrated mainly within insular and superior temporal regions (Fig. 2G). Such differences are consistent in their task correlates, timing, and left temporal distribution with previous dSPM estimates of N400m activity using similar (Marinkovic et al. 2003) or identical paradigms (Travis et al. 2011; Leonard et al. 2012). Random effects tests of dSPM values in cortical regions of interest generally confirmed these maps for both the early acoustic-phonetic response in superior temporal regions and the lexico-semantic effect in more widespread areas (Supplementary Fig. S2). Specifically, the regions selected to test for M100p responses from 90 to 110 ms exhibited activity that was significantly greater to words than noise in left superior temporal sulcus (STS) (F(1,7) = 13.50, P < 0.01, right STS (F(1,7) = 13.12, P < 0.01), and left planum temporale (PT) (F(1,7) = 12.37, P < 0.01). Only from 250 to 300 ms when lexico-semantic effects were predicted to occur did these areas (left PT (t(7) = 2.43, P < 0.046; with trends in left STS (t(7) = 2.12, P < 0.072 and right STS (t(7) = 2.30, P < 0.055) and other left temporal areas selected a priori to test for later semantic effects (anterior inferior temporal sulcus (t(7) = 2.37, P < 0.05), posterior STS (t(7) = 2.02, P < 0.083, trend) show significantly greater responses to incongruous versus congruous words.

Due to their limited spatial resolution, these MEG results are ambiguous as to whether the same locations that perform acoustic-phonetic processing at short latencies also participate in lexico-semantic processing at longer latencies. On the one hand, the cortical locations estimated with dSPM as showing an early word>noise effect (Fig. 2D) are a subset of the areas showing a late incongruous>congruous effect (Fig. 2G). This overlap is almost complete within the left posterior sTg, suggesting that top-down semantic influences may be ubiquitous at longer latencies in areas participating in acoustic-phonetic processing at short latencies. However, it remains possible that small areas within the sTg are specialized for early acoustic-phonetic processing, and adjacent small areas are devoted to late lexico-semantic processing, but their projections to the MEG sensors are too overlapping to be resolved. In order to test for this possibility, we turned to the high spatial resolution afforded by recordings in the same task made directly from the cortical surface (ECoG). Intracranial recordings also allow HGP to be recorded in addition to local field potentials (LFPs; Jerbi et al. 2009). To determine the timing and locations of the onset of the M100p and N400 effects, we used the same Monte Carlo resampling statistic described above for the MEG data.

Clear evidence was found for a partial spatial segregation of sites in left posterior sTg responding to words>noise at short latencies, versus incongruous>congruous words at long latencies (Fig. 4, Supplementary Figs S3 and 4). For example, in Figure 4, cortex underlying contact 4 generates words>noise HGP and LFP responses in the 80–120 ms range (orange and brown arrows), but no late incongruous>congruous LFP difference until after 600 ms, following the behavioral response. In contrast, the cortex underlying contact 3 (~1 cm anterior to contact 4) responds to incongruous>congruous words with LFP and HGP starting at ~200 ms, but does not show a words>noise effect until after 400 ms. Contact 2, 1 cm anterior to contact 3, shows both the early acoustic-phonetic and late lexico-semantic effects, in LFP but not HGP. Contact 1, 1 cm anterior to contact 2, shows the early acoustic-
phonetic effects in HGP but not LFP. In addition, those sTg contacts which showed the early words>noise response in HGP also showed significantly different HGP responses at similar latencies to different initial consonants in the same words, providing further evidence that this early response is related to phoneme processing (Fig. 5). Thus, the intracranial recordings validate the MEG results, and further demonstrate that the cortical domains devoted to early acoustic-phonetic and later lexico-semantic processing are anatomically distinct, at least in part, but intermingled within the posterior sTg.

Discussion

The present study combined MEG and MRI, and ECoG in patients with semi-chronic subdural electrodes, to distinguish in latency, anatomy, and task correlates 2 neural components reflecting distinct stages during speech comprehension. Within the same evoked cortical response to words, activity reflecting acoustic-phonetic processing (M100p) was separated from activity indexing lexico-semantic encoding (N400m). A words>noise difference isolated acoustic-phonetic activity as beginning at ~60 ms and peaking ~100 ms after word onset, localized to posterior superior temporal cortex (M100p; Fig. 2A, D). This response was followed by more widespread fronto-temporal activity beginning at ~200 ms, sustained for ~300 ms, and associated with lexico-semantic processing (“N400m”; Fig. 2B, G). Both components were stronger in the left hemisphere. Despite individual differences in the timing of M100p and N400m (Fig. 2H, I), we found no evidence for interactions from top-down lexico-semantic processing during the initial period of words>noise effects. These findings were validated with ECoG recordings obtained from 2 additional subjects who had been implanted with electrodes for clinical purposes. Acoustic-phonetic and lexico-semantic responses were located in distinct domains of the superior temporal gyrus separated by <1 cm.

To isolate an acoustic-phonetic processing stage, we contrasted the responses evoked by words to those elicited by their acoustically matched noise controls. This comparison revealed a differential cortical response which began 61 ms, on average, after the sound onset. Considering that it takes ~13 ms for auditory information to arrive in the cortex (Liégeois-Chauvel et al. 1994), we infer that the distinguishing acoustic information reflected in the words>noise response must be contained within the first ~48 ms of the word sound (61–13 ms). This requires that the distinctive feature be at a relatively low segmental level, at least initially. Presumably, like early fusiform responses to visual words (McCandliss et al. 2003; Dehaene and Cohen 2011) and faces (Halgren et al. 2006), the M100p likely encodes essential acoustic-phonetic elements contained within the initial segment of a word which are later combined arbitrarily into symbols pointing to semantics. Indeed, it is likely that the present words>noise response reflects overlapping or even identical acoustic-phonetic processes previously found to peak at ~100 ms in the MEG activity evoked by acoustic-phonetic and phonological aspects of speech sounds (Eulitz et al. 1995; Poeppel et al. 1996; Gootjes et al. 1999; Vihla and Salmelin 2003; Parviainen et al. 2005; Frye et al. 2007). However, further studies are needed to establish the specific sensitivity of the M100p to prelexical acoustic features.

While it is impossible to completely eliminate any possible contribution of sensory differences to the M100p, it is unlikely that the M100p reflects only low-level sensory processing. This is evidenced by its similar spatiotemporal characteristics when evoked by words spoken by voices of different speakers (Supplementary Fig. S1), and its clear differentiation from the M100 to tones (Fig. 3). Rather, the M100p has a similar

**Figure 5.** Differential response to different consonants with similar timing and locations as differential responses to word versus noise in the left sTg. In the upper row are shown the HGP responses from macroelectrode contacts in the sTg to word versus noise are reproduced from pt. A (Fig. 4 and Supplementary Fig. S3) and pt. B (Supplementary Fig. S4). In the lower row, the HGP responses to words are separately averaged according to their initial consonants. The locations (Pt. A contact 1; Pt. B contacts 1 and 3) which show significant differences at early latencies between words and noise (indicated by gray shading), also show significant differences at similar latencies between different initial consonants (peach shading).

Downloaded from https://academic.oup.com/cercor/article-abstract/23/10/2370/293215 by guest on 13 April 2019
latency and anatomical location to previously identified acoustic-phonetic responses in MEG (see above), hemodynamic studies, and intracranial recordings (reviewed below). Further evidence that the words > noise difference reflects processing at the acoustic-phonetic level was obtained from intracranial HGP recordings: ECoG contacts that responded differentially to words > noise also responded differentially at a similar latency to different initial consonants (Fig. 5). Unlike MEG and LFP, where a larger response may result from either inhibition or excitation of the generating neurons, HGP reflects integrated high-frequency synaptic activity and/or action potentials (Crone et al. 2011), and is highly correlated with the BOLD response (Ojemann et al. 2010). Thus, even if there is some contribution of sensory differences to the M100p, these considerations indicate that its major generators are early phonetic selective synaptic activity that performs the acoustic-phonetic encoding which ultimately leads to lexical identification and semantic integration. However, it is important that future studies continue to characterize the specific perceptual attributes responsible for evoking the M100p response by employing a variety of acoustic controls.

The ability of some subjects to rapidly shadow a recorded passage (Marslen-Wilson 1975), and the priming effects on visual words when presented at different points in an auditory passage (Zwitserlood 1989), both suggest that some lexico-semantic information becomes available at ~150 ms after word onset, in reasonably good agreement with the average 217 ms latency of the lexico-semantic effects reported here. By ~200 ms when semantic effects are seen, enough of the word has been presented so that it is possible to predict how it might be completed. Specifically, our results are consistent with several lexical processing models that have proposed that at least the initial syllable of a word (~150 ms) must be analyzed before contact is initiated with the lexicon (Frauenfelder and Tyler 1987; Marslen-Wilson 1987; Norris et al. 2000). However, it is long before the acoustic stimulus contains enough information to definitively and uniquely identify the word. Thus, lexico-semantic modulation observed here likely reflects the multiple lexical possibilities consistent with the initial ~204 ms (~217 − 13) of the stimulus, as predicted by some models of speech understanding (Marslen-Wilson 1987; Norris et al. 2000).

While it is possible to infer from previous M/EEG and ECoG studies when acoustic-phonetic and lexico-semantic stages may occur during speech comprehension, to our knowledge, our study is the first to directly compare their relative spatial and temporal characteristics within the same task, subjects, and using the same word stimuli. Indeed, our evidence for the timing and anatomy of acoustic-phonetic and lexico-semantic effects is consistent with both neurophysiological and hemodynamic activity associated with these processing stages studied in separate tasks. Here, the localization of early words > noise effects estimated from MEG and ECoG to the posterior and middle levels of the superior temporal gyrus and sulcus correspond closely to the areas showing hemodynamic activation associated with prelexical processing (Hickok and Poeppel 2007; Price 2010). Similarly, the localization of later incongruous > congruous word effects estimated from MEG correspond to those found with hemodynamic methods to be active during lexico-semantic processing, reflecting a hypothesized ventral and anterior pathway for speech recognition (Hickok and Poeppel 2007; Binder et al. 2009). Both words > noise and incongruous > congruous MEG differences are bilateral with left predominance, consistent with hemodynamic activations (Hickok and Poeppel 2007; Binder et al. 2009; Price 2010).

The timing and sources of acoustic-phonetic effects seen here are also consistent with previous studies that have found that LFP and HGP in the left posterior sfg distinguish between different phonemes at ~100 ms latency (Chang et al. 2010; Steinschneider et al. 2011) and between words and noise at ~120 ms (Canolty et al. 2007). However, these studies did not determine whether this activity is sensitive to top-down lexico-semantic influences. Conversely, repetition-modulated N400-like activity has been recorded in this region with LFP (Halgren et al. 1994a) and HGP (McDonald et al. 2010), at a latency of ~240–300 ms, but the sensitivity of these areas to acoustic-phonetic processing was not determined. The onset of lexico-semantic effects in the current study is consistent with previous N400 recordings that do not observe semantic priming effects until ~200 ms post-stimulus even when the initial phoneme of an auditory word presented in a sentential context is mismatched to the predicted completion of a congruous sentence (Van Petten et al. 1999). The timing of lexico-semantic effects seen here is also compatible with the latency from word onset of MEG activity associated with lexical (Pulvermüller et al. 2001) and semantic (Pulvermüller et al. 2005) processing isolated during a mismatch negativity paradigm. Taken together, the present findings provide strong evidence, within the same task and subjects, for distinct stages of auditory word processing, representing early acoustic-phonetic versus later lexico-semantic speech processing, and distinguished by their latency, location, and task correlates.

To summarize, our study demonstrates that, on average, the first ~150 ms of acoustic-phonetic activity is unaffected by the presence of a strong lexico-semantic context. This reveals a stage in processing where language-relevant properties of the speech signal have been identified (this is usually considered the acoustic-phonetic stage), but which is unaffected by top-down influences from context-driven lexico-semantic representations. The present data do not rule out the potential for interactions between prelexical and lexico-semantic processes at longer latencies (Figs 2 and 4, Supplementary Fig. S3), which may support the effects of lexico-semantic context on phoneme identification (Samuel 2011). Statistical correlations found between MEG activity estimated to the supramarginal gyrus and the posterior superior temporal gyrus indicate that top-down influences may occur during the time period from 160 to 220 ms following word onset (Gow et al. 2008). However, the current results indicate that initial processing of the word is not affected by lexico-semantic information. The present findings establish the neural basis for an acoustic-phonetic level of processing that can be studied using lexical stimuli, and provide a strong physiological constraint on the role of top-down projections in computational models of speech processing (McClelland and Elman 1986).

**Supplementary Material**

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.
Authors’ Contributions

Funding
This study was supported by the Kavli Institute for Brain and Mind, NIH R01 NS018741, and NSF BCS-0924539. K.E.T. and M.K.L. have been supported by NIH pre-doctoral training grants DC000041 and MH020002 and the Chancellor’s Collaboratories Award, UCSD. J.L. Evans supported the development of stimuli with funding from NIH R01-DC005650.

Notes
The authors thank J. Sherfey and D. Hagler for their generous technical support and M. Borzello and J. Naftulin for assisting in data collection. Conflict of Interest: None declared.

References


