Phoneme and word recognition in the auditory ventral stream

Iain DeWitt and Josef P. Rauschecker

Laboratory of Integrative Neuroscience and Cognition, Department of Neuroscience, Georgetown University Medical Center, Washington, DC 20007

Spoken word recognition requires complex, invariant representations. Using a meta-analytic approach incorporating more than 100 functional imaging experiments, we show that preference for complex sounds emerges in the human auditory ventral stream in a hierarchical fashion, consistent with nonhuman primate electrophysiology. Examining speech sounds, we show that activation associated with the processing of short-timescale patterns (i.e., phonemes) is consistently localized to left mid-superior temporal gyrus (STG), whereas activation associated with the integration of phonemes into temporally complex patterns (i.e., words) is consistently localized to left anterior STG. Further, we show left mid-temporal STG is reliably implicated in the invariant representation of phonetic forms and that this area also responds preferentially to phonetic sounds, above artificial control sounds or environmental sounds. Together, this shows increasing encoding specificity and invariance along the auditory ventral stream for temporally complex speech sounds.

functional MRI | meta-analysis | auditory cortex | object recognition | language

Spoken word recognition presents several challenges to the brain. Two key challenges are the assembly of complex auditory representations and the variability of natural speech (SI Appendix, Fig. S1) (1). Representation at the level of primary auditory cortex is precise: fine-grained in scale and local in spectrotemporal space (2, 3). The recognition of complex spectrotemporal forms, like words, in higher areas of auditory cortex requires the transformation of this granular representation into Gestalt-like, object-centered representations. In brief, local features must be bound together to form representations of complex spectrotemporal contours, which are themselves the constituents of auditory “objects” or complex sound patterns (4, 5). Next, representations must be generalized and abstracted. Coding in primary auditory cortex is sensitive even to minor physical transformations. Object-centered coding in higher areas, however, must be invariant (i.e., tolerant of natural stimulus variation) (6). For example, whereas the phonemic structure of a word is fixed, there is considerable variation in physical, spectrotemporal form—attributable to accent, pronunciation, body size, and the like—among utterances of a given word. It has been proposed for visual cortical processing that a feed-forward, hierarchical architecture (7) may be capable of simultaneously solving the problems of complexity and variability (8–12). Here, we examine these issues in the context of auditory cortex.

In a hierarchical pattern-recognition scheme (8), coding in the earliest cortical field would reflect the tuning and organization of primary auditory cortex (or core) (2, 3, 13). That is, single-neuron receptive fields (more precisely, frequency-response areas) would be tuned to particular center frequencies and would have minimal spectrotemporal complexity (i.e., a single excitation zone and one-to-two inhibitory side bands). Units in higher fields would be increasingly pattern selective and invariant to natural variation. Pattern selectivity and invariance respectively arise from neural computations similar in effect to “logical-AND” and “logical-OR” gates. In the auditory system, neurons whose tuning is combination sensitive (14–21) perform the logical-AND gate–like operation, conjoining structurally simple representations in lower-order units into the increasingly complex representations (i.e., multiple excitatory and inhibitory zones) of higher-order units. In the case of speech sounds, these neurons conjoin representations for adjacent speech formants or, at higher levels, adjacent phonemes. Although the mechanism by which combination sensitivity (CS) is directionally selective in the temporal domain is not fully understood, some propositions exist (22–26). As an empirical matter, direction selectivity is clearly present early in auditory cortex (19, 27). It is also observed to operate at time scales (50–250 ms) sufficient for phoneme concatenation, as long as 250 ms in the zebra finch (15) and 100 to 150 ms in macaque lateral belt (18). Logical-OR gate–like computation, technically proposed to be a soft maximum operation (28–30), is posited to be performed by spectrotemporal-pooling units. These units respond to supra-threshold stimulation from any member of their connected lower-order pool, thus creating a superposition of the connected lower-order representations and abstracting them. With respect to speech, this might involve the pooling of numerous, rigidly tuned representations of different exemplars of a given phoneme into an abstracted representation of the entire pool. Spatial pooling is well documented in visual cortex (7, 31, 32) and there is some evidence for its analog, spectrotemporal pooling, in auditory cortex (33–35), including the observation of complex cells when AI is developmentally reprogrammed as a surrogate V1 (36). However, a formal equivalence is yet to be demonstrated (37, 38).

Auditory cortex’s predominant processing pathways, ventral and dorsal (39, 40), appear to be optimized for pattern recognition and action planning, respectively (17, 18, 40–44). Speech-specific models generally concur (45–48), creating a wide consensus that word recognition is performed in the auditory ventral stream (refs. 42, 45, 47–50, but see refs. 51–53). The hierarchical model predicts an increase in neural receptive field size and complexity along the ventral stream. With respect to speech, there is a discontinuity in the processing demands associated with the recognition of elemental phonetic units (i.e., phonemes or something phone-like) and concatenated units (i.e., multi-segmental forms, both sublexical forms and word forms). Phoneme recognition requires sensitivity to the arrangement of constellations of spectrotemporal features (i.e., the presence and absence of energy at particular center frequencies and with particular temporal offsets). Word-form recognition requires sensitivity to the temporal arrangement of phonemes. Thus, phoneme recognition requires spectrotemporal CS and operates...
on low-level acoustic features (SI Appendix, Fig. S1B, second layer), whereas word-form recognition requires only temporal CS (i.e., concatenation of phonemes) and operates on higher-order features that may also be perceptual objects in their own right (SI Appendix, Fig. S1B, top layer). If word-form recognition is implemented hierarchically, we might expect this discontinuity in processing to be mirrored in cortical organization, with concatenative phonetic recognition occurring distal to elemental phonetic recognition.

Primate electrophysiology identifies CS as occurring as early as core’s supragranular layers and in lateral belt (16, 17, 19, 37). In the macaque, selectivity for communication calls—similar in spectrotemporal structure to phonemes or consonant-vowel (CV) syllables—is observed in belt area AL (54) and, to an even greater degree, in a more anterior field, RTP (55). Further, for macaques trained to discriminate human phonemes, categorical coding is present in the single-unit activity of AL neurons as well as in the population activity of area AL (1, 56). Human homologs to these sites putatively lie on or about the anterior-lateral aspect of Heschl’s gyrus and in the area immediately posterior to it (13, 57–59). Macaque PET imaging suggests there is also an evolutionary predisposition to left-hemisphere processing for communication call-specific communication calls (60). Consistent with macaque electrophysiology, human electrocorticography recordings from superior temporal gyrus (STG), in the region immediately posterior to the anterior-lateral aspect of Heschl’s gyrus (i.e., mid-STG), show the site to code for phoneme identity at the population level (61). Mid-STG is also the site of peak high-gamma activity in response to CV sounds (62–64). Similarly, human functional imaging studies suggest left mid-STG is involved in processing elemental speech sounds. For instance, in subtractive functional MRI (fMRI) comparisons, after partialing out variance attributable to acoustic factors, Leaver and Rauschecker (2010) showed selectivity in left mid-STG for CV speech sounds as opposed to other natural sounds (5). This implies the presence of a local density of neurons with receptive-field tuning optimized for the recognition of elemental phonetic sounds [i.e., areal specialization (AS)].

Furthermore, the region exhibits fMRI-adaptation phenomena consistent with invariant representation (IR) (65, 66). That is, response diminishes when the same phonetic content is repeatedly presented even though a physical attribute of the stimulus, one unrelated to phonetic content, is changed; here, the speaker’s voice (5). Similarly, using speech sound stimuli on the /ga — /da/ continuum and comparing response to exemplar pairs that varied only in acoustics or which varied both in acoustics and in phonetic content, Joannis and colleagues (2007) found adaptation specific to phonetic content in left mid-STG, again implying IR (67).

The site downstream of mid-STG, performing phonetic concatenation, should possess neurons that respond to late components of multisegmental sounds (i.e., latencies >60 ms). These units should also be selective for specific phoneme orderings. Nonhuman primate data for regions rostral to A1 confirm that latencies increase rostrally along the ventral stream (34, 55, 68, 69), with the median latency to peak response approaching 100 ms in area RT (34), consistent with the latencies required for phonetic concatenation. In a recent human electrophysiology study, Creutzfeldt and colleagues (1989) report vigorous single-unit responses to words and sentences in mid- to anterior STG (70). This included both feature-tuned units and late-component-tuned units. Although the relative location of feature and late-component units is not reported, and the late component units do not clearly evince temporal CS, the mixture of response types supports the supposition of temporal combination-sensitive units in human STG. Imaging studies localize processing of multisegmental forms to anterior STG/superior temporal sulcus (STS). This can be seen in peak activation to word-forms in electrocorticography (71) and magnetoencephalography (72). FMRI investigations of stimulus complexity, comparing activation to word-form and pure-tone stimuli, report similar localization (47, 73, 74).

Invariant tuning for word forms, as inferred from fMRI-adaptation studies, also localizes to anterior STG/STS (75–77). Studies investigating cross-modal repetition effects for auditory and visual stimuli confirm anterior STG/STS localization and, further, show it to be part of unimodal auditory cortex (78, 79). Finally, application of electrical cortical interference to anterior STG disrupts auditory comprehension, producing patient reports of speech as being like “a series of meaningless utterances” (80).

Here, we use a coordinate-based meta-analytic approach [activation likelihood estimation (ALE)] (81) to make an unbiased assessment of the robustness of functional-imaging evidence for the aforementioned speech-recognition model. In short, the method assesses the stereotaxic concordance of reported effects. First, we investigate the strength of evidence for the predicted anatomical dissociation between elemental phonetic recognition (mid-STG) and concatenative phonetic recognition (anterior STG). To assess this, two functional imaging paradigms are meta-analyzed: speech vs. acoustic-control sounds (a proxy for CS, as detailed later) and repetition suppression (RS). For each paradigm, separate analyses are performed for studies of elemental phonetic processing (i.e., phoneme- and CV-length stimuli) and for studies involving concatenative phonetic processing (i.e., word-length stimuli). Although the aforementioned model is principally concerned with word-from recognition, for comparative purposes, we meta-analyze studies of phrase-length stimuli as well. Second, we investigate the strength of evidence for the predicted ventral-stream colocalization of CS and IR phenomena. To assess this, the same paradigms are reanalyzed with two modifications: (i) for IR, a subset of RS studies meeting heightened criteria for fMRI-adaptation designs is included (Methods); (ii) to attain sufficient sample size, analyses are collapsed across stimulus lengths.

We also investigate the strength of evidence for AS, which has been suggested as an organizing principle in higher-order areas of the auditory ventral stream (5, 82–85) and is a well established organizing principle in the visual system’s analogous pattern recognition pathway (86–89). In the interest of comparing the organizational properties of the auditory ventral stream with those of the visual ventral stream, we assess the colocalization of CS with AS and IR phenomena. CS and IR are examined as described earlier. AS is examined by meta-analysis of speech vs. nonspeech natural-sound paradigms. At a deep level, both our AS and CS analyses putatively examine CS-dependent tuning for complex patterns of spectrotemporal energy. Acoustic-control sounds lack the spectrotemporal feature combinations requisite for driving combination-sensitive neurons tuned to speech sounds. For nonspeech natural sounds, the same is true, but there should also exist combination-sensitive neurons tuned to these stimuli, as they have been repeatedly encountered over development. For an effect to be observed in the AS analyses, not only must there be a population of combination-sensitive speech-tuned neurons, but these neurons must also cluster together such that a differential response is observable at the macroscopic scale of fMRI and PET.

Results

Phonetic-length-based analyses of CS studies (i.e., speech sounds vs. acoustic control sounds) were performed twice. In the first analyses, tonal control stimuli were excluded on grounds that they do not sufficiently match the spectrotemporal energy distribution of speech. That is, for a strict test of CS, we required acoustic control stimuli to model low-level properties of speech (i.e., contain spectrotemporal features coarsely similar to speech), not merely to drive primary and secondary auditory cortex. Under this preparation, spatial concordance was greatest in STG/STS across each phonetic length-based analysis (Table 1).
Within STG/STS, results were left-biased across peak ALE-statistic value, cluster volume, and the percentage of studies reporting foci within a given cluster, hereafter “cluster concordance.” The predicted differential localization for phoneme- and word-length processing was confirmed, with phoneme-length effects most strongly associated with left mid-STG and word-length effects with left anterior STG (Fig. 1 and SI Appendix, Fig. S2). Phrase-length studies showed a similar leftward processing bias. Further, peak processing for phrase-length stimuli localized to a site anterior and subjacent to that of word-length stimuli, suggesting a processing gradient for phonetic stimuli that progresses from mid-STG to anterior STG and then into STS.

Table 1. Results for phonetic length-based analyses

<table>
<thead>
<tr>
<th>Analysis/anatomy</th>
<th>BA</th>
<th>Cluster Concordance</th>
<th>Volume, mm³</th>
<th>Center of mass</th>
<th>Peak coordinates</th>
<th>Peak ALE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>y</td>
<td>z</td>
</tr>
<tr>
<td>CS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phoneme length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STG</td>
<td>42/22</td>
<td>0.93</td>
<td>3,624</td>
<td>−57</td>
<td>−25</td>
<td>1</td>
</tr>
<tr>
<td>Right STG/RT</td>
<td>42/22</td>
<td>0.21</td>
<td>512</td>
<td>56</td>
<td>−11</td>
<td>−2</td>
</tr>
<tr>
<td>Word length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STG</td>
<td>42/22</td>
<td>0.56</td>
<td>2,728</td>
<td>−57</td>
<td>−17</td>
<td>−1</td>
</tr>
<tr>
<td>Right STG</td>
<td>22</td>
<td>0.13</td>
<td>192</td>
<td>55</td>
<td>−17</td>
<td>0</td>
</tr>
<tr>
<td>Phrase length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STS</td>
<td>21</td>
<td>0.58</td>
<td>2,992</td>
<td>−56</td>
<td>−8</td>
<td>−8</td>
</tr>
<tr>
<td>Left STS</td>
<td>21</td>
<td>0.42</td>
<td>1,456</td>
<td>−52</td>
<td>7</td>
<td>−16</td>
</tr>
<tr>
<td>Right STS</td>
<td>21</td>
<td>0.32</td>
<td>2,264</td>
<td>54</td>
<td>−3</td>
<td>−9</td>
</tr>
<tr>
<td>Left STS</td>
<td>22</td>
<td>0.32</td>
<td>840</td>
<td>−54</td>
<td>−35</td>
<td>1</td>
</tr>
<tr>
<td>Left PreCG</td>
<td>6</td>
<td>0.32</td>
<td>664</td>
<td>−47</td>
<td>−7</td>
<td>47</td>
</tr>
<tr>
<td>Left IFG</td>
<td>47</td>
<td>0.21</td>
<td>456</td>
<td>−42</td>
<td>25</td>
<td>−12</td>
</tr>
<tr>
<td>Left IFG</td>
<td>44</td>
<td>0.16</td>
<td>200</td>
<td>−48</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>RS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phoneme length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STG</td>
<td>42/22</td>
<td>0.33</td>
<td>640</td>
<td>−58</td>
<td>−21</td>
<td>4</td>
</tr>
<tr>
<td>Word length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STG</td>
<td>42/22</td>
<td>0.50</td>
<td>1408</td>
<td>−56</td>
<td>−9</td>
<td>−3</td>
</tr>
<tr>
<td>Left STG</td>
<td>42/22</td>
<td>0.19</td>
<td>288</td>
<td>−58</td>
<td>−28</td>
<td>2</td>
</tr>
</tbody>
</table>

BA, Brodmann area; IFG, inferior frontal gyrus; PreCG, precentral gyrus; RT, rostrotemporal area.

Fig. 1. Foci meeting inclusion criteria for length-based CS analyses (A–C) and ALE-statistic maps for regions of significant concordance (D–F) ($p < 10^{-3}$, $k > 150$ cm³). Analyses show leftward bias and an anterior progression in peak effects with phoneme-length studies showing greatest concordance in left mid-STG (A and D; $n = 14$), word-length studies showing greatest concordance in left anterior STG (B and E; $n = 16$), and phrase-length analyses showing greatest concordance in left anterior STS (C and F; $n = 19$). Sample size is given with respect to the number of contrasts from independent experiments contributing to an analysis.
Although individual studies report foci for left frontal cortex in each of the length-based cohorts, only in the phrase-length analysis do focus densities reach statistical significance.

Second, to increase sample size and enable lexical status-based subanalyses, we included studies that used tonal control stimuli. Under this preparation the same overall pattern of results was observed with one exception: the addition of a pair of clusters in left ventral prefrontal cortex for the word-length analysis (SI Appendix, Fig. S3 and Table S1). Next, we further subdivided word-length analyses according to lexical status: real word or pseudoword. A divergent pattern of concordance was observed in left STG (Fig. 2 and SI Appendix, Fig. S4 and Table S1). Peak processing for real-word stimuli robustly localized to anterior STG. For pseudoword stimuli, a bimodal distribution was observed, peaking both in mid- and anterior STG and coextensive with the real-word cluster.

Third, to assess the robustness of the predicted STG stimulus-length processing gradient, length-based analyses were performed on foci from RS studies. For both phoneme- and word-length stimuli, concordant foci were observed to be strictly left-lateralized and exclusively within STG (Table 1). The predicted processing gradient was also observed. Peak concordance for phoneme-length stimuli was seen in mid-STG, whereas peak concordance for word-length stimuli was seen in anterior STG (Fig. 3 and SI Appendix, Fig. S5). For the word-length analysis, a secondary cluster was observed in mid-STG. This may reflect repetition effects concurrently observed for phoneme-level representation or, as the site is somewhat inferior to that of phoneme-length effects, it may be tentative evidence of a secondary processing pathway within the ventral stream (63, 90).

Fourth, to assess colocalization of CS, IR, and AS, we performed length-pooled analyses (Fig. 4, Table 2, and SI Appendix, Fig. S6). Robust CS effects were observed in STG/STS. Again, they were left-biased across peak ALE-statistic value, cluster volume, and cluster concordance. Significant concordance was also found in left frontal cortex. A single result was observed in the IR analysis, localizing to left mid- to anterior STG. This cluster was entirely coextensive with the primary left-STG CS cluster. Finally, analysis of AS foci found concordance in STG/STS. It was also left-biased in peak ALE-statistic value, cluster volume, and cluster concordance. Further, a left-lateralized ventral prefrontal result was observed. The principal left STG/STS cluster was coextensive with the region of overlap between the CS and IR analyses. Within superior temporal cortex, the AS analysis was also generally coextensive with the CS analysis. In left ventral prefrontal cortex, the AS and CS results were not coextensive but were nonetheless similarly localized. Fig. 5 shows exact regions of overlap across length-based and pooled analyses.

Discussion

Meta-analysis of speech processing shows a left-hemisphere optimization for speech and an anterior-directed processing gradient. Two unique findings are presented. First, dissociation is observed for the processing of phonemes, words, and phrases: elemental phonetic processing is most strongly associated with mid-STG; auditory word-form processing is most strongly associated with anterior STG, and phrasal processing is most strongly associated with anterior STS. Second, evidence for CS, IR, and AS colocalize in mid- to anterior STG. Each finding supports the presence of an anterior-directed ventral-stream pattern-recognition pathway. This is in agreement with Leaver and Rauschecker (2010), who tested colocalization of AS and IR in a single sample using phoneme-length stimuli (5). Recent meta-analyses that considered related themes affirm aspects of the present work. In a study that collapsed across phoneme and pseudoword processing, Turkeltaub and Coslett (2010) localized sublexical processing to mid-STG (91). This is consistent with our more specific localization of elemental phonetic processing. Samson and colleagues (2011), examining preferential tuning for speech over music, reported peak concordance in left anterior STG/STS (92), consistent with our more general area-specialization analysis. Finally, our results support Binder and colleagues’ (2000) anterior-directed, hierarchical account of word recognition (47) and Cohen and colleagues’ (2004) hypothesis of an auditory word-form area in left anterior STG (78).

Classically, auditory word-form recognition was thought to localize to posterior STG/STS (93). This perspective may have been biased by the spatial distribution of middle cerebral artery accidents. The artery’s diameter decreases along the Sylvian fissure, possibly increasing the prevalence of posterior infarcts. Current methods in aphasia research are better controlled and more precise. They implicate mid- and anterior temporal regions in speech comprehension, including anterior STG (94, 95). Although evidence for an anterior STG/STS localization of auditory word-form processing has been present in the functional imaging literature since inception (96–99), perspectives advancing this view have been controversial and the localization is still not uniformly accepted. We find strong agreement among word-
processing experiments, both within and across paradigms, each supporting relocation of auditory word-form recognition to anterior STG. Through consideration of phoneme- and phrasal-processing experiments, we show the identified anterior-STG word form-recognition site to be situated between sites robustly associated with phoneme and phrase processing. This comports with hierarchical processing and thereby further supports anterior-STG localization for auditory word-form recognition.

It is important to note that some authors define “posterior” STG to be posterior of the anterior-lateral aspect of Heschl’s gyrus or of the central sulcus. These definitions include the region we discuss as “mid-STG,” the area lateral of Heschl’s gyrus. We differentiate mid- from posterior STG on the basis of proximity to primary auditory cortex and the putative course of the ventral stream. As human core auditory fields lie along or about Heschl’s gyrus (13, 57–59, 100), the ventral streams’ course can be inferred to traverse portions of planum temporale. Specifically, the ventral stream is associated with macaque areas RTp and AL (54–56), which lie anterior to and lateral of A1 (13). As human A1 lies on or about the medial aspect of Heschl’s gyrus, with core running along its extent (57, 100), a processing cascade emanating from core areas, progressing both laterally, away from core itself, and anteriorly, away from A1, will necessarily traverse the anterior-lateral portion of planum temporale. Further, this implies mid-STG is the initial STG waypoint of the ventral stream.

Nominal issues aside, support for a posterior localization could be attributed to a constellation of effects pertaining to
aspects of speech or phonology that localize to posterior STG/STS (69), for instance: speech production (101–108), phonological/articulatory working memory (109, 110), reading (111–113) [putatively attributable to orthography-to-phonology translation (114–116)], and aspects of audiovisual language processing (117–122). Although these findings relate to aspects of speech and phonology, they do so in terms of multisensory processing and sensorimotor integration and are not the key paradigms indicated by computational theory for demonstrating the presence of pattern recognition networks (8–12, 123). Those paradigms (CS and adaptation), systematically meta-analyzed here, find anterior localization.

**Table 2. Results for aggregate analyses**

<table>
<thead>
<tr>
<th>Analysis/anatomy</th>
<th>BA</th>
<th>Cluster Concordance</th>
<th>Volume, mm³</th>
<th>Center of Mass (x, y, z)</th>
<th>Peak Coordinates (x, y, z)</th>
<th>Peak ALE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STG</td>
<td>42/22</td>
<td>0.82</td>
<td>11,944</td>
<td>–57 –19 –1</td>
<td>–58 –18 0</td>
<td>0.056</td>
</tr>
<tr>
<td>Right STG</td>
<td>42/22</td>
<td>0.47</td>
<td>6,624</td>
<td>55 –10 –3</td>
<td>56 –6 –6</td>
<td>0.045</td>
</tr>
<tr>
<td>Left STS</td>
<td>21</td>
<td>0.18</td>
<td>1,608</td>
<td>–51 8 –14</td>
<td>–50 8 –14</td>
<td>0.039</td>
</tr>
<tr>
<td>Left PreCG</td>
<td>6</td>
<td>0.12</td>
<td>736</td>
<td>–47 –7 48</td>
<td>–48 –8 48</td>
<td>0.031</td>
</tr>
<tr>
<td>Left IFG</td>
<td>44</td>
<td>0.10</td>
<td>744</td>
<td>–45 12 21</td>
<td>–46 12 20</td>
<td>0.025</td>
</tr>
<tr>
<td>Left IFG</td>
<td>47</td>
<td>0.08</td>
<td>240</td>
<td>–42 25 –12</td>
<td>–42 24 –12</td>
<td>0.022</td>
</tr>
<tr>
<td>Left IFG</td>
<td>45</td>
<td>0.04</td>
<td>200</td>
<td>–50 21 12</td>
<td>–50 22 12</td>
<td>0.020</td>
</tr>
<tr>
<td><strong>IR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STG</td>
<td>22/21</td>
<td>0.45</td>
<td>1,200</td>
<td>–58 –16 –1</td>
<td>–56 –14 –2</td>
<td>0.020</td>
</tr>
<tr>
<td><strong>AS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left STG</td>
<td>42/22</td>
<td>0.87</td>
<td>3,976</td>
<td>–58 –22 2</td>
<td>–58 –24 2</td>
<td>0.031</td>
</tr>
<tr>
<td>Right STG</td>
<td>42/22</td>
<td>0.53</td>
<td>2,032</td>
<td>51 –23 2</td>
<td>54 –16 0</td>
<td>0.026</td>
</tr>
<tr>
<td>Left IFG</td>
<td>47/45</td>
<td>0.13</td>
<td>368</td>
<td>–45 17 3</td>
<td>–44 18 2</td>
<td>0.018</td>
</tr>
</tbody>
</table>

*Broader inclusion criteria for the IR analysis (SI Appendix, Table S3) yield equivalent results with the following qualifications: cluster volume 1,008 mm³ and cluster concordance 0.33.

---

**Fig. 5.** Flat-map presentation of ALE cluster overlap for (A) the CS analyses shown in Fig. 1, (B) the word-length lexical status analyses shown in Fig. 2, (C) the RS analyses shown in Fig. 3, and (D) the length-pooled analyses shown in Fig. 4. For orientation, prominent landmarks are shown on the left hemisphere of A, including the circular sulcus (CirS), central sulcus (CS), STG, and STS.
The segregation of phoneme and word-form processing along STG implies a growing encoding specificity for complex phonetic forms by higher-order ventral-stream areas. More specifically, it suggests the presence of a hierarchical network performing phonetic concatenation at a site anatomically distinct from and downstream of the site performing elemental phonetic recognition. Alternatively, the phonetic-length effect could be attributed to semantic confound; semantic content increases from phonemes to word forms. In an elegant experiment, Thierry and colleagues (2003) report evidence against this (82). After controlling for acoustics, they show that left anterior STG responds more to speech than to semantically matched environmental sounds. Similarly, Belin and colleagues (2000, 2002), after controlling for acoustics, show that left anterior STG is not merely responding to the vocal quality of phonetic sounds; rather, it responds preferentially to the phonetic quality of vocal sounds (83, 84).

Additional comment on the localization and laterality of auditory word and pseudoword processing, as well as on processing gradients in STG/STS, is provided in SI Appendix, Discussion.

The auditory ventral stream is proposed to use CS to conjoin lower-order representations and thereby to synthesize complex representations. As the tuning of higher-order combination-sensitive units is contingent upon sensory experience (124, 125), phrases and sentences would not generally be processed as Gestalt-like objects. Although we have analyzed studies involving phrase- and sentence-level processing, their inclusion is for context and because word-form recognition is a constituent part of sentence processing. In some instances, however, phrases are processed as objects (126). This status is occasionally recognized in orthography (e.g., “nonetheless”). Such phrases ought to be recognized by the ventral-stream network. This, however, would be the exception, not the rule. Hypothetically, the opposite may also occur: a word form’s length might exceed the network’s integrative capacity (e.g., “antidisestablishmentarianism”). We speculate the network is capable of concatenating sequences of at least five to eight phonemes: five to six phonemes is the modal length of English word forms and seven- to eight-phoneme-long word forms comprise nearly one fourth of English words (SI Appendix, Fig. S7 and Discussion). This estimate is also consistent with the time constant of echoic memory (~2 s). (Notably, there is a similar issue concerning the processing of text in the visual system’s ventral stream, where, for longer words, fovea-width representations must be “temporally” conjoined across microsaccades.) Although some phrases may be recognized in the word-form recognition network, the majority of STS activation associated with phrase-length stimuli (Fig. 1F) is likely related to aspects of syntax and semantics. This observation enables us to subdivide the intelligibility network, broadly defined by Scott and colleagues (2000) (127). The first two stages involve elemental and concatenative phonetic recognition, extending from mid-STG to anterior STG and, possibly, into subjacent STS. Higher-order syntactic and semantic processing is conducted throughout STS and continues into prefrontal cortex (128–133).

A qualification to the propositions advanced here for word-form recognition is that this account pertains to perceptually fluent speech recognition (e.g., native language conversational discourse). Both left ventral and dorsal networks likely mediate nonfluent speech recognition (e.g., when processing neologisms or recently acquired words in a second language). Whereas ventral networks are implicated in pattern recognition, dorsal networks are implicated in forward- and inverse-model computation (42, 44), including sensorimotor integration (42, 45, 48, 134). This supports a role for left dorsal networks in mapping auditory representations onto the somatomotor frame of reference (135–139), yielding articulator-encoded speech. This ventral–dorsal dissociation is illustrated in an experiment by Buchsbaum and colleagues (2005) (110). Using a verbal working memory task, they demonstrated the time course of left anterior STG/STS activation to be consistent with strictly auditory encoding: activation was locked to auditory stimulation and it was not sustained throughout the late phase of item rehearsal. In contrast, they observed the activation time course in the dorsal stream to be modality independent and to coincide with late-phase rehearsal (i.e., it was associated with verbal rehearsal independent of input modality, auditory or visual). Importantly, late-phase rehearsal can be demonstrated behaviorally, by articulatory suppression, to be mediated by subvocalization (i.e., articulatory rehearsal in the phonological loop) (140).

There are some notable differences between auditory and visual word recognition. Spoken language was intensely selected for during evolution (141), whereas reading is a recent cultural innovation (111). The age of acquisition of phoneme representation is in the first year of life (124), whereas it is typically in the third year for letters. A similar developmental lag is present with respect to acquisition of the visual lexicon. Differences aside, word recognition in each modality requires similar processing, including the concatenation of elemental forms, phonemes or letters, into sublexical forms and word forms. If the analogy between auditory and visual ventral streams is correct, our results predict a similar anatomical dissociation for elemental and concatenative representation in the visual ventral stream. This prediction is also made by models of text processing (10). Alternatively, we are aware of studies that have interpreted letter and word recognition in a single sample, for support of the dissociation is present in the literature. The visual word-form area, the putative site of visual word-form recognition (142), is located in the left fusiform gyrus of inferior temporal cortex (IT) (143). Consistent with expectation, the average site of peak activation to single letters in IT (144–150) is more proximal to V1, by approximately 13 mm. A similar anatomical dissociation can be seen in paradigms probing IR. Ordinarily, nonhuman primate IT neurons exhibit a degree of mirror-symmetric invariant tuning (151). Letter recognition, however, requires nonmirror IR (e.g., to distinguish “b” from “d”). When assessing identity-specific RS (i.e., repetition effects specific to non–mirror-inverted repetitions), letter and word effects differentially localize: effects for word stimuli localize to the visual word-form area (152), whereas effects for single-letter stimuli localize to the lateral occipital complex (153), a site closer to V1. Thus, the anatomical dissociation observed in auditory cortex for phonemes and words appears to reflect a general hierarchical processing architecture also present in other sensory cortices.

In conclusion, our analyses show the human functional imaging literature to support a hierarchical model of object recognition in auditory cortex, consistent with nonhuman primate electrophysiology. Specifically, our results support a left-biased, two-stage model of auditory word-form recognition with analysis of phonemes occurring in mid-STG and word recognition occurring in anterior STG. A third stage extends the model to phrase-level processing in STS. Mechanistically, left mid- to anterior STG exhibits core qualities of a pattern recognition network, including CS, IR, and AS.

Methods
To identify prospective studies for inclusion, a systematic search of the PubMed database was performed for variations of the query, “(phonetics OR ‘speech sounds’ OR phoneme OR ‘auditory word’ AND (MRI OR fMRI OR PET)).” This yielded more than 550 records (as of February 2011). These studies were screened for compliance with formal inclusion criteria: (i) the publication of stereotaxic coordinates for group-wise fMRI or PET results in a peer-reviewed journal and (ii) report of a contrast of interest (as detailed later). Exclusion criteria were the use of pediatric or clinical samples. Inclusion exclusion criteria admitted 115 studies. For studies reporting multiple suitable contrasts per sample, to avoid sampling bias, a single contrast was selected. For CS analyses, contrasts of interest compared activation to speech stimuli (i.e., phonemes/syllables, words/pseudowords, and phrases/sentences/
pseudoword sentences) with activation to matched, nonnaturalistic acoustic control stimuli (i.e., various tonal, noise, and complex artificial nonspeech stimuli). Analysis of BSA eligible contrasts were identified, representing 1,211 subjects and 541 foci. For RS analyses, contrasts compared activation to repeated and nonrepeated speech stimuli. A total of 31 eligible contrasts were identified, representing 471 subjects and 145 foci. For IR analyses, a subset of the RS cohort was selected that used designs in which “repeated” stimuli also varied acoustically but not phonetically (e.g., two different utterances of the same word). The RS cohort was used for phonetic length-based analyses as the more restrictive criteria for IR yielded insufficient sample sizes (as detailed later). For AS analyses, contrasts compared activation to speech stimuli and to other naturalistic stimuli (e.g., animal calls, music, tool sounds). A total of 17 eligible contrasts were identified, representing 239 subjects and 100 foci. All retained contrasts were binned for phonetic length-based analyses according to the estimated mean number of phonemes (i) “phoneme length,” one or two phonemes, (ii) “word length,” three to 10 phonemes, and (iii) “phrase length,” more than 10 phonemes. SI Appendix, Tables S2–S4, identify the contrasts included in each analysis.

The minimum sample size for meta-analyses was 10 independent contrasts. Foci reported in Montreal Neurological Institute coordinates were transformed into Talairach coordinates according to the ICBM2TAL transformation (154). Foci concordance was assessed by the method of ALE (81) in a random-effects implementation (155) that controls for within-experiment effects (156). Under the ALE, foci are treated as Gaussian probability distributions, which reflect localisation uncertainty. Pooled Gaussian focus maps were tested against a null distribution reflecting a random spatial association between different experiments. Correction for multiple comparisons was obtained through estimation of false discovery rate (157). Two significance criteria were used: minimum p value was set at 10−3 and minimum cluster extent was set at 150 mm². Analyses were conducted in GINGERALE (Research Imaging Institute, AFNI (National Institute of Mental Health), and MATLAB (Mathworks)). For visualization, CARET (Washington University in St. Louis) was used to project foci and ALE clusters from volumetric space onto the cortical surface of the Population-Average, Landmark- and Surface-based atlas (158). Readers should note that this procedure can introduce slight localization artifacts (e.g., projection may distribute one volumetric cluster discontinuously over two adjacent gyri).

ACKNOWLEDGMENTS. We thank Max Riesenhuber, Marc Ettlinger, and two anonymous reviewers for comments helpful to the development of this manuscript. This work was supported by National Science Foundation Grants BCS-0519127 and OISE-0730255 (to J.P.R.) and National Institute on Deafness and Other Communication Disorders Grant 1R1C1DC010720 (to J.P.R.).
DeWitt and Rauschecker


