Published in final edited form as:
Brain Lang. 2014 July ; 134: 1–10. doi:10.1016/j.bandl.2014.03.008.

The role of visual representations during the lexical access of spoken words

Gwyneth Lewis and **David Poeppel**
Department of Psychology New York University 6 Washington Place, 2nd floor New York, NY 10003

Abstract

Do visual representations contribute to spoken word recognition? We examine, using MEG, the effects of sublexical and lexical variables at superior temporal (ST) areas and the posterior middle temporal gyrus (pMTG) compared with that of word imageability at visual cortices. Embodied accounts predict early modulation of visual areas by imageability - concurrently with or prior to modulation of pMTG by lexical variables. Participants responded to speech stimuli varying continuously in imageability during lexical decision with simultaneous MEG recording. We employed the linguistic variables in a new type of correlational time course analysis to assess trial-by-trial activation in occipital, ST, and pMTG regions of interest (ROIs). The linguistic variables modulated the ROIs during different time windows. Critically, visual regions reflected an imageability effect prior to effects of lexicality on pMTG. This surprising effect supports a view on which sensory aspects of a lexical item are not a consequence of lexical activation.

Introduction

Speech perception can be intuitively described as a sequential process involving the piecemeal mapping of continuous acoustic signals onto phonetic units of some form. Less straightforward are the transitional processes and representations leading to lexical retrieval (Poeppel, Idsardi, & van Wassenhove, 2008). One particularly thorny problem in the context of lexical processing concerns the hypothesized role of perceptual representations, an issue emphasized by embodiment models in lexical semantic access (e.g., Pulvermüller, 1999). How and when do the acoustic signals and/or phonetic units of speech activate visual representations of a word's real world referent? Does the word *strawberry*, for example, automatically activate a mental image of the color red (for example in the sense of feature spreading), and if so, is such activation a requirement of or merely incidental to lexical access? One especially important issue concerns the temporal dynamics of lexical access. To answer such questions, we examine these issues in the context of the most notable models of

Corresponding Author: Gwyneth Lewis, gwyneth@nyu.edu, phone: (212) 998-8072.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

lexical processing, which differently emphasize the temporal dynamics and access stages of speech recognition.

Models of lexical processing

Neurocognitive models of lexical access suggest the participation of distinct neural regions in the activation of, competition between, and selection of basic sound-meaning representations. In the visual domain, studies of lexical access have determined that certain MEG components are sensitive to the orthographic, morphological, and semantic features of words during different time windows. For example, the frequency of a word's adjacent letter strings (bigram frequency) modulates occipital activation at ~100 ms post word onset (Solomyak & Marantz, 2010). Around 50 ms later, the morphological transition probability of words (the probability of the whole word form, given the stem) modulates responses in the fusiform gyrus (Lewis, Solomyak, & Marantz, 2011; Solomyak & Marantz, 2010). At ~300 ms, properties of the whole word form modulate a superior temporal response (Lewis, Solomyak, & Marantz, 2011; Pylkkänen & Marantz, 2003; Simon, Lewis, & Marantz; Solomyak & Marantz, 2009; Solomyak and Marantz, 2010). How do the stages of spoken word recognition compare with those involved in visual word recognition?

Interaction-competition models of lexical access and comprehension have attempted to describe the mappings between phonetic units and lexical representations in terms of activation and competition between lexical competitors. The original Cohort model, for example, describes lexical retrieval as a strictly bottom-up selective process wherein the incoming speech signal activates all words beginning with the initial phoneme and gradually winnows the selection as word information accrues, eventually discarding all competitors until only the target word remains (Marslen-Wilson, 1987). Another interaction-competition model, TRACE, depicts speech comprehension as a cumulative process wherein a new representation is generated for previous and incoming elements over time, with bidirectional feedback between entries at featural, phonemic, and lexical levels (McClelland & Elman, 1986). The Neighborhood Activation Model (Luce & Pisoni, 1998) conceptualizes the cohort of competitors in terms of phonological neighbors, which are words that differ by one phone from the target word. Similar to Cohort, incoming acoustic information activates "word decision units" (the neighborhood). As in TRACE, there is bi-directional feedback between higher-level information (e.g., context and frequency) and the lower level sensory input.

Embodied perspectives on language processing focus less on activation and competition and more on the role of perceptual (or sensorimotor) representations in, for example, (lexical) semantic access. Strong theories of embodiment view semantic knowledge as grounded in perceptual experience rather than in the relationships between words (Bickhard, 2008). Semantic access is thought to require perceptual simulation and directly engage areas of the brain that are active while perceiving the referent in the real world (Gallese & Lakoff, 2005). Weak-embodiment theories view lexical-semantic access as only moderately dependent on the participation of sensory and motor systems. On such models, semantics may be grounded in sensory and motor information but may also be accessed from higher-level representations (Meteyard & Vigliocco, 2008). In opposition to embodied-based accounts, abstract, symbolic theories view semantic knowledge as derived from correspondences between internal symbols and their extensions to objects in the real world (Fodor & Pylyshyn, 1988). New data could shed light on these theories and disambiguate among some of these predictions.

Recent empirical findings

Results from fMRI studies of visual perception and mental imagery suggest that the same occipital regions active while perceiving objects are similarly active while mentally ‘simulating’ visual images of objects (Ganis et al., 2004). Evidence that occipital (visual) regions are involved in simulating perceptual visual features during language comprehension also comes from recent fMRI experiments. One study showed, for example, that occipital regions processed shape information of sounds, wherein the stimulus impact sound of an object (such as a ball bouncing) modulated occipital activation when the hearer’s instructions were to name the shape (e.g., *round*) rather than the material (e.g., *rubber*) of the object generating the sound (James et al., 2011).

Results from Pulvermüller and colleagues’ EEG, fMRI, and MEG experiments support the hypothesis that (conceptual or lexical) semantic knowledge activation requires (or at least co-occurs with) perceptual simulation. Pulvermüller (2005) reported that action words involving the feet, hands, and face (e.g., *kick*, *pick*, *lick*) elicited cortical activation in motor regions associated with performing those actions with the respective body parts, by argument during early recognition stages. Similar results were reported for other novel sensory modalities during fMRI reading experiments, where scent-words such as *cinnamon* activated olfactory cortices (González et al., 2006) and taste-words such as *salt* activated gustatory cortices (Barrós-Loscertales et al., 2011). While Pulvermüller et al. (1996) argued that imageable nouns and verbs elicited the visual and motor cortices (respectively) in EEG, the results from a later fMRI reading experiment failed to indicate any effect of shape- and color-words such as *square* and *bronze* on activation in the visual cortex (Pulvermüller & Hauk, 2006).

Motivation of the current experiment

Based on such findings, we assume that the visual cortex is at least possibly active during spoken word recognition. Whether and when such activation contributes to meaning-based resolution remains controversial. In previous work, we found that the meaning-based resolution of *visual* words can be verified at around 300 ms post-stimulus onset (Simon, Lewis, & Marantz, 2012). This is reflected in the modulation of a superior temporal response (the MEG M350, comparable to the N400/N400m of Helenius et al., 2007) by the meaning-entropy (semantic ambiguity) of visually presented words. An absence of earlier semantic effects does *not* mean, however, that lexical resolution (selection of the appropriate representation) does not begin much earlier. But can one diagnose lexical resolution and perceptual simulation at earlier stages, and on which brain regions should one focus?

Previous studies of language processing have employed magnetoencephalography (MEG) combined with structural MRIs to examine the various stages of visual word recognition. The rationale is that MEG provides fine-grained temporal resolution (unlike fMRI) and, when enriched by source modeling that is constrained by structural MRI data, provides good spatial resolution (unlike EEG), which allows for examination on a millisecond level of the neural contributors to word recognition. Such work has determined that in *visual* word recognition, occipital brain regions process orthographic features at ~100 ms, inferior temporal regions decompose morphological properties at ~150 ms, and superior temporal regions contribute to lexical access (of the whole word form) at meaning based-resolution by as early as ~300 ms (Lewis, Solomyak, & Marantz, 2011; Simon, Lewis, & Marantz, 2012; Solomyak & Marantz, 2009; Solomyak & Marantz, 2010).

Of particular interest, in the functional anatomic sense, is the posterior middle temporal gyrus (pMTG), which previous work implicates as an indicator of lexical access in spoken word recognition (Hickok & Poeppel, 2007). While traditional accounts of verbal comprehension emphasize the role of Wernicke's (superior temporal) area in speech processing, there is considerable evidence that the MTG plays a central role in lexical processing (see reviews in, e.g., Dronkers et al. 2004; Hickok & Poeppel, 2007; Lau et al. 2008). Evidence that the pMTG is a critical node in the language comprehension network comes from lesion studies that find that, compared with patients with lesions in Wernicke's (superior temporal) and Broca's areas, patients with lesions to pMTG demonstrate poor performance in comprehending and naming single words. The pMTG may therefore link conceptual information to lexical representations (Dronkers et al., 2004). Further evidence comes from a study of connectivity profiles of brain areas within the language comprehension network, which determined that the MTG connectivity pattern is extensively integrated with areas of the network previously found to be critical to sentence comprehension (Turken & Dronkers, 2011). Results from neuroimaging show that MTG activation increases as a function of speech intelligibility (Davis & Johnsrude, 2003) and is also modulated by increasing semantic ambiguity (Rodd, Davis, & Johnsrude, 2005). The MTG therefore provides an ideal testing ground for the study of the processes leading to lexical access of spoken words.

Following spectrotemporal analysis of auditory input in the early auditory cortex and phonological analysis in the superior temporal gyrus (STG) and superior temporal sulcus (STS), the circuitry of the pMTG links phonological information to semantic representations. Based on this heuristic model, we may test predictions about embodiment by examining the temporal influence of continuous variables indexing a given word's "sensory information" on responses in occipital/visual brain areas, compared with the influence of acoustic, phonemic, and lexical variables on superior temporal and pMTG responses.

We can identify the stage of spectrotemporal analysis by examining the STG response to biphone frequency (BF), which is the frequency of occurrence of two adjacent sounds. The phonemic stage of recognition should be evident in STS responses to cohort competition (CC), which is based on cohort size as the summed frequency of all items beginning with the same two phonemes. The STS should also be sensitive to the cohort entropy (ENT) of words, which is the uncertainty of a word based on the number of words beginning with the same phonemes.

To identify lexical access, we examine responses of the pMTG to phonological neighborhood density (ND), which is based on the number of words that differ from the word by one phoneme (Luce & Pisoni, 1998). We employed this metric because words are thought to be organized based on their phonetic similarity to other words. As an example, the word *bat* has a dense neighborhood because it phonetically resembles many other words such as bate, ban, bit, etc. We recognize words within dense neighborhoods more slowly because high neighborhood density words activate more lexical representations than do low density items, which entails greater competition among entries (Vitevitch & Luce, 1998). We also examine responses to whole word form (surface) frequency (SF), which has previously been shown to modulate middle and superior temporal areas that are involved in lexical access during the later stages of word recognition (Lewis, Solomyak, & Marantz, 2011; Simon, Lewis, & Marantz, 2012).

How do we identify activation of visual representations? One prominent measure of sensory information is given by “imageability,” which indexes the extent to which a particular meaning can be perceptually experienced. The higher the imageability of a word, the easier it is to evoke a mental image of the associated meaning. Given this, imageability has been termed a measure of “semantic richness,” as it is an indicator of the number of perceptual features attached to the meaning of the word (Tyler et al., 2000). Various studies find that higher imageability leads to faster response times and stronger activation in pMTG (e.g., Zhuang et al., 2007). In a reversible lesion experiment, temporary disruption of the anterior temporal lobe (ATL) (via rTMS) led to comprehension difficulties for low- but not high-imageability words, suggesting that lower imageability word recognition depends more on language areas per se rather than occipital/visual areas (Pobric et al., 2009).

We also employ concreteness, which is another metric of the sensory information attached to the word. Concrete words (e.g., *apple*) are thought to be more easily encoded and retrieved than abstract words (e.g. *freedom*). For example, concrete words are recognized significantly faster and omitted significantly less from recall memory (Holmes & Langford, 1976). Concrete words additionally induce more negative N400s, which may be because concrete words evoke more sensory information attached to the representation (West & Holcomb, 2000). While concreteness and imageability strongly correlate, we include concreteness as a measure because imageability ratings are based only on visual aspects of the item (e.g., instructions require the rater to evoke a mental image of the item (Paivo et al., 1968).

The broad objective of this experiment is to thus test whether visual representations contribute to the lexical access of auditory words. The hypothesis, derived from Hebbian learning based neural accounts, is that semantic access requires perceptual simulation and directly engages the same areas of the brain that are active while perceiving the referent in the real world. MEG data are used to quantify the visual cortical, ST, and pMTG responses of subjects responding to lexical stimuli varying continuously in imageability (a quantification of the number of perceptual features attached to a word) as well as in acoustic, phonemic, and lexical properties. The spatial and temporal resolution provided by combining MEG and structural MRIs allows us to examine the localization and timing of word recognition stages. We asked whether the visual sensory information associated with lexical stimuli will modulate visual cortical activation simultaneously with or prior to modulation of the pMTG by lexical variables. An embodied account predicts that imageability will modulate occipital activation in parallel with or prior to surface frequency effects at the pMTG. A non-embodied account would predict effects of imageability in occipital regions only after semantic access has occurred.

Materials and methods

Participants

The study included 12 right-handed native English speakers (six males) from the New York University student population, with normal or corrected to normal vision. Two subjects were excluded from the source space analysis due to poor digitization data. The source space analysis therefore included 10 subjects and the behavioral analysis included 12 subjects.

Stimuli

We accessed all 1,324 monosyllabic nouns with imageability ratings from the MRC Psycholinguistic Database (Coltheart, 1981). The exclusion procedure removed items with the following characteristics:

- Multiple Part-of-Speech (POS) classes (e.g., *yawn* that is both a noun and a verb) based on the coding in the English Lexicon Project (ELP) (Balota et al., 2007).
- More than one morpheme, as coded by the ELP.
- Lexical decision accuracy below 70%, as coded by the ELP.
- Shared phonology with orthographically different word(s) (e.g., *cent* and *scent*), as based on homophony coding in the program Neighborhood Watch (NW). Because imageability rating tasks are based on visual words, it was necessary to exclude orthographically different items with identical phonologies to ensure that the subject accessed the correct meaning of the word.
- Heteronymy, where orthographically identical words have multiple meanings but different phonologies (e.g., *sow*, which refers to a female pig or the act of planting by seed), which was determined by accessing the number of dictionary headwords from the Wordsmyth Online Dictionary (Parks, Kennedy, & Broquist, 1998). While imageability ratings tasks make explicit the part of speech class an item belongs to, there are some instances where words have multiple meanings under the same speech class (e.g., the noun *yard* may denote a unit of measurement or an area of ground). We also excluded homonyms, where orthographically identical words have multiple meanings and the same phonology (e.g., *bank*, which might refer to a river bank or a financial institution), based on the number of dictionary headwords. We did not remove items where the alternative meaning was obscure, as in *mare*, which refers obscurely to "a large flat dark area on the moon or Mars
- Items outside the range of 3–4 phonemes (83% were within this range).
- Unusual consonant-vowel sequences.

The exclusion procedure reduced the set to 287 items. To increase the stimulus set, we included all 423 nouns not already in the MRC set from the Cortese & Fugett (2004) corpus, which includes imageability ratings for 3,000 monosyllabic words. After applying the same exclusion criteria to these items, we further winnowed down this new list to just 113 items (closely matched to items from the other corpus in terms of phonemic and orthographic frequency and length), for a total of 400 total target items. Imageability ratings in both corpora fall between 100 (lowest-imageability) and 700 (highest-imageability). For example, the imageability of *whim* = 180, *hag* = 400, and *goose* = 690. Our stimulus imageability ranged from 160–690.

The target stimulus list consisted of monosyllabic, monomorphemic, unambiguous, familiar nouns consisting of 3–4 phonemes. Nonwords were selected from the Online ARC Nonword Database (Rastle et al., 2002). The items were chosen to include only legal bigrams and to resemble the target stimuli in terms of bigram frequency, letter length, and phoneme length. Each of the 400 target stimuli was matched to a nonword of equal phoneme length, for a total of 800 items. All target stimuli are listed in the Appendix.

Variables

One way of isolating acoustic, phonemic, lexical, and visual effects is to employ the variables in a regression model. By regressing each variable onto the competing variables, one can ensure that any early cortical activation in visual cortices can not be attributed to phonemic, phonetic, or lexical effects. The variables are described below.

Biphone Frequency—This measures the segment-to-segment frequency of occurrence of two sounds as calculated by the program Neighborhood Watch.

Cohort Competition—This was the ratio of a word's CELEX frequency to its cohort-size, multiplied by 100, so that lower values indicate greater competition (as calculated in Zhuang et al., 2011)

Cohort Entropy—This is an estimate of uncertainty based on the number of words beginning with the same phonemes. The Penn Forced Aligner (Yuan & Liberman, 2008) was used to obtain a phoneme-by-phoneme measure of millisecond-by-millisecond neural activation.

Phonological neighborhood density—This measures the density of phonological neighbors that differ from the word by one phoneme (Luce & Pisoni, 1998). Measures were accessed from the English Lexicon Project (Balota et al., 2007).

Surface frequency—This was the log value of the written frequency of the whole word form in the CELEX corpus.

Imageability—The measure is based on ratings from the MRC online database and the Cortese and Fugett corpus.

Concreteness—Values are based on ratings from the MRC online. Ratings were only available for around half of the target items. Means and standard deviations of the target stimuli's properties are provided in Table 1.

Procedure and Recording

Words and nonwords were converted to synthetic speech files with Mac Text-to-Speech. The speech files were edited for pauses at the beginning and end, and verified for intelligibility. Stimuli were presented diotically at a loudness level of ~72 dB through foam insert earphones. The presentation script used Psychtoolbox helper scripts programmed in MATLAB (MathWorks, Inc., Natick, MA, USA). Subjects responded to individually presented speech stimuli (with 800 ms ITIs) in a lexical decision task with button presses over the course of 800 trials randomized over four blocks. Subjects lay supine with their eyes closed during the presentation blocks. MEG data were acquired continuously via a whole-head MEG system with 157 axial gradiometer sensors (Kanazawa Institute of Technology, Kanazawa, Japan) with the recording parameters of 1,000 Hz sampling rate, 60 HZ band pass filter 60 HZ, and DC high pass filter. Structural MRIs were separately acquired during a separate experiment at the Center for Brain Imaging at New York University (3T Siemens Allegra scanner with T1-weighted MPRAGE sequences).

Analysis

We followed the same procedure for MEG data processing for source space analyses described in Lewis, Solomyak, and Marantz (2011. Noise reduction with software MEG160 (Yokogawa Electric Corporation and Eagle Technology Corporation, Tokyo, Japan) and data from three MEG reference sensors involved the Continuously Adjusted Least-Squares Method (Calm; Adachi, Shimogawara, Higuchi, Haruta, & Ochiai, 2001). Further processing of the noise reduced data was in MNE (MGH/HMS/MIT Athinoula A. Martinos Center for Biomedical Imaging, Charleston, MA). We reconstructed each subject's structural MRI using FreeSurfer routines (CorTechs Lab Inc., La Jolla, CA). The reconstructions were used to estimate the cortically constrained minimum-norm solutions of the MEG data. The forward solution (magnetic field estimates at each MEG sensor) was estimated from a source space of 5124 activity points with a boundary-element model (BEM) method. The inverse solution, which is an estimate of the temporal and spatial distribution of the MEG data, was calculated from the forward solution. The data was then converted into dynamic statistical parameter map (dSPM) values (Dale, et al., 2000).

Regions of interest—We morphed each subject's cortex to a standard FreeSurfer brain to visualize grand average activation across subjects. We defined the occipital ROI around visible peak activation in the occipital region, whereas anatomical FreeSurfer labels, including the superior temporal gyrus (STG), superior temporal sulcus, (STS) and middle temporal gyrus (MTG), constrained the selection of ROIs based on peaks in the visible grand average activation. The ROIs are pictured in Figure 1.

MNE routines morphed labels back to individual subject brains, and grand average ROI activation within each subject's label was employed in trial-by-trial correlational analyses with the stimulus variables (including imageability, biphone frequency, cohort competition, neighborhood density, entropy, and surface frequency). We focused on the left hemisphere because neurophysiological evidence suggests that speech and language perception is lateralized here, however, we do acknowledge that this may depend on the technique, as hemodynamic and electrophysiological imaging data has indicated that processing may be more bilateral in nature (Price, 2012; Schirmer, Fox, & Grandjean, 2012; Turkeltaub & Coslett, 2010).

Exclusion Criteria: We applied the exclusion procedure from Lewis, Solomyak, and Marantz (2011) to data from target trials, beginning with removal of trials in which responses were incorrect and/or exceeded 5s. Normalization converted the rest of the trials into z-scores for each subject. We then excluded trials in which a subject's response time surpassed three standard deviations from that subject's overall mean. The exclusion procedure removed ~18% of the data.

Time course analyses—Our analysis examined effects of the stimulus variables on millisecond level neural activation as the speech played. Specifically, we correlated millisecond level activation within the ROIs with the various stimulus variables. A multiple comparisons correction (Maris & Oostenveld, 2007) was performed on temporal clusters of the point-by-point regressions that were significant prior to correction at the $p < .05$ significance level. An •*r* statistic was constructed by summing coefficients of temporally continuous effects. We tested the significance of the statistic with Monte-Carlo *p*-values. First, we computed a correlation wave by permuting the random variable 10,000 times and then calculated the •*r* statistic at significant clusters at each of the 10,000 permutations. A distribution of •*r* values was constructed from the highest •*r* value at the individual permutations. We defined our Monte-Carlo *p*-value based on the ratio of new values that were higher than the initial statistic.

Results

Behavioral results

Higher values of imageability significantly reduced response latencies ($p < .01$, $r = -0.046$) and increased response accuracies ($p < .01$, $r = 0.103$). Consistent with previous findings, we found that concreteness also facilitated response times ($p = .015$, $r = -0.0371$). Like Zhuang et al. (2011), we found that higher competition resulted in slower response times ($p < .01$, $r = -0.0681$). Again, note that lower values of cohort competition (the ratio of a word's frequency to the sum of its competitors') indicate greater competition. Zhuang et al. (2011) found that higher imageability sped up response times only when competition was high. We, however, found that that higher imageability significantly sped up response times regardless of competition level.

Neural data

We report here the significant findings. Each subject displayed the typical auditory M100 response. Contour maps and the grand average waveform for all subjects and trials of the raw MEG data at the M100 response are shown in Figure 2. Early occipital activation was primarily positive (outgoing from the cortex), while peak activation within the STG, STS, and MTG labels was negative (ingoing toward the cortex). Figure 2 also presents the labels along with the grand average time courses of activation.

Neurophysiological Timing Results

We investigated the millisecond-by-millisecond, trial-by-trial activation within each subject's ROI in a mixed effects model analysis with subjects and items as random factors. The variables were residual values from linear regressions that removed effects from other variables. Figure 3 displays the correlations and Table 2 provides a summary of the significant correlations.

Superior temporal gyrus—Activation significantly correlated with token biphone frequency over the 160–191 ms time window ($\sum r$ =1.4463 for 31 time points, p = .04 following correction for multiple comparisons (CMC) over the 1–200 ms window time window), and also over the 217–255 time window ($\sum r$ =1.7320 for 39 time points, p = .03 following CMC over the 200–500 ms window time window), with higher values of biphone frequency resulting in stronger activation.

Superior temporal sulcus—Activation significantly correlated with cohort competition from 255–276 ms ($\sum r$ =1.0677 for 21 time points, p = .02 following CMC over the 150–300ms time window), where higher competition had an inhibitory effect on activation (note that lower values denote higher competition). The linear mixed effects model analysis of entropy examined the effect of the millisecond entropy values on each millisecond of STS activation. The analysis identified a large cluster of significant t-values between ~250–280 ms post stimulus onset (significance threshold of t > 1.96, p < .05).

Posterior middle temporal gyrus—Phonological neighborhood density significantly modulated activation over the 327–347 ms time window, however, the correlation was just at the threshold of significance following CMC ($\sum r$ =.91 for 21 time points, p = .05). Activation significantly correlated with surface frequency between 415–442 ms ($\sum r$ =1.2875 for 28 time points, p = .04 following CMC over the 200–500ms time window), with higher values resulting in stronger activation.

Occipital—Activation significantly correlated with residual imageability over the 161–191 ms time window ($\sum r$ =1.2415 for 31 time points, p < .05 following CMC over the 100–300 ms window ms), with higher values resulting in stronger activation. We included additional variables in the model to rule out alternative plausible explanations for the effect on occipital activation. First, visual activation could signify contact with a visual word form rather than semantic content. To rule this out, we included the words' bigram frequencies from the English Lexicon Project (Balota et al., 2007). Second, the age of acquisition (AoA) of a concept may be earlier for more imageable words. Previous work shows an association between AoA and visual activation (e.g., Ellis et al., 2006). We acquired AoA ratings from the Cortese and Khanna corpus (2008), which lists ratings for over 30,000 monosyllabic words. Third, occipital activation could be attributed to low-level features that happen to correlate with imageability, such as duration or some other property of the sound. Sound symbolism in general is the idea that certain units of sound may share something in common. To investigate this, we coded the words by their phonetic descriptors (stop-plosive, fricative, nasal, affricative, glide, lateral, and rhotic).

We first ran individual correlations with the new variables and the imageability variable. Imageability was only significantly correlated with AoA, with higher imageability associated with lower AoA ($r = -.4149$, $p < .001$), which indicates that more imageable words are learned at an earlier age. However, a correlation with AoA and the *residual* imageability values was not significant ($p > .05$). We included the new variables in correlational analyses with occipital activation. None of the variables were found to significantly modulate activity following correction for multiple corrections ($p > .05$). Additionally, including the new variables in a regression model with imageability did not affect the significance of the correlation between imageability and occipital activation.

Discussion

This study focused on the temporal organization of the mapping from sound to meaning in lexical processing. We found that perceptual and lexical variables modulated different brain regions during different time windows. Importantly, and somewhat counter-intuitively, visual regions were maximally sensitive to imageability early on in speech processing, prior to effects of cohort competition and surface frequency, typical lexical-level effects. Token biphone frequency modulated STG activation from ~160–190 ms and from ~215–255 ms. While the effect at this region might be predicted, the direction of the effect is surprising as higher frequency was not predicted to result in stronger activation. During the very same time window, imageability modulated occipital activation. The direction of this effect has two interpretations: 1) imageable words have a *stronger* (single) visual representation, or, 2) imageable words have *more* visual representations. We base the latter explanation on the timing of the effect, which occurs prior to lexical access during the activation of multiple competitors. Given the temporal overlap of the imageability and biphone effects, we must assume that incoming sound automatically results in contact with low-level representations of the sound and associated visual properties.

As predicted, cohort competition modulated STS activation prior to lexical access (between ~255–275 ms). Based on the direction of the effect, we hypothesize that greater competition inhibits activation. We additionally found that higher entropy facilitated recognition, presumably because when uncertainty is high (e.g., more competitors), we devote less resources in accessing the representation (Ettinger, Linzen, & Marantz, 2014).

Surface frequency modulated pMTG activation between ~415–440 ms. The presence of lexical effects during only this later stage is supportive of “late access” models of lexical resolution. The direction of the effect (higher frequency yielded greater activation) is counter intuitive yet consistent with previous MEG results (e.g., Lewis, Solomyak, & Marantz, 2011; Simon, Lewis, & Marantz, 2012; Solomyak & Marantz, 2009). As stated earlier, we do acknowledge that failure to find earlier lexical effects does not necessarily mean that lexical resolution has not already begun. Figure 4 depicts the stages of spoken word recognition based on the results reported here.

Although the words were fairly short (mean of ~370 ms), it is unlikely that subjects reached the uniqueness point of the items by as early as 160 ms (where the effect of imageability emerged). Based on these findings, we must conclude that perceptual representations became activated before selection of the lexical entry. We hypothesize that the incoming sound simultaneously activates both segmental sound representations at the STG as well as the associated visual representations at occipital regions. Immediately prior to the recognition point of the word, the phonemic representations activate a cohort of competitors at the STS. Once the winning competitor has been selected, we then activate representations of the item's phonological family. Only after completing these processes do we activate and select the lexical entry. There is tension in these findings as to the extent of their ecological validity. Single spoken word recognition may operate at a much slower pace in the absence of contextual information (compared to, for instance, comprehension of a conversation).

While the imageability effect indicates early contact with visual representations, we cannot conclude from these data whether visual representations causally contribute to lexical access. Because we employed residual imageability values from a regression model including biphone frequency, surface frequency, onset phoneme frequency, and other variables, we can at least assume that early cortical activation in visual cortices cannot be attributed to phonemic, phonetic, or lexical effects.

Similar to Zhuang et al. (2011), we found that imageability correlated positively with activation early on in recognition, suggesting that words that are more "semantically rich" (where richness is simply a measure of featural distinctiveness) will activate more perceptual representations, as indicated by stronger activation for such words. Also consistent with Zhuang et al. (2011), we found that higher imageability led to faster and more accurate responses. Both Zhuang et al. and Tyler et al. (2000) found that higher imageability contributed to recognition only when items came from a large cohort of competitors, and therefore concluded that we more easily recognize lower competition words based on their phonemic properties rather than the semantic properties of their cohorts. We, however, found that imageability contributed to recognition across the board, with higher imageability resulting in faster responses to both high and low competition words. This pervasive effect (which partialled out confounding involvement of sublexical and lexical variables) at least indicates that early activation of perceptual representations plays *some* role in computing the meaning of a word, although the nature of this role remains to be understood.

Our findings provide support for claims about language processing that propose an automatic activation of perceptual representations during word recognition. Spoken word recognition implicates neuronal ensembles in brain regions responsible for processing visual and acoustic information. These areas become automatically activated simultaneously by acoustic attributes of the stimulus and by perceptual properties tied to the real world referent, and later by phonemic information culminating in the retrieval of the whole word form representation.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation Grant No. DGE-1342536 to GL and NIH 2R01DC05660 to DP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank Liina Pylkkänen for her advice during the design phase of the experiment and Alec Marantz for his considerable advice throughout the study. We also thank Tal Linzen for his help in computing the cohort entropy variable.

References

Baayen, R.; Piepenbrock, R.; Gulikers, L. The celex lexical database (release 2) [CD ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania; 1995.

Balota D, Yap M, Cortese M, Hutchison K, Kessler B, Loftis B, Neely J, Nelson D, Simpson G, Treiman R. The English lexicon project. Behavior Research Methods. 2007; 39:445. [PubMed: 17958156]

Barrós-Loscertales A, González J, Pulvermüller F, Ventura-Campos N, Bustamante J, Costumero V, Parcet M, Ávila C. Reading salt activates gustatory brain regions: fMRI evidence for semantic grounding in a novel sensory modality. Cerebral Cortex. 2011

Bickhard, M. Is embodiment necessary?. In: Calvo, P.; Gomila, T., editors. Handbook of cognitive science: An embodied approach. Amsterdam: Elsevier; 2008. p. 29-40.

Cortese M, Fugett A. Imagery ratings for 3,000 monosyllabic words. Behavior Research Methods. 2004; 3:384–387.

Cortese M, Khanna M. Age of acquisition ratings for 3,000 monosyllabic words. Behavior Research Methods. 2008; 40:791–794. [PubMed: 18697675]

Coltheart M. The MRC Psycholinguistic Database. Quarterly Journal of Experimental Psychology. 1981; 33A:497–505.

Dale A, Liu A, Fischl B, Buckner R, Belliveau J, Lewine J, Halgren E. Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron. 2000; 26:55–67. [PubMed: 10798392]

Davis M, Johnsrude I. Hierarchical processing in spoken language comprehension. Journal of Neuroscience. 2003; 23:2423–3431.

Dronkers N, Wilkins D, Van Valin R, Redfern B, Jaeger J. Lesion analysis of the brain areas involved in language comprehension. Cognition. 2004; 92:145–177. [PubMed: 15037129]

Fodor J, Pylyshyn Z. Connectionism and cognitive architecture: a critical analysis. Cognition. 1988; 28:3–71. [PubMed: 2450716]

Gallese V, Lakoff G. The brain's concepts: The role of the sensory-motor system in reason and language. Cognitive Neuropsychology. 2005; 22:455–479. [PubMed: 21038261]

Ganis G, Thompson W, Kosslyn S. Brain areas underlying visual mental imagery and visual perceptual. Cognitive Brain Research. 2004; 20:226–241. [PubMed: 15183394]

González J, Barrós-Loscertales A, Pulvermüller F, Meseguer V, Sanjúan A, et al. Reading cinnamon activates olfactory brain regions. Neuroimage. 2006; 32:906–912. [PubMed: 16651007]

Helenius P, Salmelin R, Service E, Connolly J, Leinonen S, Lyytinen H. Cortical activation during spoken-word segmentation in nonreading impaired and dyslexic adults. Journal of Neuroscience. 2002; 22:2936–2944. [PubMed: 11923458]

Hickok G, Poeppel D. The cortical organization of speech processing. Nature Reviews Neuroscience. 2007; 8:393–402.

Holmes V, Langford J. Comprehension and recall of abstract and concrete sentences. Journal of Verbal Learning and Verbal Behavior. 1976; 15:559–566.

James T, Stevenson R, Kim S, VanDerKlok R, James K. Shape from sound: Evidence for a shape operator in the lateral occipital cortex. Neuropsychologia. 2011; 49:1807–1815. [PubMed: 21397616]

Lau E, Phillips C, Poeppel D. A cortical network for semantics: (de)constructing the N400. Nature Reviews Neuroscience. 2008; 9:920–933.

Ettinger A, Linzen T, Marantz A. The role of morphology in phoneme prediction: Evidence from MEG. Brain and Language. 2014; 129:14–23. [PubMed: 24486600]

Luce P, Pisoni B. Recognizing spoken words: The neighborhood activation model. Ear and Hearing. 1998; 19:1–36. [PubMed: 9504270]

Lewis G, Solomyak O, Marantz A. The neural basis of obligatory decomposition of suffixed words. Brain & Language. 2011; 118:118–127. [PubMed: 21620455]

Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods. 2007; 164:177–190. [PubMed: 17517438]

Marslen-Wilson W. Functional parallelism in spoken word recognition. Cognition. 1987; 25:71–102. [PubMed: 3581730]

McClellend J, Elman J. The TRACE model of speech perception. Cognitive Psychology. 1986; 18:1–86. [PubMed: 3753912]

Meteyard, L.; Vigliocco, G. The role of sensory and motor information in semantic representation: A review. In: Calvo, P.; Gomila, A., editors. Handbook of cognitive science: An embodied approach. London, United Kingdom: Academic Press, Elsevier; 2008. p. 293-312.

Paivo A, Yuille J, Madigan S. Concreteness, imagery, and meaningfulness for 925 nouns. Journal of Experimental Psychology. 1968; 76:1–25.

Pobric G, Ralph M, Jefferies E. The role of the anterior temporal lobes in the comprehension of concrete and abstract words: rTMS evidence. Cortex. 2009; 45:1104–1110. [PubMed: 19303592]

Poeppel D, Idsardi W, Wassenhove V. Speech perception at the interface of neurobiology and linguistics. Philisophical Transactions of the Royal Society of London Biological Sciences. 2008; 363:1071–1086.

Price C. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage. 2012; 62:816–847. [PubMed: 22584224]

Pulvermüller F. Words in the brain’s language. Behavioral and Brain Sciences. 1999; 22:253–336. [PubMed: 11301524]

Pulvermüller F. Brain mechanisms linking language and action. Nature Reviews Neuroscience. 2005; 6:576–582.

Pulvermüller F, Hauk O. Category-specific conceptual processing of color and form in left fronto-temporal cortex. Cerebral Cortex. 2006; 8:1193–1201. [PubMed: 16251506]

Pulvermüller F, Preissl H, Lutzenberger W, Birbaumer N. Brain rhythms of language: Nouns versus verbs. European Journal of Neuroscience. 1996; 8:937–941. [PubMed: 8743741]

Pylkkänen L, Marantz A. Tracking the time course of word recognition with MEG. Trends in Cognitive Science. 2003; 7:187–189.

Rastle K, Harrington J, Coltheart M. 358,534 nonwords: The ARC Nonword Database. Quarterly Journal of Experimental Psychology. 2002; 55A:1339–1362. [PubMed: 12420998]

Rodd J, Davis H, Johnsrude I. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cerebral Cortex. 2005; 15:1261–1269. [PubMed: 15635062]

Schirmer A, Fox P, Grandjean D. On the spatial organization of sound processing in the human temporal lobe: a meta-analysis. Neuroimage. 2012; 63:137–147. [PubMed: 22732561]

Simon D, Lewis G, Marantz A. Disambiguating form and lexical frequency of MEG responses using homonyms. Language & Cognitive Processes. 2012; 27:275–287.

Solomyak O, Marantz A. Lexical access in early stages of visual word processing: A single-trial correlational MEG study of heteronym recognition. Brain and Language. 2009; 108:191–196. [PubMed: 19004492]

Solomyak O, Marantz A. MEG evidence for early morphological decomposition in visual word recognition: A single-trial correlational MEG study. Journal of Cognitive Neuroscience. 2010; 22:2042–2057. [PubMed: 19583463]

Turkeltaub P, Coslett H. Localization of sublexical speech perception components. Brain and Language. 2010; 114:1–15. [PubMed: 20413149]

Turken U, Dronkers N. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Frontiers in Systems Neuroscience. 2011; 5

Tyler L, Voice J, Moss H. The interaction of meaning and sound in spoken word recognition. Psychonomic Bulletin & Review. 2000; 7:320–326. [PubMed: 10909140]

Vitevitch M, Luce P. When words compete: Levels of processing in spoken word perception. Psychological Science. 1998; 9:325–329.

West W, Holcomb P. Imaginal, semantic, and surface-level processing of concrete and abstract words: An electrophysiological investigation. Journal of Cognitive Neuroscience. 2000; 12:1024–1037. [PubMed: 11177422]

Yuan J, Liberman M. Speaker identification on the SCOTUS corpus. Journal of the Acoustical Society of America. 2008; 123

Zhuang J, Randall B, Stamatakis E, Marslen-Wilson W, Tyler L. The interaction of lexical semantics and cohort competition in spoken word recognition: an fMRI study. Cognitive Neuroscience. 2011; 12:778–3790.

Appendix

Target stimuli (in decreasing order of imageability)

goose
pill
bulb
chimp
sled
jeep
fist
gym
lid
shelf
wasp
cheese
dice
pearl
wrist
car
girl
boat
beard
tub
kite
tooth
yacht
kilt
tongue
blood
chef
clerk
fern
nut
skull
broom
lawn
hill
house
keg
mink
roof
web
bun
dorm
smog
spine
thorn
door
pond
fork
hoof
hen
sheep
tomb
blouse
sheet
juice
noose
tent
stove
grape
rug
dime
stool
globe
mud
flute
vest
crutch
cub
grain
prom
gown
song
moth
snail
wheat
lung
shawl
disc
dusk
lamp
wife
beak
brick
desk
mug
scab
brain
gem
van
hemp
moss
snack
cone
fang
beast
mast
slime
dome
chest
dot
goal
shrub
bench
hive
knob
gift
town
ghost
fuzz
path
sleeve
sperm
cloth
dirt
sleet
nymph
ridge
cob
den
mound
niece
quill
yeast
thief
womb
hog
brass
pork
dude
gauze
reed
wench
badge
ranch
veal
tang
filth
groin
tribe
lint
pouch
wand
mile
fig
ledge
plaque
shrine
silk
rim
rice
mace
gap
marsh
thug
bib
pint
booth
nerve
news
life
brute
truce
breath
duke
grief
hunk
scope
smudge
twine
yam
height
crime
chive
grove
hearth
rump
self
latch
mosque
dean
grid
clove
pest
knoll
lust
germ
folk
jazz
josh
lobe
loft
slate
theft
health
zone
balm
cod
dill
finch
punk
slaw
gloom
year
growth
smut
blotch
chore
greed
math
rink
slab
sheath
luck
dell
length
theme
deed
noise
skit
tab
cult
debt
lymph
fraud
batch
bunt
clique
creed
drought
gene
grime
runt
scum
wrath
truth
drake
husk
musk
wad
noun
bile
zest
bliss
grail
fact
douche
dud
guild
sham
siege
thirst
whiz
wisp
welt
niche
shank
fad
pox
prude
choice
mead
pun
quirk
crude
norm
vogue
farce
stein
beck
sheen
bout

yarn
child
lip
cat
trout
church
golf
lake
mouse
bird
clock
mouth
peach
queen
shirt
crate
jug
monk
rod
barn
clown
morgue
rat
fruit
wig
porch
rib
wool
cage
glass
goat
king
blade
flesh
slush
horn
scar
tube
vine
vase
hat
birch
cheek
throat
hoop
hut
pub
rum
tweed
wart
wick
wreath
food
tool
tusk
brat
couch
graph
lad
gas
lice
crotch
fin
mutt
mob
snout
spud
stag
swine
valve
wealth
death
crumb
font
belt
chrome
gasp
gulf
lard
clan
hick
pal
gust
lair
broth
chunk
cove
dune
loin
malt
smock
snob
barb
month
thong
speech
steed
trough
haste
volt
slang
grub
jab
sloth
stooge
loss
threat
spite
chic
hag
hutch
noon
kin
nook
myth
thing
verb
crock
depth
faith
fame
zeal
curd
fright
rift
romp
realm
glade
pence
whim

Highlights

- We examine the temporal role of visual cortices in lexical access of speech
- We contrasted word imageability and word frequency effects on cortical activation
- Word imageability modulated visual cortices early in recognition
- Word frequency modulated posterior temporal gyrus activation later in recognition
- Sensory aspects of a lexical item are not a consequence of lexical activation

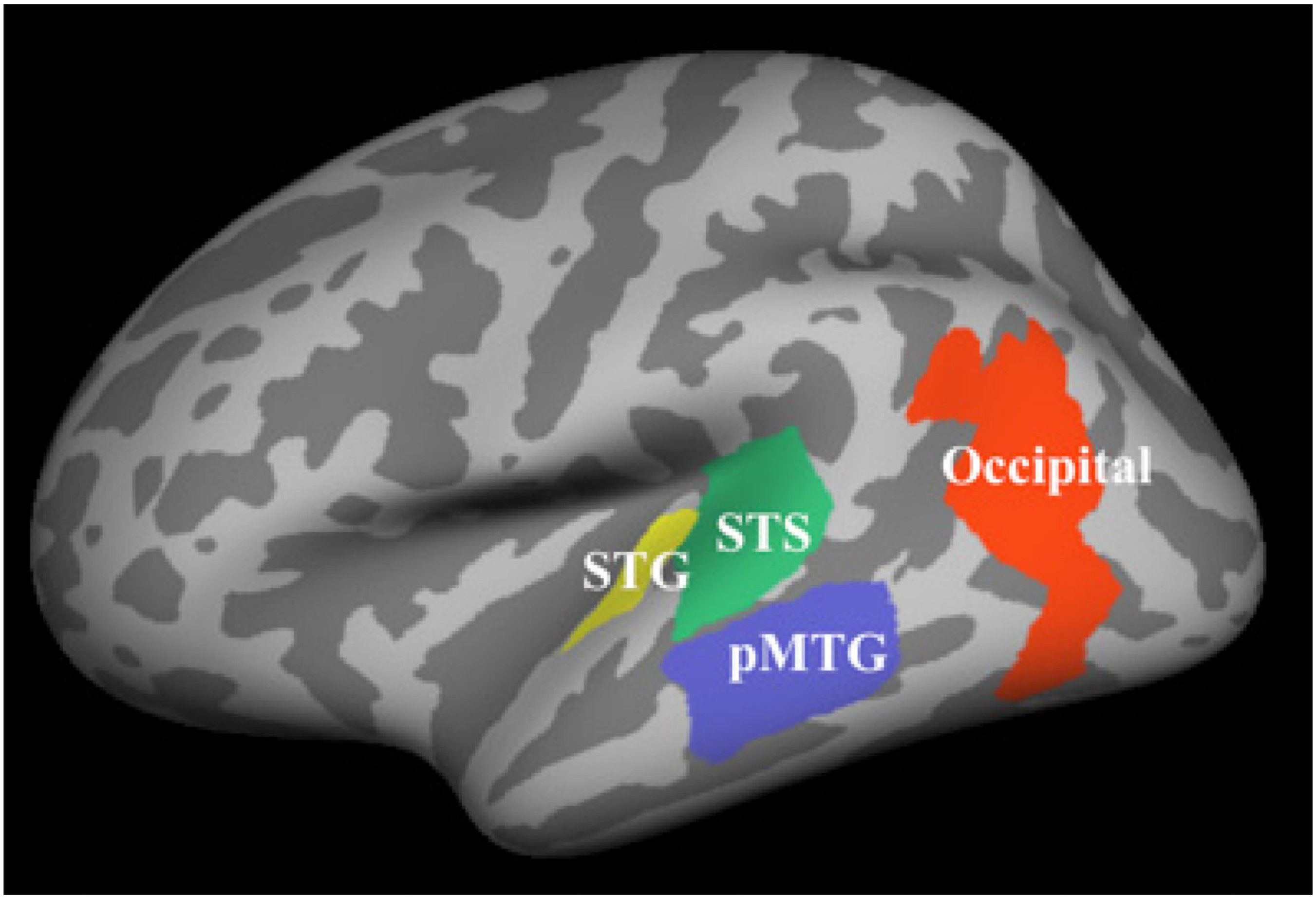

Figure 1.

The three functionally defined anatomically constrained ROIs presented on the inflated standard brain (STS = superior temporal sulcus, STG = superior temporal gyrus, pMTG = posterior middle temporal gyrus). Data within ROIs were employed in correlational analyses with stimulus variables. The ROIs include, roughly, BA areas 42, 22, 21, and 19 (STG, STS, pMTG, and Occipital, respectively).

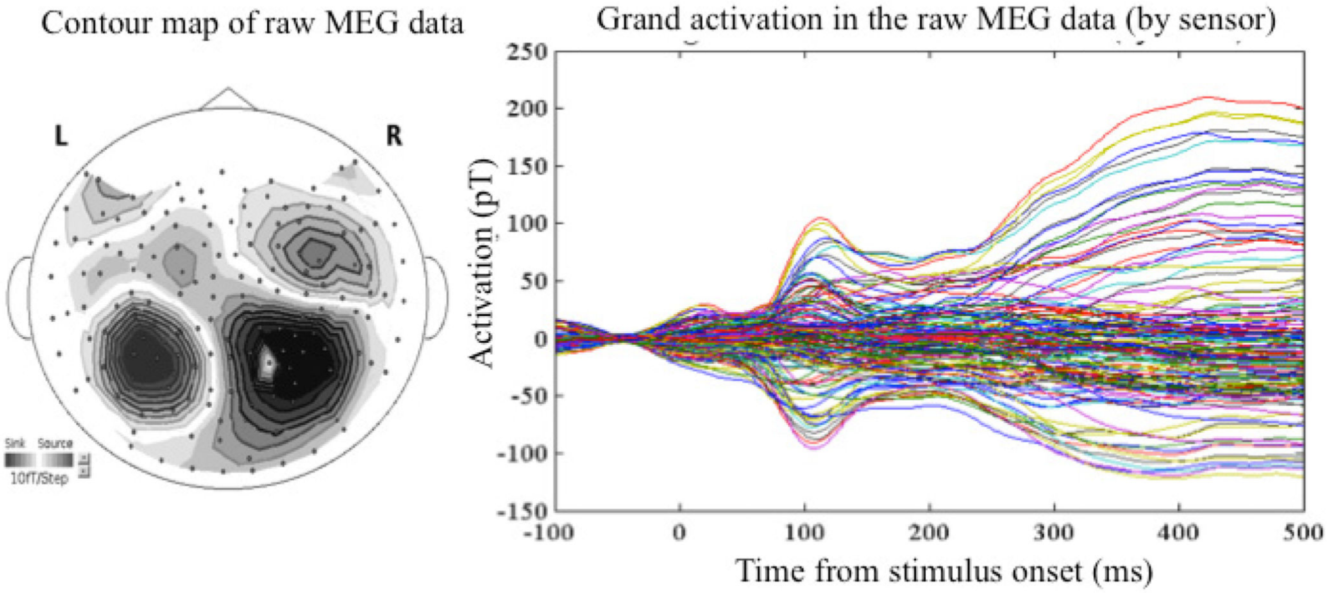

Figure 2.

Top left: Contour map of the grand average auditory M100 component; Top right: Grand average waveforms of the raw MEG data by sensor averaged over all trials and subjects; Center: Regions of interest (faint green blobs) and average activation displayed on the standard inflated brain in MNE. Bottom: Average time course of activation within ROIs in arbitrary dSPM units.

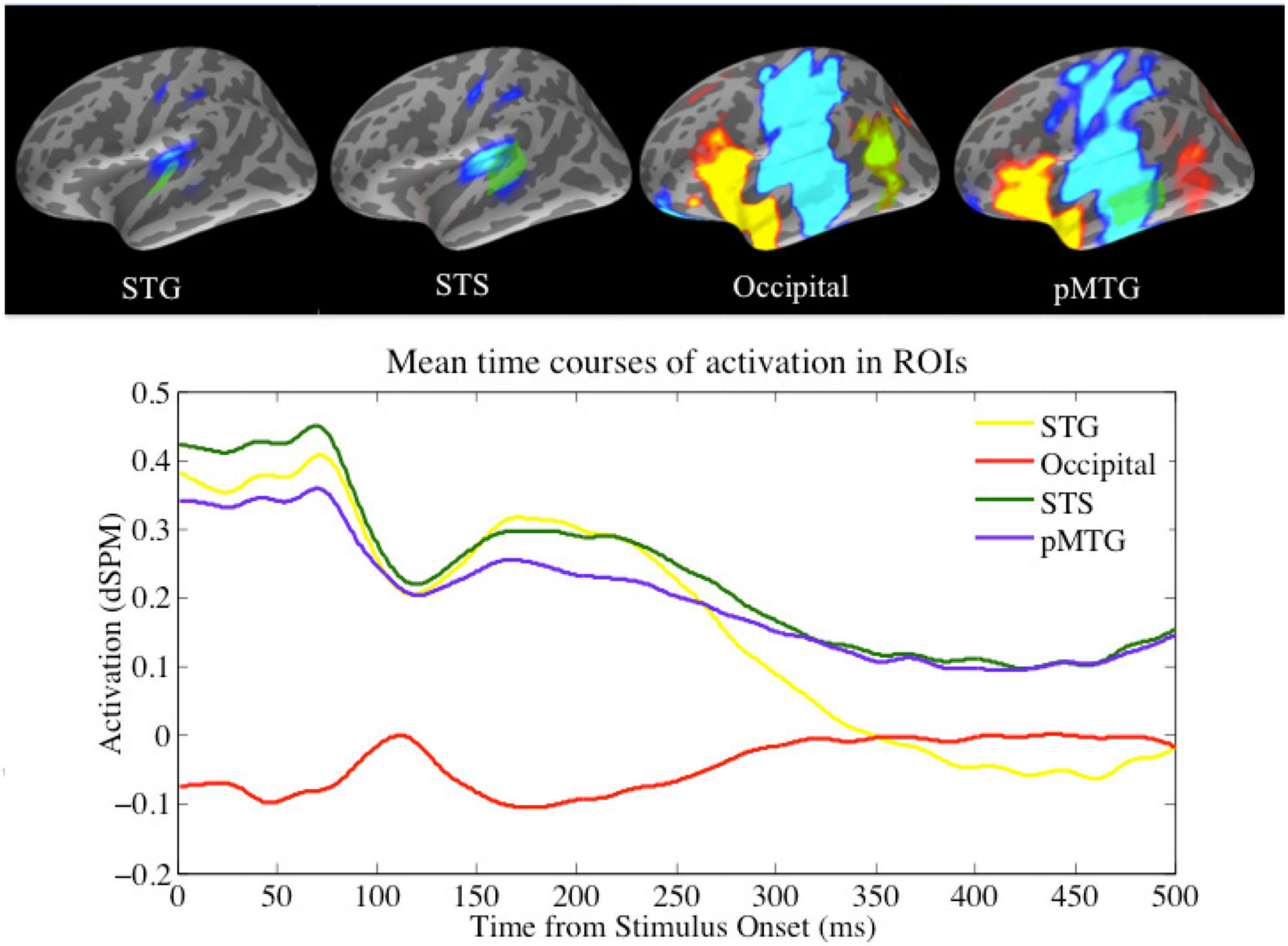

Figure 3.

Effects of stimulus variables on ROI activation. Correlations are plotted over time, with the *p* < .05 significance level (prior to CMC) indicated by the dotted line. Bold lines identify temporal clusters that survived the Monte-Carlo CMC. Note BF = biphone frequency, CC = cohort competition, IMG = imageability, SF = surface frequency, ND = neighborhood density, STG= superior temporal gyrus, STS = superior temporal sulcus, pMTG = posterior middle temporal gyrus.

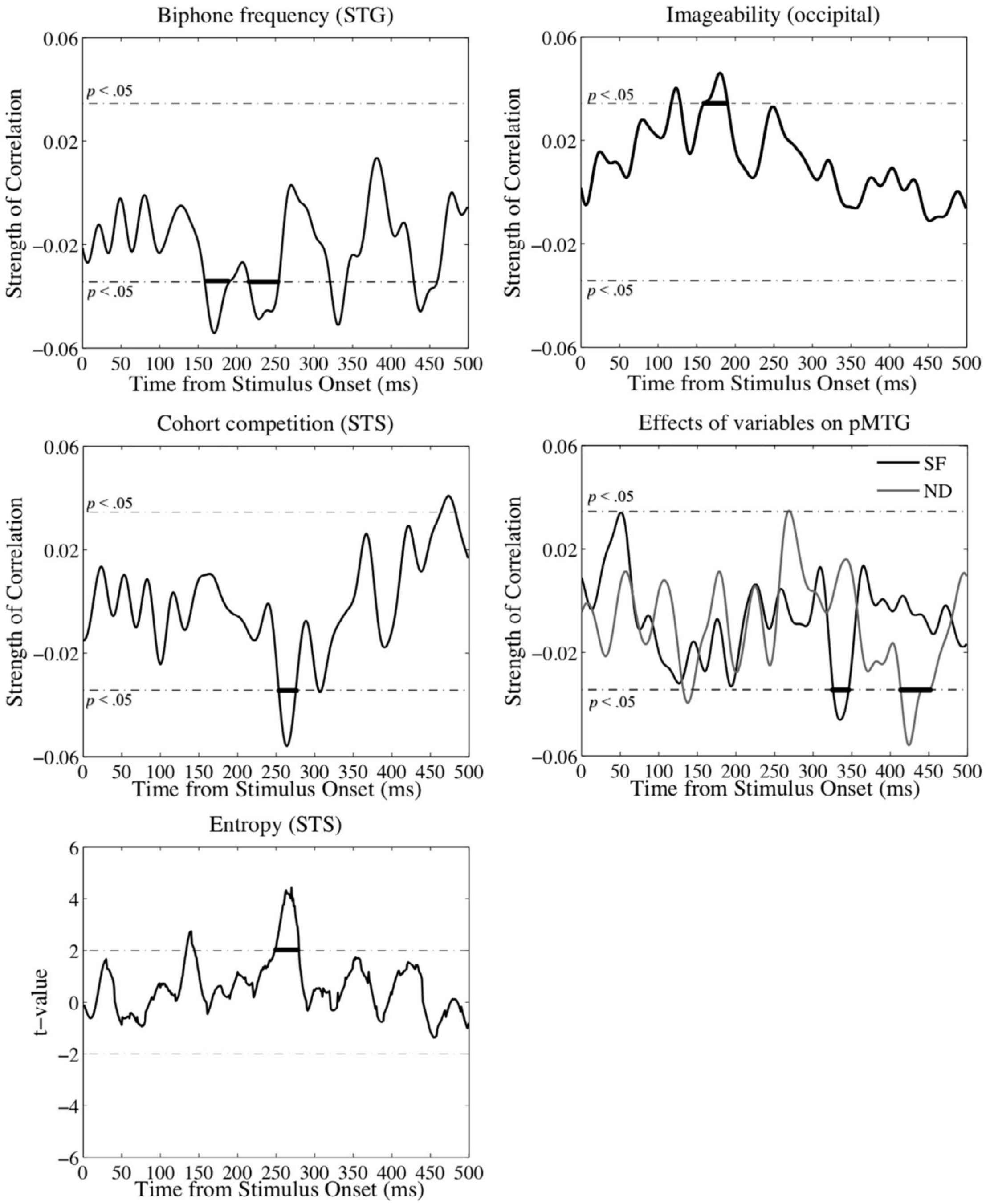

Figure 4.

Flowchart of sublexical and lexical stages of spoken word recognition, based on the results of the present study (BF = biphone frequency, IMG = imageability, ENT = entropy, CC = cohort competition, ND = neighborhood density, SF = surface frequency). In the model, the incoming speech waveform (bottom panel) activates segmental sound representations (middle panel) at STG and visual representations at visual regions. Before the recognition point (in this example, at ~350 ms), phonemes activate a cohort of competitors at STS. After selection of the representation, the item’s phonological family becomes activated at the pMTG. After these processes are complete, the lexical entry is selected at the pMTG.

Table 1

Properties for target items

Variable	Mean	SD
Biphone Frequency	2.50	0.66
Cohort Competition	6.85	18.68
Concreteness	527.28	95.04
Duration (ms)	371.82	48.37
Entropy	4.01	1.24
Imageability	493.71	107.76
Length	4.41	0.83
Number Phonemes	3.51	0.50
Phonological Density	13.76	8.56
Surface Frequency	7.97	1.61

Table 2

Significant correlations between ROI activation and stimulus variables

ROI	Variable	*p*	*r*	Time window (ms)
STG	Biphone Frequency	0.04	1.45	160–191
STG	Biphone Frequency	0.03	1.73	217–255
Occipital	Imageability	0.03	1.24	100–300
STS	Cohort Competition	0.02	1.06	255–276
pMTG	Neighborhood Density	0.05	0.91	327–347
pMTG	Surface Frequency	0.04	1.28	415–442

STG = superior temporal gyrus, STS = superior temporal sulcus, pMTG = posterior middle temporal gyrus.