Phoneme and word recognition in the auditory ventral stream

Iain DeWitt[1] and Josef P. Rauschecker[1]

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

Edited by Mortimer Mishkin, National Institute for Mental Health, Bethesda, MD, and approved December 19, 2011 (received for review August 17, 2011)

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- to anterior 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 ideas 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 excitatory 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 suprathreshold 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 A1 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., multisegmental 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

Author contributions: I.D. designed research; I.D. performed research; I.D. analyzed data; and I.D. and J.P.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: id32@georgetown.edu or rauschej@georgetown.edu.

See Author Summary on page 2709 (volume 109, number 8).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1113427109/-/DCSupplemental.

on low-level acoustic features (*SI Appendix*, Fig. S1*B*, 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. S1*B*, 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 conspecific 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, Joanisse 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 rare 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 AS phenomena with CS 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).

Table 1. Results for phonetic length-based analyses

Analysis/anatomy	BA	Cluster Concordance	Volume, mm^3	Center of mass			Peak coordinates			Peak ALE
				x	*y*	*z*	*x*	*y*	*z*	
CS										
Phoneme length										
Left STG	42/22	0.93	3,624	−57	−25	1	−58	−20	2	0.028
Right STG/RT	42/22	0.21	512	56	−11	−2	54	−2	2	0.015
Word length										
Left STG	42/22	0.56	2,728	−57	−17	−1	−56	−16	−2	0.021
Right STG	22	0.13	192	55	−17	0	56	−16	0	0.014
Phrase length										
Left STS	21	0.58	2,992	−56	−8	−8	−56	−8	−8	0.038
Left STS	21	0.42	1,456	−52	7	−16	−52	8	−16	0.035
Right STS	21	0.32	2,264	54	−3	−9	56	−6	−6	0.032
Left STS	22	0.32	840	−54	−35	1	−54	−34	0	0.028
Left PreCG	6	0.32	664	−47	−7	47	−48	−8	48	0.025
Left IFG	47	0.21	456	−42	25	−12	−42	24	−12	0.021
Left IFG	44	0.16	200	−48	11	20	−48	10	20	0.020
RS										
Phoneme length										
Left STG	42/22	0.33	640	−58	−21	4	−58	−20	4	0.018
Word length										
Left STG	42/22	0.50	1408	−56	−9	−3	−56	−10	−4	0.027
Left STG	42/22	0.19	288	−58	−28	2	−58	−28	2	0.017

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

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.

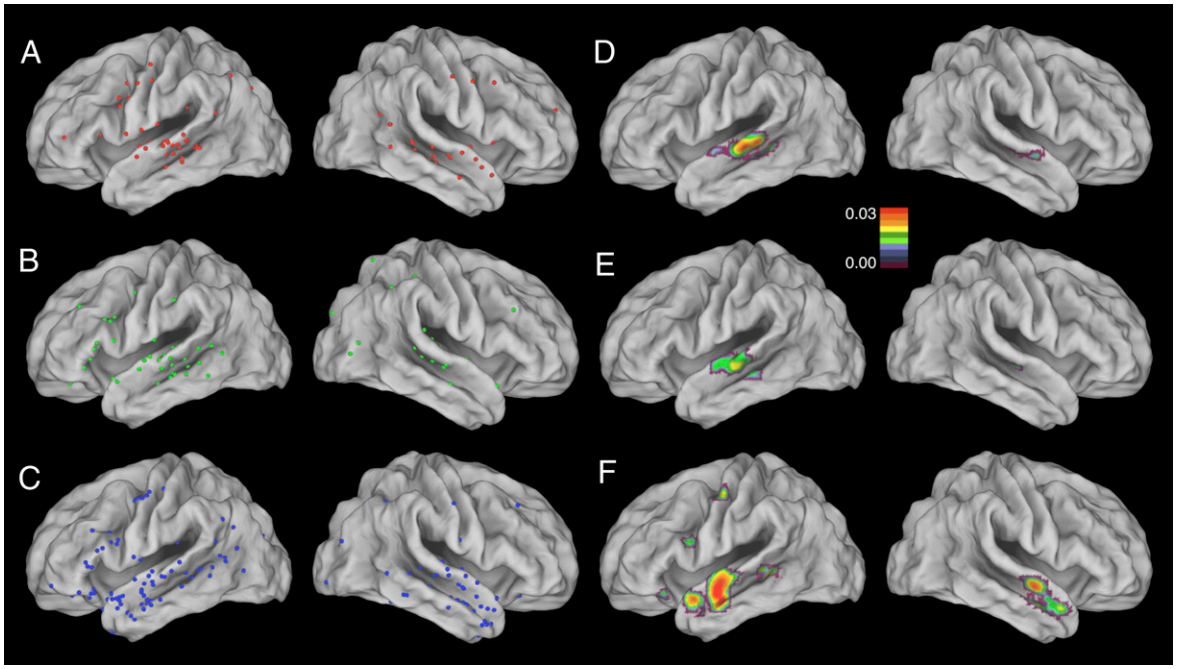

Fig. 1. Foci meeting inclusion criteria for length-based CS analyses (*A*–*C*) and ALE-statistic maps for regions of significant concordance (*D*–*F*) ($p < 10^{-3}$, $k > 150\ cm^3$). Analyses show leftward bias and an anterior progression in peak effects with phoneme-length studies showing greatest 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 studies 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, report peak concordance in left anterior STG/STS (92), consistent with our more general areal-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-

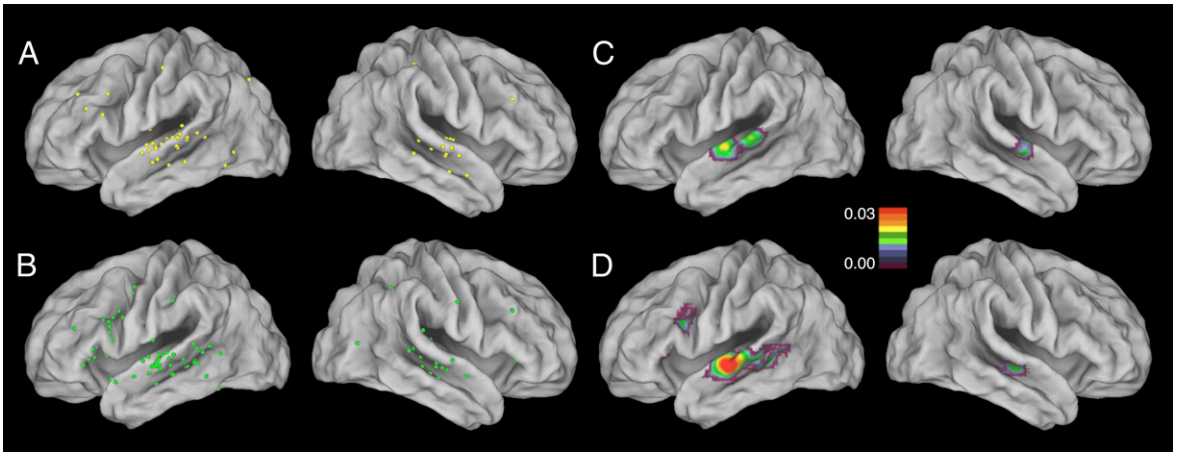

Fig. 2. Foci meeting liberal inclusion criteria for lexically based word-length CS analyses (*A* and *B*) and ALE-statistic maps for regions of significant concordance (*C* and *D*) ($p < 10^{-3}$, $k > 150$ cm^3). Similar to the CS analyses in Fig. 1, a leftward bias and an anterior progression in peak effects are shown. Pseudoword studies show greatest concordance in left mid- to anterior STG (*A* and *C*; $n = 13$). Notably, the distribution of concordance effects is bimodal, peaking both in mid- (−60, −26, 6) and anterior (−56, −10, 2) STG. Real-word studies show greatest concordance in left anterior STG (*B* and *D*; $n = 22$).

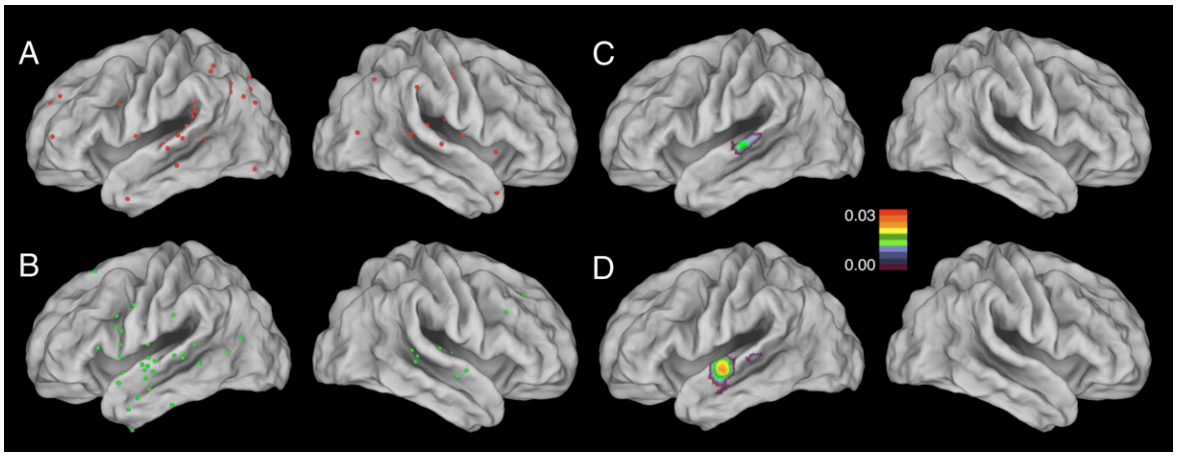

Fig. 3. Foci meeting inclusion criteria for length-based RS analyses (*A* and *B*) and ALE-statistic maps for regions of significant concordance (*C* and *D*) ($p < 10^{-3}$, $k > 150$ cm^3). Analyses show left lateralization and an anterior progression in peak effects with phoneme-length studies showing greatest concordance in left mid-STG (*A* and *C*; $n = 12$) and word-length studies showing greatest concordance in left anterior STG (*B* and *D*; $n = 16$). Too few studies exist for phrase-length analyses ($n = 4$).

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

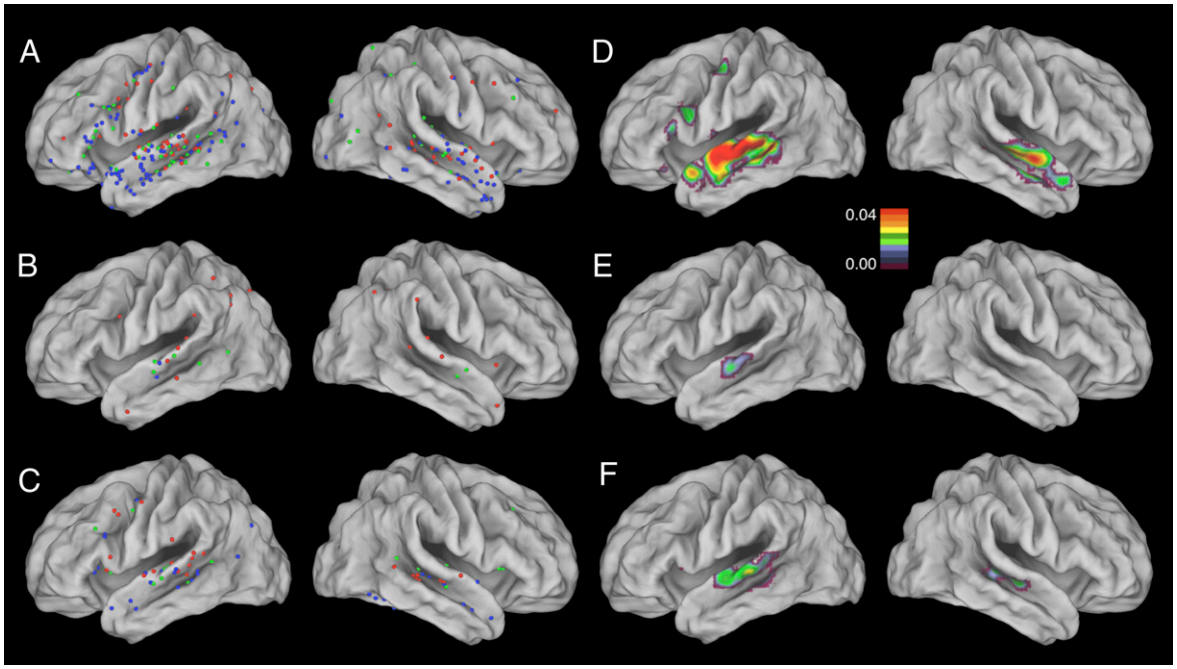

Fig. 4. Foci meeting inclusion criteria for length-pooled analyses (*A*–*C*) and ALE-statistic maps for regions of significant concordance (*D*–*F*) ($p < 10^{-3}$, $k > 150$ cm^3). Analyses show leftward bias in the CS (*A* and *D*; $n = 49$) and AS (*C* and *F*; $n = 15$) analyses and left lateralization in the IR (*B* and *E*; $n = 11$) analysis. Foci are color coded by stimulus length: phoneme length, red; word length, green; and phrase length, blue.

Table 2. Results for aggregate analyses

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

*Broader inclusion criteria for the IR analysis (*SI Appendix*, Table S3) yield equivalent results with the following qualifications: cluster volume 1,008 mm^3 and cluster concordance 0.33.

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.

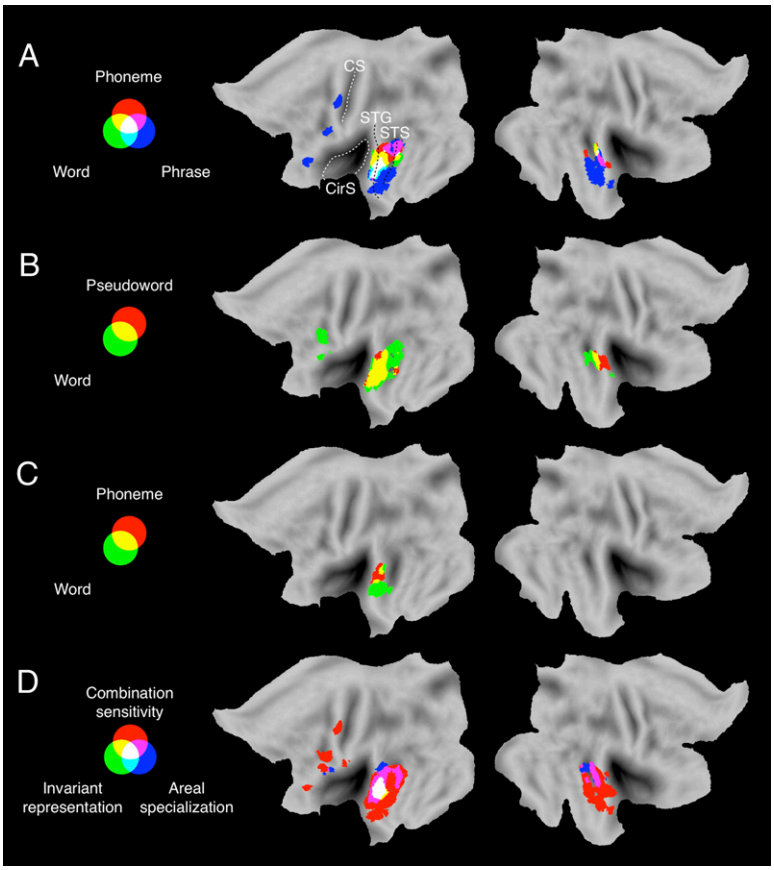

Fig. 5. Flat-map presentation of ALE cluster overlap for (*A*) the CS analyses shown in Fig. 1, (*B*) the word-length lexical status analyses shown in Fig. 2, (*C*) the RS analyses shown in Fig. 3, and (*D*) the length-pooled analyses shown in Fig. 4. For orientation, prominent landmarks are shown on 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. 1*F*) 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). Although we are aware of no study that has investigated letter and word recognition in a single sample, support for 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). A total of 84 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 in their stimuli: (*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 ALE, foci are treated as Gaussian probability distributions, which reflect localization 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 mm^3. 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 1RC1DC010720 (to J.P.R.).

1. Steinschneider M (2011) Unlocking the role of the superior temporal gyrus for speech sound categorization. *J Neurophysiol* 105:2631–2633.
2. Brugge JF, Merzenich MM (1973) Responses of neurons in auditory cortex of the macaque monkey to monaural and binaural stimulation. *J Neurophysiol* 36: 1138–1158.
3. Bitterman Y, Mukamel R, Malach R, Fried I, Nelken I (2008) Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature* 451:197–201.
4. Griffiths TD, Warren JD (2004) What is an auditory object? *Nat Rev Neurosci* 5: 887–892.
5. Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J Neurosci* 30:7604–7612.
6. Luce P, McLennan C (2005) Spoken word recognition: The challenge of variation. *Handbook of Speech Perception*, eds Pisoni D, Remez R (Blackwell, Malden, MA), pp 591–609.
7. Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160:106–154.
8. Riesenhuber M, Poggio TA (2002) Neural mechanisms of object recognition. *Curr Opin Neurobiol* 12:162–168.
9. Husain FT, Tagamets M-A, Fromm SJ, Braun AR, Horwitz B (2004) Relating neuronal dynamics for auditory object processing to neuroimaging activity: A computational modeling and an fMRI study. *Neuroimage* 21:1701–1720.
10. Dehaene S, Cohen L, Sigman M, Vinckier F (2005) The neural code for written words: a proposal. *Trends Cogn Sci* 9:335–341.
11. Hoffman KL, Logothetis NK (2009) Cortical mechanisms of sensory learning and object recognition. *Philos Trans R Soc Lond B Biol Sci* 364:321–329.
12. Larson E, Billimoria CP, Sen K (2009) A biologically plausible computational model for auditory object recognition. *J Neurophysiol* 101:323–331.
13. Hackett TA (2011) Information flow in the auditory cortical network. *Hear Res* 271: 133–146.
14. Suga N, O'Neill WE, Manabe T (1978) Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the mustache bat. *Science* 200: 778–781.
15. Margoliash D, Fortune ES (1992) Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. *J Neurosci* 12:4309–4326.
16. Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268:111–114.
17. Rauschecker JP (1997) Processing of complex sounds in the auditory cortex of cat, monkey, and man. *Acta Otolaryngol Suppl* 532:34–38.
18. Rauschecker JP (1998) Parallel processing in the auditory cortex of primates. *Audiol Neurootol* 3:86–103.
19. Sadagopan S, Wang X (2009) Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. *J Neurosci* 29:11192–11202.
20. Medvedev AV, Chiao F, Kanwal JS (2002) Modeling complex tone perception: grouping harmonics with combination-sensitive neurons. *Biol Cybern* 86:497–505.
21. Willmore BDB, King AJ (2009) Auditory cortex: representation through sparsification? *Curr Biol* 19:1123–1125.
22. Voytenko SV, Galazyuk AV (2007) Intracellular recording reveals temporal integration in inferior colliculus neurons of awake bats. *J Neurophysiol* 97: 1368–1378.
23. Peterson DC, Voytenko S, Gans D, Galazyuk A, Wenstrup J (2008) Intracellular recordings from combination-sensitive neurons in the inferior colliculus. *J Neurophysiol* 100:629–645.
24. Ye CQ, Poo MM, Dan Y, Zhang XH (2010) Synaptic mechanisms of direction selectivity in primary auditory cortex. *J Neurosci* 30:1861–1868.
25. Rao RP, Sejnowski TJ (2000) Predictive sequence learning in recurrent neocortical circuits. *Advances in Neural Information Processing Systems*, eds Solla SA, Leen TK, Muller KR (MIT Press, Cambridge), Vol 12.
26. Carr CE, Konishi M (1988) Axonal delay lines for time measurement in the owl's brainstem. *Proc Natl Acad Sci USA* 85:8311–8315.
27. Tian B, Rauschecker JP (2004) Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J Neurophysiol* 92:2993–3013.
28. Fukushima K (1980) Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern* 36: 193–202.
29. Riesenhuber M, Poggio TA (1999) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2:1019–1025.
30. Kouh M, Poggio TA (2008) A canonical neural circuit for cortical nonlinear operations. *Neural Comput* 20:1427–1451.
31. Lampl I, Ferster D, Poggio T, Riesenhuber M (2004) Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J Neurophysiol* 92:2704–2713.
32. Finn IM, Ferster D (2007) Computational diversity in complex cells of cat primary visual cortex. *J Neurosci* 27:9638–9648.
33. Bendor D, Wang X (2007) Differential neural coding of acoustic flutter within primate auditory cortex. *Nat Neurosci* 10:763–771.
34. Bendor D, Wang X (2008) Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *J Neurophysiol* 100:888–906.
35. Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative nonlinearities in auditory cortical neurons. *Neuron* 58:956–966.
36. Roe AW, Pallas SL, Kwon YH, Sur M (1992) Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex. *J Neurosci* 12:3651–3664.
37. Atencio CA, Sharpee TO, Schreiner CE (2009) Hierarchical computation in the canonical auditory cortical circuit. *Proc Natl Acad Sci USA* 106:21894–21899.
38. Ahmed B, Garcia-Lazaro JA, Schnupp JWH (2006) Response linearity in primary auditory cortex of the ferret. *J Physiol* 572:763–773.
39. Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc Natl Acad Sci USA* 97:11800–11806.
40. Romanski LM, et al. (1999) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat Neurosci* 2:1131–1136.
41. Kaas JH, Hackett TA (1999) 'What' and 'where' processing in auditory cortex. *Nat Neurosci* 2:1045–1047.
42. Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci* 12:718–724.
43. Romanski LM, Averbeck BB (2009) The primate cortical auditory system and neural representation of conspecific vocalizations. *Annu Rev Neurosci* 32:315–346.
44. Rauschecker JP (2011) An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hear Res* 271:16–25.
45. Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393–402.
46. Scott SK, Wise RJS (2004) The functional neuroanatomy of prelexical processing in speech perception. *Cognition* 92:13–45.
47. Binder JR, et al. (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex* 10:512–528.
48. Wise RJ, et al. (2001) Separate neural subsystems within 'Wernicke's area'. *Brain* 124: 83–95.
49. Patterson RD, Johnsrude IS (2008) Functional imaging of the auditory processing applied to speech sounds. *Philos Trans R Soc Lond B Biol Sci* 363:1023–1035.
50. Weiller C, Bormann T, Saur D, Musso M, Rijntjes M (2011) How the ventral pathway got lost: and what its recovery might mean. *Brain Lang* 118:29–39.
51. Whalen DH, et al. (2006) Differentiation of speech and nonspeech processing within primary auditory cortex. *J Acoust Soc Am* 119:575–581.
52. Nelken I (2008) Processing of complex sounds in the auditory system. *Curr Opin Neurobiol* 18:413–417.

53. Recanzone GH, Cohen YE (2010) Serial and parallel processing in the primate auditory cortex revisited. *Behav Brain Res* 206:1–7.
54. Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. *Science* 292:290–293.
55. Kikuchi Y, Horwitz B, Mishkin M (2010) Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *J Neurosci* 30:13021–13030.
56. Tsunada J, Lee JH, Cohen YE (2011) Representation of speech categories in the primate auditory cortex. *J Neurophysiol* 105:2634–2646.
57. Galaburda AM, Sanides F (1980) Cytoarchitectonic organization of the human auditory cortex. *J Comp Neurol* 190:597–610.
58. Chevillet M, Riesenhuber M, Rauschecker JP (2011) Functional correlates of the anterolateral processing hierarchy in human auditory cortex. *J Neurosci* 31: 9345–9352.
59. Glasser MF, Van Essen DC (2011) Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *J Neurosci* 31:11597–11616.
60. Poremba A, et al. (2004) Species-specific calls evoke asymmetric activity in the monkey's temporal poles. *Nature* 427:448–451.
61. Chang EF, et al. (2010) Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13:1428–1432.
62. Chang EF, et al. (2011) Cortical spatio-temporal dynamics underlying phonological target detection in humans. *J Cogn Neurosci* 23:1437–1446.
63. Steinschneider M, et al. (2011) Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb Cortex* 21:2332–2347.
64. Edwards E, et al. (2009) Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex. *J Neurophysiol* 102: 377–386.
65. Miller EK, Li L, Desimone R (1991) A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* 254:1377–1379.
66. Grill-Spector K, Malach R (2001) fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychol (Amst)* 107:293–321.
67. Joanisse MF, Zevin JD, McCandliss BD (2007) Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using FMRI and a short-interval habituation trial paradigm. *Cereb Cortex* 17:2084–2093.
68. Scott BH, Malone BJ, Semple MN (2011) Transformation of temporal processing across auditory cortex of awake macaques. *J Neurophysiol* 105:712–730.
69. Kusmierek P, Rauschecker JP (2009) Functional specialization of medial auditory belt cortex in the alert rhesus monkey. *J Neurophysiol* 102:1606–1622.
70. Creutzfeldt O, Ojemann G, Lettich E (1989) Neuronal activity in the human lateral temporal lobe. I. Responses to speech. *Exp Brain Res* 77:451–475.
71. Pei X, et al. (2011) Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage* 54:2960–2972.
72. Marinkovic K, et al. (2003) Spatiotemporal dynamics of modality-specific and supramodal word processing. *Neuron* 38:487–497.
73. Binder JR, Frost JA, Hammeke TA, Rao SM, Cox RW (1996) Function of the left planum temporale in auditory and linguistic processing. *Brain* 119:1239–1247.
74. Binder JR, et al. (1997) Human brain language areas identified by functional magnetic resonance imaging. *J Neurosci* 17:353–362.
75. Dehaene-Lambertz G, et al. (2006) Functional segregation of cortical language areas by sentence repetition. *Hum Brain Mapp* 27:360–371.
76. Sammler D, et al. (2010) The relationship of lyrics and tunes in the processing of unfamiliar songs: A functional magnetic resonance adaptation study. *J Neurosci* 30: 3572–3578.
77. Hara NF, Nakamura K, Kuroki C, Takayama Y, Ogawa S (2007) Functional neuroanatomy of speech processing within the temporal cortex. *Neuroreport* 18: 1603–1607.
78. Cohen L, Jobert A, Le Bihan D, Dehaene S (2004) Distinct unimodal and multimodal regions for word processing in the left temporal cortex. *Neuroimage* 23:1256–1270.
79. Buchsbaum BR, D'Esposito M (2009) Repetition suppression and reactivation in auditory-verbal short-term recognition memory. *Cereb Cortex* 19:1474–1485.
80. Matsumoto R, et al. (2011) Left anterior temporal cortex actively engages in speech perception: A direct cortical stimulation study. *Neuropsychologia* 49:1350–1354.
81. Turkeltaub PE, Eden GF, Jones KM, Zeffiro TA (2002) Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* 16: 765–780.
82. Thierry G, Giraud AL, Price CJ (2003) Hemispheric dissociation in access to the human semantic system. *Neuron* 38:499–506.
83. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. *Nature* 403:309–312.
84. Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res* 13:17–26.
85. Petkov CI, et al. (2008) A voice region in the monkey brain. *Nat Neurosci* 11:367–374.
86. Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci* 4:2051–2062.
87. Gaillard R, et al. (2006) Direct intracranial, FMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Neuron* 50:191–204.
88. Tsao DY, Freiwald WA, Tootell RBH, Livingstone MS (2006) A cortical region consisting entirely of face-selective cells. *Science* 311:670–674.
89. Kanwisher N, Yovel G (2006) The fusiform face area: A cortical region specialized for the perception of faces. *Philos Trans R Soc Lond B Biol Sci* 361:2109–2128.
90. Edwards E, et al. (2010) Spatiotemporal imaging of cortical activation during verb generation and picture naming. *Neuroimage* 50:291–301.
91. Turkeltaub PE, Coslett HB (2010) Localization of sublexical speech perception components. *Brain Lang* 114:1–15.
92. Samson F, Zeffiro TA, Toussaint A, Belin P (2011) Stimulus complexity and categorical effects in human auditory cortex: An activation likelihood estimation meta-analysis. *Front Psychol* 1:241.
93. Geschwind N (1970) The organization of language and the brain. *Science* 170: 940–944.
94. Bates E, et al. (2003) Voxel-based lesion-symptom mapping. *Nat Neurosci* 6:448–450.
95. Dronkers NF, Wilkins DP, Van Valin RD, Jr., Redfern BB, Jaeger JJ (2004) Lesion analysis of the brain areas involved in language comprehension. *Cognition* 92: 145–177.
96. Mazziotta JC, Phelps ME, Carson RE, Kuhl DE (1982) Tomographic mapping of human cerebral metabolism: Auditory stimulation. *Neurology* 32:921–937.
97. Petersen SE, Fox PT, Posner MI, Mintun M, Raichle ME (1988) Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* 331: 585–589.
98. Wise RJS, et al. (1991) Distribution of cortical neural networks involved in word comprehension and word retrieval. *Brain* 114:1803–1817.
99. Démonet JF, et al. (1992) The anatomy of phonological and semantic processing in normal subjects. *Brain* 115:1753–1768.
100. Rademacher J, et al. (2001) Probabilistic mapping and volume measurement of human primary auditory cortex. *Neuroimage* 13:669–683.
101. Hamberger MJ, Seidel WT, Goodman RR, Perrine K, McKhann GM (2003) Temporal lobe stimulation reveals anatomic distinction between auditory naming processes. *Neurology* 60:1478–1483.
102. Hashimoto Y, Sakai KL (2003) Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. *Hum Brain Mapp* 20:22–28.
103. Warren JE, Wise RJS, Warren JD (2005) Sounds do-able: Auditory-motor transformations and the posterior temporal plane. *Trends Neurosci* 28:636–643.
104. Guenther FH (2006) Cortical interactions underlying the production of speech sounds. *J Commun Disord* 39:350–365.
105. Tourville JA, Reilly KJ, Guenther FH (2008) Neural mechanisms underlying auditory feedback control of speech. *Neuroimage* 39:1429–1443.
106. Towle VL, et al. (2008) ECoG gamma activity during a language task: Differentiating expressive and receptive speech areas. *Brain* 131:2013–2027.
107. Takaso H, Eisner F, Wise RJS, Scott SK (2010) The effect of delayed auditory feedback on activity in the temporal lobe while speaking: A positron emission tomography study. *J Speech Lang Hear Res* 53:226–236.
108. Zheng ZZ, Munhall KG, Johnsrude IS (2010) Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production. *J Cogn Neurosci* 22:1770–1781.
109. Buchsbaum BR, Padmanabhan A, Berman KF (2011) The neural substrates of recognition memory for verbal information: Spanning the divide between short- and long-term memory. *J Cogn Neurosci* 23:978–991.
110. Buchsbaum BR, Olsen RK, Koch P, Berman KF (2005) Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron* 48:687–697.
111. Vinckier F, et al. (2007) Hierarchical coding of letter strings in the ventral stream: dissecting the inner organization of the visual word-form system. *Neuron* 55: 143–156.
112. Dehaene S, et al. (2010) How learning to read changes the cortical networks for vision and language. *Science* 330:1359–1364.
113. Pallier C, Devauchelle A-D, Dehaene S (2011) Cortical representation of the constituent structure of sentences. *Proc Natl Acad Sci USA* 108:2522–2527.
114. Graves WW, Desai R, Humphries C, Seidenberg MS, Binder JR (2010) Neural systems for reading aloud: A multiparametric approach. *Cereb Cortex* 20:1799–1815.
115. Jobard G, Crivello F, Tzourio-Mazoyer N (2003) Evaluation of the dual route theory of reading: A metanalysis of 35 neuroimaging studies. *Neuroimage* 20:693–712.
116. Turkeltaub PE, Gareau L, Flowers DL, Zeffiro TA, Eden GF (2003) Development of neural mechanisms for reading. *Nat Neurosci* 6:767–773.
117. Hamberger MJ, Goodman RR, Perrine K, Tamny T (2001) Anatomic dissociation of auditory and visual naming in the lateral temporal cortex. *Neurology* 56:56–61.
118. Hamberger MJ, McClelland S, III, McKhann GM, II, Williams AC, Goodman RR (2007) Distribution of auditory and visual naming sites in nonlesional temporal lobe epilepsy patients and patients with space-occupying temporal lobe lesions. *Epilepsia* 48:531–538.
119. Blau V, van Atteveldt N, Formisano E, Goebel R, Blomert L (2008) Task-irrelevant visual letters interact with the processing of speech sounds in heteromodal and unimodal cortex. *Eur J Neurosci* 28:500–509.
120. van Atteveldt NM, Blau VC, Blomert L, Goebel R (2010) fMR-adaptation indicates selectivity to audiovisual content congruency in distributed clusters in human superior temporal cortex. *BMC Neurosci* 11:11.
121. Beauchamp MS, Nath AR, Pasalar S (2010) fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *J Neurosci* 30:2414–2417.
122. Nath AR, Beauchamp MS (2011) Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *J Neurosci* 31:1704–1714.
123. Ison MJ, Quiroga RQ (2008) Selectivity and invariance for visual object perception. *Front Biosci* 13:4889–4903.
124. Kuhl PK (2004) Early language acquisition: Cracking the speech code. *Nat Rev Neurosci* 5:831–843.
125. Glezer LS, Jiang X, Riesenhuber M (2009) Evidence for highly selective neuronal tuning to whole words in the "visual word form area". *Neuron* 62:199–204.
126. Cappelle B, Shtyrov Y, Pulvermüller F (2010) Heating up or cooling up the brain? MEG evidence that phrasal verbs are lexical units. *Brain Lang* 115:189–201.

127. Scott SK, Blank CC, Rosen S, Wise RJS (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123:2400–2406.
128. Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb Cortex* 19:2767–2796.
129. Rogalsky C, Hickok G (2011) The role of Broca's area in sentence comprehension. *J Cogn Neurosci* 23:1664–1680.
130. Obleser J, Meyer L, Friederici AD (2011) Dynamic assignment of neural resources in auditory comprehension of complex sentences. *Neuroimage* 56:2310–2320.
131. Humphries C, Binder JR, Medler DA, Liebenthal E (2006) Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J Cogn Neurosci* 18:665–679.
132. Tyler LK, Marslen-Wilson W (2008) Fronto-temporal brain systems supporting spoken language comprehension. *Philos Trans R Soc Lond B Biol Sci* 363:1037–1054.
133. Friederici AD, Kotz SA, Scott SK, Obleser J (2010) Disentangling syntax and intelligibility in auditory language comprehension. *Hum Brain Mapp* 31:448–457.
134. Guenther FH (1994) A neural network model of speech acquisition and motor equivalent speech production. *Biol Cybern* 72:43–53.
135. Cohen YE, Andersen RA (2002) A common reference frame for movement plans in the posterior parietal cortex. *Nat Rev Neurosci* 3:553–562.
136. Hackett TA, et al. (2007) Sources of somatosensory input to the caudal belt areas of auditory cortex. *Perception* 36:1419–1430.
137. Smiley JF, et al. (2007) Multisensory convergence in auditory cortex, I. Cortical connections of the caudal superior temporal plane in macaque monkeys. *J Comp Neurol* 502:894–923.
138. Hackett TA, et al. (2007) Multisensory convergence in auditory cortex, II. Thalamocortical connections of the caudal superior temporal plane. *J Comp Neurol* 502:924–952.
139. Dhanjal NS, Handunnetthi L, Patel MC, Wise RJS (2008) Perceptual systems controlling speech production. *J Neurosci* 28:9969–9975.
140. Baddeley A (2003) Working memory: Looking back and looking forward. *Nat Rev Neurosci* 4:829–839.
141. Fitch WT (2000) The evolution of speech: A comparative review. *Trends Cogn Sci* 4: 258–267.
142. McCandliss BD, Cohen L, Dehaene S (2003) The visual word form area: Expertise for reading in the fusiform gyrus. *Trends Cogn Sci* 7:293–299.
143. Wandell BA, Rauschecker AM, Yeatman JD (2012) Learning to see words. *Ann Rev Psychol* 63:31–53.
144. Turkeltaub PE, Flowers DL, Lyon LG, Eden GF (2008) Development of ventral stream representations for single letters. *Ann N Y Acad Sci* 1145:13–29.
145. Joseph JE, Cerullo MA, Farley AB, Steinmetz NA, Mier CR (2006) fMRI correlates of cortical specialization and generalization for letter processing. *Neuroimage* 32: 806–820.
146. Pernet C, Celsis P, Démonet J-F (2005) Selective response to letter categorization within the left fusiform gyrus. *Neuroimage* 28:738–744.
147. Callan AM, Callan DE, Masaki S (2005) When meaningless symbols become letters: Neural activity change in learning new phonograms. *Neuroimage* 28:553–562.
148. Longcamp M, Anton J-L, Roth M, Velay J-L (2005) Premotor activations in response to visually presented single letters depend on the hand used to write: A study on left-handers. *Neuropsychologia* 43:1801–1809.
149. Flowers DL, et al. (2004) Attention to single letters activates left extrastriate cortex. *Neuroimage* 21:829–839.
150. Longcamp M, Anton J-L, Roth M, Velay J-L (2003) Visual presentation of single letters activates a premotor area involved in writing. *Neuroimage* 19:1492–1500.
151. Logothetis NK, Pauls J (1995) Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb Cortex* 5:270–288.
152. Dehaene S, et al. (2010) Why do children make mirror errors in reading? Neural correlates of mirror invariance in the visual word form area. *Neuroimage* 49: 1837–1848.
153. Pegado F, Nakamura K, Cohen L, Dehaene S (2011) Breaking the symmetry: Mirror discrimination for single letters but not for pictures in the visual word form area. *Neuroimage* 55:742–749.
154. Lancaster JL, et al. (2007) Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Hum Brain Mapp* 28:1194–1205.
155. Eickhoff SB, et al. (2009) Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum Brain Mapp* 30:2907–2926.
156. Turkeltaub PE, et al. (2012) Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Hum Brain Mapp* 33:1–13.
157. Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878.
158. Van Essen DC (2005) A Population-Average, Landmark- and Surface-based (PALS) atlas of human cerebral cortex. *Neuroimage* 28:635–662.