





Evidence for a Spoken Word Lexicon in the Auditory Ventral Stream

Srikanth R. Damera¹ , Lillian Chang¹, Plamen P. Nikolov¹, James A. Mattei¹,
Suneel Banerjee¹ , Laurie S. Glezer², Patrick H. Cox¹, Xiong Jiang¹,
Josef P. Rauschecker¹, and Maximilian Riesenhuber¹

¹Department of Neuroscience, Georgetown University Medical Center, Washington, DC, USA

²Department of Speech, Language, and Hearing Sciences, San Diego State University, San Diego, CA, USA

Keywords: auditory lexicon, auditory ventral stream, speech recognition, superior temporal gyrus

ABSTRACT

The existence of a neural representation for whole words (i.e., a lexicon) is a common feature of many models of speech processing. Prior studies have provided evidence for a visual lexicon containing representations of whole written words in an area of the ventral visual stream known as the *visual word form area*. Similar experimental support for an auditory lexicon containing representations of spoken words has yet to be shown. Using functional magnetic resonance imaging rapid adaptation techniques, we provide evidence for an auditory lexicon in the *auditory word form area* in the human left anterior superior temporal gyrus that contains representations highly selective for individual spoken words. Furthermore, we show that familiarization with novel auditory words sharpens the selectivity of their representations in the auditory word form area. These findings reveal strong parallels in how the brain represents written and spoken words, showing convergent processing strategies across modalities in the visual and auditory ventral streams.

INTRODUCTION

Speech perception is perhaps the most remarkable achievement of the human auditory system and one that likely is critically dependent on its overall cortical architecture. It is generally accepted that the functional architecture of auditory cortex in human and nonhuman primates comprises two processing streams (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Rauschecker & Tian, 2000). There is an auditory dorsal stream that is involved in the processing of auditory space and motion (van der Heijden et al., 2019) as well as in sensorimotor transformations such as those required for speech production (Archakov et al., 2020; Hickok et al., 2011; Rauschecker, 2011, 2018). There is also an auditory ventral stream specialized for recognizing auditory objects such as spoken words. This stream is organized along a simple-to-complex feature hierarchy (Rauschecker & Scott, 2009), akin to the organization of the visual ventral stream (Kravitz et al., 2013).

Visual object recognition studies support a simple-to-complex model of cortical visual processing in which neuronal populations in the visual ventral stream are selective for increasingly complex features and ultimately visual objects along a posterior-to-anterior gradient extending from lower-to-higher-order visual areas (Felleman & Van Essen, 1991; Hubel & Wiesel, 1977). For the special case of recognizing written words, this simple-to-complex model predicts that progressively more anterior neuronal populations are selective for

Citation: Damera, S. R., Chang, L., Nikolov, P. P., Mattei, J. A., Banerjee, S., Glezer, L. S., Cox, P. H., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2023). Evidence for a spoken word lexicon in the auditory ventral stream. *Neurobiology of Language*, 4(3), 420–434. https://doi.org/10.1162/nol_a_00108

DOI:
https://doi.org/10.1162/nol_a_00108

Supporting Information:
https://doi.org/10.1162/nol_a_00108

Received: 12 December 2022
Accepted: 27 April 2023

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Srikanth R. Damera
srd49@georgetown.edu

Handling Editor:
Sophie Scott

Copyright: © 2023
Massachusetts Institute of Technology
Published under a Creative Commons
Attribution 4.0 International
(CC BY 4.0) license



The MIT Press

increasingly complex orthographic patterns (Dehaene et al., 2005; Vinckier et al., 2007). Thus, analogous to general visual processing, orthographic word representations are predicted to culminate in representations of whole visual words—an orthographic lexicon. Evidence suggests that these lexical representations are subsequently linked to concept representations in downstream areas like the anterior temporal lobe (Damera et al., 2020; Lambon Ralph et al., 2017; Liuzzi et al., 2019; Malone et al., 2016). The existence of this orthographic lexicon in the brain is predicted by neuropsychological studies of reading (Coltheart, 2004). Indeed, functional magnetic resonance imaging (fMRI; Glezer et al., 2009, 2015) and, more recently, electrocorticographic data (Hirshorn et al., 2016; Woolnough et al., 2021) have confirmed the existence of such a lexicon in a region of the posterior fusiform cortex known as the *visual word form area* (VWFA; Dehaene & Cohen, 2011; Dehaene et al., 2005).

It has been proposed that an analogous simple-to-complex hierarchy exists in the auditory ventral stream as well (Kell et al., 2018; Rauschecker, 1998) extending anteriorly from Heschl's gyrus along the superior temporal cortex (STC; DeWitt & Rauschecker, 2012; Rauschecker & Scott, 2009). Yet, the existence and location of a presumed *auditory lexicon*, that is, a neural representation for the recognition (and storage) of real words, has been quite controversial (Bogen & Bogen, 1976): The traditional *posterior* view is that the auditory lexicon should be found in posterior STC (pSTC; Geschwind, 1970). In contrast, a notable meta-analysis (DeWitt & Rauschecker, 2012) provided strong evidence for the existence of word-selective auditory representations in *anterior* STC (aSTC), consistent with imaging studies of speech intelligibility (Binder et al., 2000; Scott et al., 2000) and proposals for an *auditory word form area* (AWFA) in the human left anterior temporal cortex (Cohen et al., 2004; DeWitt & Rauschecker, 2012). Such a role of the aSTC is compatible with nonhuman primate studies that show selectivity for complex communication calls in aSTC (Ortiz-Rios et al., 2015; Rauschecker et al., 1995; Tian et al., 2001) and demonstrate, in humans and nonhuman primates, that progressively anterior neuron populations in the STC pool over longer timescales (Hamilton et al., 2018; Hullett et al., 2016; Jasmin et al., 2019; Kajikawa et al., 2015). In this anterior account of lexical processing, the pSTC and speech-responsive regions in the IPL are posited to be involved in “inner speech” and phonological reading (covert articulation), but not auditory comprehension (DeWitt & Rauschecker, 2013; Rauschecker, 2011). Yet, despite this compelling alternative to traditional theories, there is still little direct evidence for an auditory lexicon in the aSTC.

Investigating the existence and location of auditory lexica is critical for understanding the neural bases of speech processing and, consequently, the neural underpinnings of speech processing disorders. However, finely probing the selectivity of neural representations in the human brain with fMRI is challenging, in part because it is difficult to assess the selectivity of these populations. Many studies have identified speech processing areas by contrasting speech stimuli with various nonspeech controls (Evans et al., 2014; Okada et al., 2010; Scott et al., 2000). However, these coarse contrasts cannot reveal what neurons in a particular auditory word-responsive region of interest (ROI) are selective for, for example, phonemes, syllables, or whole words. More sensitive techniques such as fMRI rapid adaptation (fMRI-RA; Grill-Spector & Malach, 2001; Krekelberg et al., 2006) are needed to probe the selectivity of speech representations in the brain and resolve the question of the existence of auditory lexica. In the current study, we used fMRI-RA to test the existence of lexical representations in the auditory ventral stream. Paralleling previous work in the visual system that used fMRI-RA to provide evidence for the existence of an orthographic lexicon in the VWFA (Glezer et al., 2009, 2015, 2016), we first performed an independent auditory localizer scan that we used to identify the AWFA in individual subjects, and then conducted three fMRI-RA scans that probed

the representation in the AWFA and its plasticity. The first two scans consisted of real words (RWs) and pseudowords (PWs; i.e., pronounceable nonwords), respectively. These scans revealed an adaptation profile consistent with lexical selectivity in the putative AWFA for RWs, but not novel PWs, directly replicating results for written words in the VWFA. We then tested the lexicon hypothesis by predicting that training subjects to recognize novel PWs would add them to their auditory lexica, leading them to exhibit lexical-like selectivity in the AWFA following training, as previously shown for written words in the VWFA (Glezer et al., 2015). To do so, we conducted a third fMRI-RA scan after PW training. Results from this scan showed RW-like lexical selectivity to the now-familiar PWs following training, supporting the role of the AWFA as an auditory lexicon shaped by experience with auditory words.

MATERIALS AND METHODS

Overall Procedure

In this study, participants completed an auditory localizer scan and three RA experiments over the course of three scanning sessions. In the first session, subjects completed the auditory localizer and RW scans. The auditory localizer scan was used to identify a candidate AWFA region in the anterior auditory ventral stream (Rauschecker & Scott, 2009), adopting the approach used in the visual/orthographic case for defining the VWFA (Glezer et al., 2009, 2015, 2016). The RW RA scan was used to test if the candidate AWFA exhibited lexical selectivity for RWs. In the remaining two scanning sessions subjects completed pre- and post-training PW scans separated by six behavioral training sessions outside of the scanner. During the pre-training scan, subjects were presented with the then *untrained pseudowords* (UTPW). Next, subjects completed six behavioral training sessions consisting of a 2-back and a novel/familiar task designed to familiarize and assess subject familiarity with the UTPW, respectively. Subjects completed a maximum of one behavioral session each day and those participants that achieved at least 80% accuracy on the novel/familiar task by the final session qualified for the post-training scan. During the post-training scan, subjects were presented with the same set of PWs, now called *trained pseudowords* (TPW). The pre- and post-training scans were used to test whether the candidate AWFA exhibited lexical selectivity for TPW (i.e., after training) but not to the UTPW (i.e., before training).

Participants

We recruited a total of 28 right-handed healthy native English speakers for this study (ages 18–34, 12 females). Georgetown University's Institutional Review Board approved all experimental procedures, and written informed consent was obtained from all subjects before the experiment. Two subjects were excluded from further analyses for performing 2 standard deviations below the average on the in-scanner task. Furthermore, two subjects dropped out of the study after completing the RW RA scan. In total, we analyzed 26 subjects for the RW RA scan and 24 of those 26 for the pre-training UTPW scan. Due to subject dropout ($n = 4$) or failure to achieve 80% accuracy on the novel/familiar task ($n = 4$), 16 out of the 24 subjects were analyzed for the post-training TPW scan.

Stimuli

Real word stimuli, for the RW RA experiment, were chosen using the English Lexicon Project (Balota et al., 2007). Analogous to studies of the neural representation of written words (Glezer et al., 2009, 2015), three sets of 50 high-frequency (>50 per million), monosyllabic RWs that were 3–5 phonemes in length were created. One set of words (target words) served as the

reference for the other two lists. The second set was created by altering each target word by a single phoneme to create another RW. The third set was created by selecting for each target word another RW that had the same number of (but no shared) phonemes. All three of these lists were matched on the number of phonemes, orthographic neighborhood, and phonological neighborhood. To create the UTPW/TPW we used MCWord (Medler & Binder, 2005) to generate four sets of 50 target PWs, 3–5 phonemes in length. One set to be trained and three sets to remain untrained and serve as foils in the training task. As with the RW stimuli, we then used the target PW as the reference to generate a set of PWs each differing by one phoneme from the target word, and another set of PWs matched to the target PWs by number of phonemes but not sharing any phonemes. RW and PW sets were matched for length, bigram and trigram frequency, and phonological neighborhood. All stimuli were recorded using a 44.1-kHz sampling rate in a sound-isolated booth by a trained female native speaker of American English.

Auditory Localizer Scan

The auditory localizer scan (Figure 1A) was used to independently identify auditory speech-selective areas. In this scan, subjects were randomly presented with trials from one of five conditions: Real Words, Pseudowords, Scrambled Real Words, Scrambled Pseudowords, and

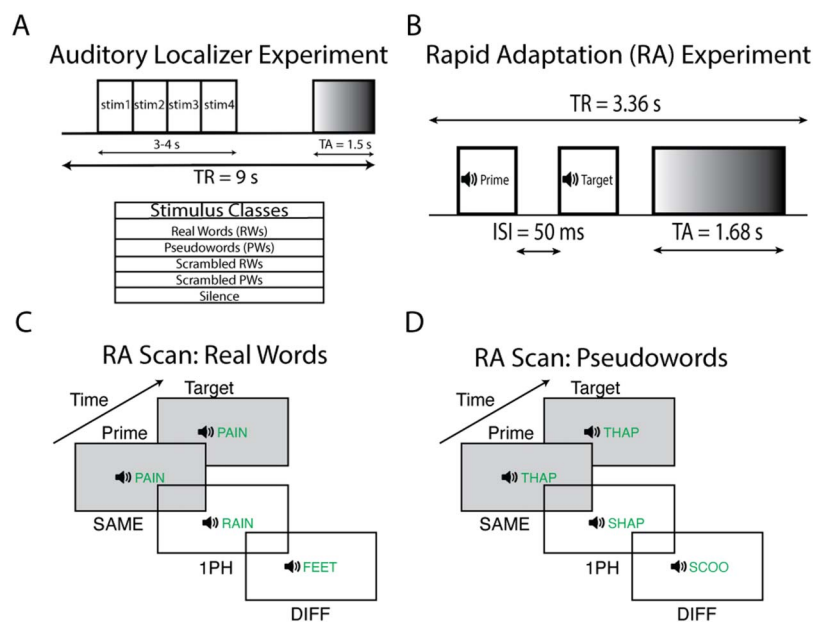


Figure 1. Rapid adaptation and auditory localizer experimental paradigms. (A) The slow clustered acquisition paradigm used in the auditory localizer scan. Each trial was 9 s long with 1.5 s of volume acquisition and 7.5 s of silence. During the silent period, the subject heard four sounds from one of five stimulus classes and performed a 1-back task. (B) The rapid clustered acquisition paradigm used for the RA scans (Chevillet et al., 2013; Jiang et al., 2018). Each trial was 3.36 s long with 1.68 s of volume acquisition. During the silent period, two spoken words were played to the subject with a 50 ms interstimulus interval. The first word acted as a prime and the second word the target. The experimental paradigms for (C) real words and (D) pseudowords. The prime was followed by a target word that was either the same word (SAME), a word that differed from the target by one phoneme (1PH), or a word that shared no phonemes with the target (DIFF). Furthermore, subjects were presented with silence trials that served as an explicit baseline. During the task, subjects were asked to attend to all the words and respond when they heard the oddball stimulus (RW or PW containing the rhyme “-ox,” e.g., “socks”) in either the prime or target position.

Silence. Real Words and Pseudowords were one syllable long; the lists were matched for length and orthographic and phonologic neighborhood. Scrambled Real Words and Scrambled Pseudowords were generated by randomly rearranging 200 ms by 1-octave tiles of the constant Q spectrogram (Brown, 1991) for each RW or PW, respectively, and then reconstructing a time-domain waveform with an inverse transform (Ortiz-Rios et al., 2015). Each trial in the auditory localizer scan was 9 s long and began with 2.5 s of silence (during which 1.68 s of scanner acquisition occurred), followed by a ~3–4 s stimulus presentation period, and concluding with silence. During the stimulus presentation period, subjects heard four stimuli from a given condition and responded with a left-handed button press if any of the four stimuli was a repeat within that block. In total, there were 145 trials per run with 25 trials of each of the five conditions and five 1-back trials for each of the four non-Silence conditions. An additional 18 s of fixation were added to the start and end of each run. Subjects completed five runs of the task.

Rapid Adaptation Scans

There were three RA scans (Figure 1B–D; i.e., RW, UTPW, TPW) performed on different days. In the RA scans, subjects heard a pair of words (prime/target) on each trial (Figure 1C–D). The words in each pair were either identical (SAME), differed by a single phoneme (1PH), or shared no phonemes at all (DIFF). These pairs were generated by using the three matched word lists described above. To engage subjects' attention, we asked subjects to perform an oddball rhyme detection task in the scanner. To do so, we created an additional condition (Oddball) in which a word or pseudoword containing the oddball rhyme “-ox” (e.g., socks, grox) was presented in lieu of either the prime or target word. Participants were asked to attentively listen to all stimuli and respond with a left-handed button press when an oddball stimulus was heard. In all three scans, the number of repetitions of each word was counterbalanced across all conditions to control for long-lag priming effects (Henson et al., 2000). Trial order and timing were adjusted using M-sequences (Buračas & Boynton, 2002). Each trial was 3.36 s long and consisted of a 1.68 s silent period followed by 1.68 s stimulus presentation period during which the word pairs were presented. Following prior auditory RA studies (Chevillet et al., 2013; Jiang et al., 2018), we presented stimuli with a 50 ms ISI. In total, there were 25 trials of each condition (SAME, 1PH, DIFF, oddball, and silence) for a total of 125 trials per run. An additional 10.08 s of fixation was added to the start and end of each run. Subjects completed four runs for each scan.

Data Acquisition

MRI data were acquired on a 3.0 Tesla Siemens Prisma-fit scanner. We used whole-head echo-planar imaging (EPI) sequences (flip angle = 70°, echo time [TE] = 35 ms, field of view [FOV] = 205 mm, 102 × 102 matrix) with a 64-channel head coil. Building off other auditory localizer (Damera et al., 2021) and RA paradigms (Chevillet et al., 2013), a slow (repetition time [TR] = 9,000 ms, acquisition time [TA] = 1,680 ms) and a fast (TR = 3,360 ms, TA = 1,680 ms) clustered acquisition paradigm were used for the auditory localizer and RA scans, respectively. Fifty-four axial slices were acquired in descending order (thickness = 1.8 mm, 0.18 mm gap; in-plane resolution = 2.0 × 2.0 mm²). A T1-weighted MPRAGE (magnetization-prepared rapid acquisition with gradient echo) image (resolution 1 × 1 × 1 mm³) was also acquired for each subject.

fMRI Data Preprocessing

Image preprocessing was performed using SPM12 (Ashburner et al., 2021; <https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>). The first three volumes of each run were discarded to allow for T1 stabilization, and the remaining EPI images were spatially realigned to the mean blood

oxygen level-dependent (BOLD) reference image. No slice-time correction was done given the presence of temporal discontinuities between successive volumes in clustered acquisition paradigms (Perrachione & Ghosh, 2013). EPI images for each subject were co-registered to the anatomical image. The anatomical image was then segmented and the resulting deformation fields for spatial normalization were used to normalize the functional data to the standard MNI (Montreal Neurological Institute) space. Next, we smoothed the normalized functional images with a 4-mm full-width at half maximum Gaussian kernel. Finally, a first-level model containing regressors for each condition and the six motion parameters from realignment was fit. In the auditory localizer scan, the regressors were Real Words, Pseudowords, Scrambled Real Words, Scrambled Pseudowords, Button Press, and Silence. In the RA scans, the regressors were SAME, 1PH, DIFF, Button Press, and Silence.

Defining the Auditory Word Form Area

The AWFA was determined at the individual-subject level using the auditory localizer scan. Analogous to studies of the visual word form area (Glezer et al., 2009, 2015, 2016), for each subject we first defined an RW vs. Silence contrast and an RW vs. Scrambled Words contrast. Next, at the group level, we masked the RW vs. Silence t statistic map with the RW vs. Scrambled Words t statistic map thresholded at $p < 0.05$ before applying a $p < 0.001$ voxel-wise threshold. The resulting map had a peak at MNI: $-62, -14, 2$. These coordinates are consistent with the hypothesized locus (MNI: $-61, -15, -5$) of the AWFA from prior studies (DeWitt & Rauschecker, 2012). We then created the same RW vs. Silence masked by RW vs. Scrambled Words maps at the individual-subject level using the same thresholds as above. Then, for each subject we identified the local peak in the resulting maps closest to the group peak (MNI: $-62, -14, 2$). Finally, to create each individual subject's AWFA, we created an ROI consisting of the 50 closest voxels to each individual's local peak.

Behavioral Training

Subjects were trained to recognize 150 auditory PWs (TPW, see above). A 2-back training task, in which subjects had to detect repeats of a PW separated by another PW, was used to familiarize subjects with the PWs. Each session of the 2-back task consisted of 15 blocks of 75 trials each with self-paced breaks between each block. Each trial lasted 1.5 s during which subjects heard a single PW and had to respond if they heard a 2-back repeat (i.e., if the current PW was the same as the PW before the last). Each block lasted 112.5 s, and each session lasted for a total task length of 28.125 min excluding breaks. Following the 2-back task, subjects' familiarity with the trained PWs was assessed using a novel/familiar task. For this task, we developed three sets of foils for each of the 150 PWs. Each foil differed from its base PW by a single phoneme. Each session of the novel/familiar task consisted of three blocks of 100 trials with self-paced breaks between each block. Each trial lasted 1.5 s during which subjects heard a single PW and had to respond with either the left or right arrow key to indicate a novel (i.e., a foil) or familiar (i.e., trained) PW. In total, each block lasted 150 s for a total task length of 7.5 min excluding breaks. Over the course of the novel/familiar task sessions each foil list was paired with the trained PW list only twice with at least two days since the last pairing. Six sessions of both tasks were performed over the course of about eight days. To proceed to the post-training RA scan, subjects had to achieve at least 80% accuracy on the novel/familiar task by their sixth session.

Task-Based Functional Connectivity

After preprocessing, we used the CONN toolbox Version 21.a (Whitfield-Gabrieli & Nieto-Castanon, 2012) to calculate seed-to-voxel stimulus-driven functional connectivity from the

AWFA during auditory word processing in the auditory localizer scans. Word processing was included as the primary task condition by combining the onset times and durations for RW and PW into one condition. The individual subject ($n = 26$) AWFA seeds were used for the analysis. We then performed denoising with confounds, which included the subject-specific motion regressors computed during preprocessing in SPM, cerebrospinal fluid and white-matter signals, and their derivatives, following the CompCor strategy (Behzadi et al., 2007) as implemented in CONN. Data were then band-pass filtered (0.008–0.09 Hz) with linear detrending. Seed-to-voxel functional connectivity was performed using a weighted general linear model computing the bivariate correlation between the AWFA seed with the whole brain. Due to the sparse acquisition type of this analysis, no hemodynamic response function weighting was performed. Group-level significance was determined with a voxel threshold of $p < 0.001$ and a cluster threshold of $p < 0.05$, false discovery rate (FDR) corrected.

RESULTS

Auditory Localizer Scan Identifies Bilateral Speech-Selective ROI in the Auditory Ventral Stream

An independent auditory localizer scan was used to identify a putative AWFA near previous literature coordinates (Figure 2A). To do so, analogous to prior studies of written word representations in the VWFA (Glezer et al., 2009, 2015, 2016), we first identified group-level auditory speech-selective areas (see Materials and Methods) by examining the (Real Words) vs. Silence contrast thresholded at $p < 0.001$ masked by the RW vs. Scrambled Real Words contrast thresholded at $p < 0.05$ at the group-level cluster-corrected at the FDR $p < 0.05$. This approach mimicked definition of the VWFA in prior studies (Glezer et al., 2009, 2015) and gave us the greatest flexibility for defining ROIs while ensuring responsiveness to sound and intelligible speech. This revealed several clusters of activation in the superior temporal, frontal,

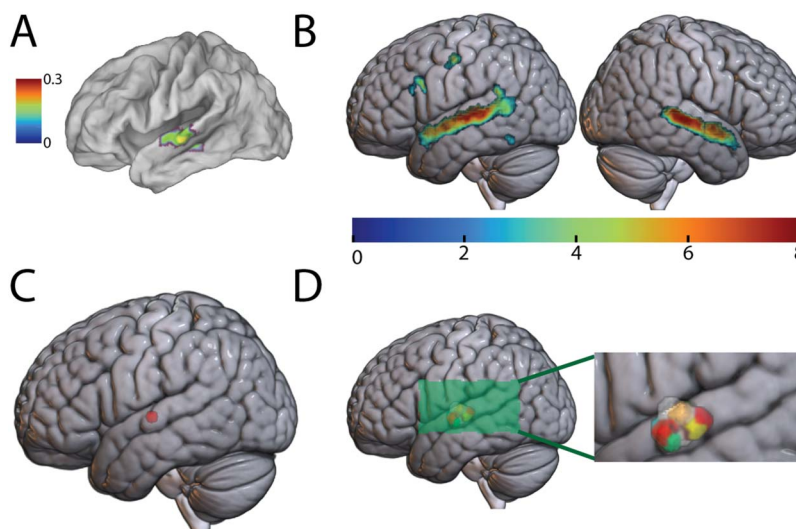


Figure 2. Identifying the auditory word form area (AWFA). (A) Proposed location of the AWFA (MNI: $-61, -15, -5$). Adapted from DeWitt and Rauschecker (2012). Color bar (arbitrary units) reflects the activation likelihood estimatoin (Laird et al., 2005) statistic. (B) The RW vs. Silence contrast ($p < 0.001$) masked by the RW vs. Scrambled Real Words contrast ($p < 0.05$) in the auditory localizer scan. Only clusters significant at the FDR $p < 0.05$ level are shown. Colors reflect t statistics. (C) The peak in the left STG (MNI: $-62, -14, 2$). (D) The AWFA defined in individual subjects. The inset zooms in on the perisylvian region to highlight the location of the AWFA.

and inferior temporal cortices (Figure 2B). A local peak in the left superior temporal gyrus (STG) was identified at MNI: $-62, -14, 2$ (Figure 2C) near the literature coordinates (DeWitt & Rauschecker, 2012) of MNI: $-61, -15, -5$. Individual subject AWFA ROIs (Figure 2D) were created by building a 50-voxel ROI (400 mm^3) around the local peak of each subject (mean \pm SD: $-62 \pm 2, -14.9 \pm 3, 2.6 \pm 2.6$) closest to the group peak (see Materials and Methods). This is similar to the size of the VWFA identified in prior studies (Glezer et al., 2009, 2015).

Lexical Selectivity for Real Words but Not Pseudowords in the AWFA

The first two fMRI-RA scans were performed with RWs and UTPWs, respectively. In these experiments, we predicted the lowest signal for the SAME condition since the two identical words presented in that condition would repeatedly activate the same neural populations, thereby causing maximal adaptation. Likewise, we predicted the least amount of adaptation for the DIFF condition because two words that share no phonemes should activate disjoint groups of neurons, irrespective of whether responsiveness to auditory words in the localizer scan in that ROI was due to neurons selective for phonemes, syllables, or whole words. Finally, we tested specific predictions regarding responses in the 1PH condition. Specifically, if neurons in the AWFA ROI are tightly tuned to whole RWs (i.e., if the AWFA contains an auditory lexicon), the two similar but nonidentical RWs in the 1PH condition should have minimal neural overlap, and therefore no adaptation should occur and response levels in the 1PH condition should be comparable to that of the DIFF condition (as found for written real words in the VWFA; Glezer et al., 2009). In contrast, if neurons in the AWFA were tuned to sublexical phoneme combinations, then there should be a gradual release from adaptation from SAME to 1PH to DIFF, with $1\text{PH} < \text{DIFF}$, as sublexical overlap would continue to increase from 1PH to DIFF. For PWs, we predicted that there would be a gradual increase in the BOLD signal paralleling the increasing dissimilarity (i.e., SAME to 1PH to DIFF). This is thought to reflect low-level activation of RW-tuned neurons to phonologically similar PWs. These predictions mirror findings in the VWFA for written words (Glezer et al., 2009, 2015, 2016) and are compatible with an experience-driven increase in selectivity of neurons in the AWFA to real words because of extensive auditory experience with and the need to discriminate among real words but not pseudowords.

To test our hypotheses, we ran a 2-way repeated measures analysis of variance (ANOVA) on AWFA responses to investigate the relationship between lexicality (RW and UTPW) and word similarity (SAME, 1PH, and DIFF). The ANOVA was run on subjects that successfully completed both the RW and UTPW scans ($n = 24$). This revealed a significant main effect of similarity ($F_{2,46} = 42.597$; $p = 3.4\text{E}-11$) but no significant main effect of lexicality ($F_{1,23} = 2.547$; $p = 0.124$). Critically, however, the analysis revealed a significant interaction between lexicality and word similarity ($F_{2,46} = 4.092$; $p = 0.023$). Planned paired t tests revealed an adaptation profile in the AWFA that was consistent with tight neural tuning for individual RWs (Figure 3A): There was a significant difference in the mean percent signal between the DIFF vs. SAME conditions ($t(23) = 7.86$; $p < 0.001$) and the 1PH vs. SAME ($t(23) = 5.71$; $p < 0.001$). However, there was no significant difference between the DIFF and 1PH conditions ($t(23) = 1.95$; $p = 0.189$). Similar results were obtained using all 26 subjects who completed the RW scan (Figure S1 in the Supporting Information). In contrast, the adaptation profile for UTPWs was not consistent with lexical selectivity (Figure 3B): There was a significant response difference between the DIFF vs. SAME conditions ($t(23) = 7.03$; $p < 0.001$), the 1PH vs. SAME ($t(23) = 3.28$; $p = 0.010$), and, critically, also for the DIFF vs. 1PH conditions ($t(23) = 3.89$; $p = 0.002$). Finally, we ran a whole-brain analysis to test whether lexical representations were localized to the anterior STG or more distributed in nature. Specifically, for RW the

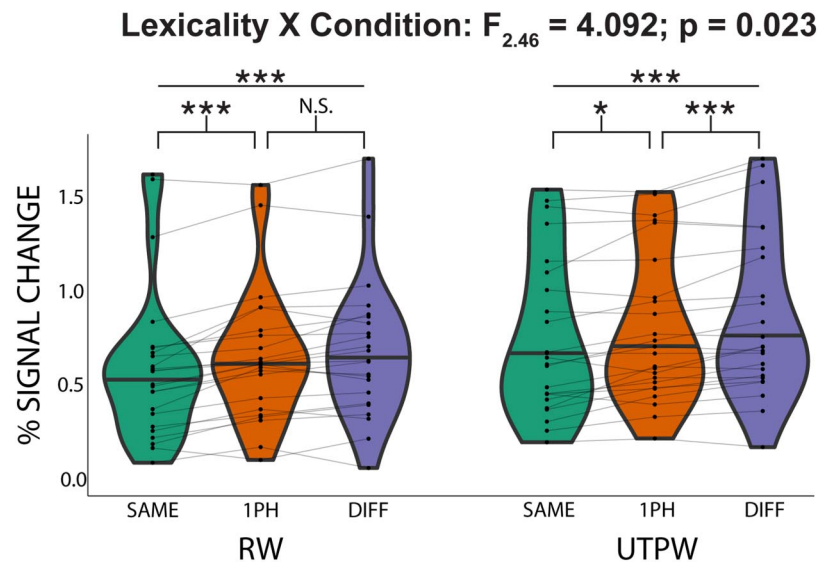


Figure 3. Evidence for auditory lexical representations in the auditory word form area. Within-subject ($n = 24$) adaptation profile for auditory real words (RWs) and untrained pseudowords (UTPWs). Patterns of release from adaptation are compatible with tight tuning to individual RWs consistent with an auditory lexicon. In contrast, UTPWs show a graded release from adaptation as a function of phonological similarity. Horizontal black line in violin plots indicates the median. ***, **, *, and N.S. mark $p < 0.001$, < 0.01 , < 0.05 , and not significant (> 0.1), all Bonferroni-corrected for multiple comparisons.

conjunction of DIFF vs. SAME and 1PH vs. SAME after excluding voxels where DIFF vs. PH was $p < 0.05$ produced a cluster in the left anterior STG (family-wise error corrected $p < 0.05$; MNI: $-62, -14, 6$) within 3.5 mm from the average coordinate (MNI: $-62, -14.9, 2.6$) of our individual ROIs. Importantly, this analysis for UTPWs produced no significant clusters. Thus, the whole-brain analysis also supports the special status of the AWFA as the location of a lexicon for spoken real words.

Planned paired t tests revealed an adaptation profile that was consistent with tight neural tuning for individual RWs. See Figure S1 in the Supporting Information available at https://doi.org/10.1162/nol_a_00108. There was a significant difference in the mean percent signal between the DIFF vs. SAME conditions ($t(25) = 6.94$; $p < 0.001$) and the 1PH vs. SAME ($t(25) = 5.36$; $p < 0.001$). However, there was no significant difference between the DIFF and 1PH conditions ($t(25) = 1.70$; $p = 0.3039$).

Adaptation Patterns to Pseudowords in the AWFA Exhibit Lexical Selectivity After but Not Before Familiarization

Next, in the pre- and post-training scans (UTPW and TPW, respectively), we tested the hypothesis that familiarization with previously novel PWs drives the formation of lexical selectivity in the AWFA. To do so, we examined the adaptation profiles for RWs, UTPWs, and TPWs in subjects who had completed all three scans ($n = 16$). We ran a 2-way repeated measures ANOVA to investigate the relationship between lexicality (RW, UTPW, and TPW) and similarity (SAME, 1PH, and DIFF). This revealed a significant main effect of similarity ($F_{2,30} = 43.023$; $p = 2.95E-9$) but not a significant main effect of lexicality ($F_{2,30} = 2.398$; $p = 0.116$). Critically, there was again a significant interaction between lexicality and similarity ($F_{4,60} = 4.144$; $p = 0.012$).

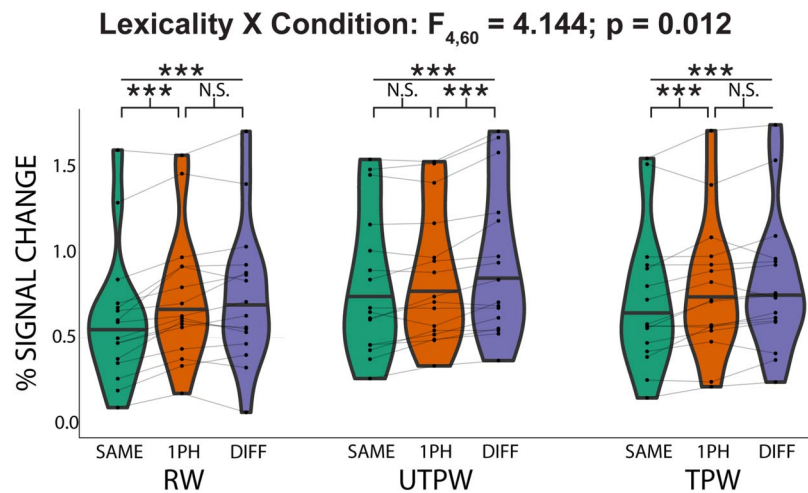


Figure 4. Auditory lexical representations emerge in the auditory word form area (AWFA) for pseudowords after familiarization training. Within-subject ($n = 16$) adaptation profile for auditory real words (RW), untrained pseudowords (UTPW), and trained pseudowords (TPW). RW adaptation profile shows tuning to individual RWs consistent with an auditory lexicon. UTPWs show a graded release from adaptation as a function of phonological similarity. Importantly, following familiarization training, adaptation patterns in the AWFA to the same pseudowords (now TPW) reveal tight lexical tuning, similar to RW. Horizontal black line in violin plots indicates the median. ***, **, *, and N.S. mark $p < 0.001$, < 0.01 , < 0.05 , and not significant (> 0.1), all Bonferroni-corrected.

Consistent with the full data set (Figure 3), planned paired t tests revealed an adaptation profile that was consistent with tight neural tuning for individual RWs (Figure 4). There was a significant difference in the mean percent signal change between the DIFF vs. SAME conditions ($t(15) = 7.28$; $p < 0.001$) and 1PH vs. SAME ($t(15) = 5.85$; $p < 0.001$). However, there was no significant difference between the DIFF and 1PH conditions ($t(15) = 0.750$; $p = 1$). In contrast, the adaptation profile for UTPW was not consistent with lexical selectivity (Figure 4): There was a significant difference between the DIFF vs. SAME conditions ($t(15) = 7.84$; $p < 0.001$) and the DIFF vs. 1PH conditions ($t(15) = 4.57$; $p = 0.001$), but not 1PH vs. SAME ($t(15) = 2.16$; $p = 0.141$). Crucially, the adaptation profile for PW *after* training (i.e., TPW) was consistent with lexical selectivity (Figure 4): There was a significant difference in the mean percent signal between the DIFF vs. SAME conditions ($t(15) = 5.55$; $p < 0.001$) and the 1PH vs. SAME ($t(15) = 3.17$; $p = 0.019$). However, there was no significant difference between the DIFF and 1PH conditions ($t(15) = 1.71$; $p = 0.327$).

The AWFA Connects to the Language Network

We next calculated task-based functional connectivity in the auditory localizer data set ($n = 26$) to examine the connections between the AWFA and the rest of the brain. This seed-to-voxel analysis (Figure 5) showed that the AWFA is highly connected with brain regions previously shown (Wilson et al., 2004) to be involved in language processing such as the inferior frontal gyrus (local peak at MNI $-46, 12, 26$) and premotor cortex (local peak at MNI $-50, -8, 48$). Especially noteworthy is the connectivity between the AWFA and a cluster in the left posterior fusiform cortex (circled in red; MNI: $-44, -46, -14$) that encompasses the reported location of the VWFA (MNI: $-45, -54, -20$; Dehaene et al., 2005; Kronbichler et al., 2004). Crucially, subjects never saw any words in print during the auditory localizer scan, thereby precluding the possibility of spurious correlations between these regions.

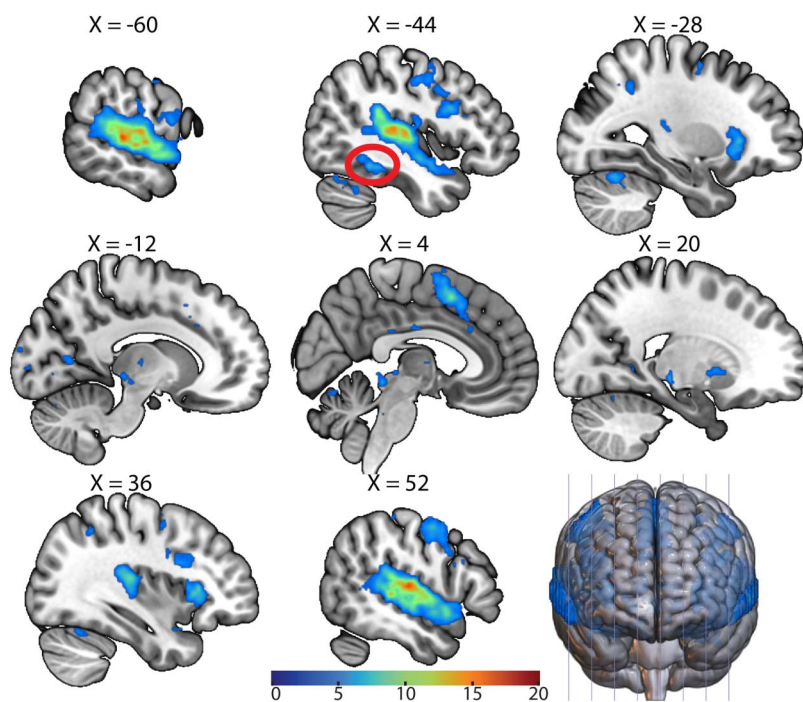


Figure 5. Functional connectivity of the auditory word form area (AWFA). Whole-brain functional connectivity of the AWFA during the auditory localizer task ($n = 26$). Results are thresholded at a voxel-wise $p < 0.001$ and cluster-level $p < 0.05$, family-wise error corrected. Cluster corresponding to the literature coordinates of the visual word form area is circled in red. Color bar represents t statistic.

DISCUSSION

Cognitive models of speech comprehension (Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Morton, 1969) have proposed the existence of an auditory lexicon. More recent models of speech comprehension in the brain (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009) have posited the existence of such an auditory lexicon in an auditory ventral *what* stream. Yet, significant disagreement exists between these models about the location of such an auditory lexicon, and no studies to date have directly tested the existence of an auditory lexicon for speech comprehension in the brain. Prior work (Glezer et al., 2009, 2015, 2016), has used fMRI RA techniques to establish the existence of an orthographic lexicon in a region of the visual ventral stream—the VWFA (subsequently confirmed by human neurophysiological recordings; Hirshorn et al., 2016; Lochy et al., 2018; Woolnough et al., 2021). In the present study, we leveraged these techniques to test the existence of an auditory lexicon in the auditory ventral stream, particularly in the AWFA of the anterior STG (Cohen et al., 2004; DeWitt & Rauschecker, 2012). We first defined the ROI for an individual AWFA through an independent auditory localizer scan. We then showed in RW and UTPW RA scans that, consistent with an auditory lexicon, spoken RWs engaged distinct neural populations in the AWFA, with even a single phoneme change causing full release from adaptation, whereas PWs did not exhibit this lexical adaptation profile but instead showed a graded release from adaptation as a function of phoneme overlap. This graded release from adaptation suggests that neurons in the AWFA, like the VWFA (Glezer et al., 2009), exhibit broader tuning to PWs due to experience-driven refinement of tuning of neurons to RWs but not PWs. Thus, these results directly replicated analogous findings of highly selective lexical tuning to RWs

but not PWs in the VWFA (Glezer et al., 2009). Then, in the TPW RA scan, we showed that training subjects to recognize a set of PWs led to the development of lexical selectivity for these TPWs, again replicating previous results for written words in the VWFA (Glezer et al., 2015). Finally, the AWFA was connected to an area in the left posterior fusiform cortex coincident with the reported location of the VWFA, further supporting analogous roles of the AWFA and VWFA in the processing of spoken and written words, respectively.

Our novel evidence of auditory lexical representations in the brain informs current cognitive models of speech comprehension. While all these models map acoustic input to meaning, they disagree on whether auditory lexica exist (Coltheart, 2004; Woollams, 2015). Our data not only present compelling evidence for the existence of an auditory lexicon—they also place its location in the anterior STG where it is ideally suited to interface with semantic representations located further anteriorly in the temporal lobe (Lambon Ralph et al., 2017; Ueno et al., 2011), thereby completing the mapping of speech sounds to meaning. Furthermore, this anterior STG location is consistent with prior studies demonstrating other familiar auditory objects (Griffiths & Warren, 2004), such as human voices (Belin et al., 2000; Bodin et al., 2021; Staib & Frühholz, 2023) or musical instruments (Leaver & Rauschecker, 2010). Moreover, such a simple-to-complex progression in selectivity from simple perceptual features over lexical representations to semantic representations in the anteroventral auditory processing stream along the STC is a direct counterpart of the ventral visual stream in the inferior temporal cortex (Damera et al., 2020; Kravitz et al., 2013), revealing convergent processing strategies across speech modalities.

FUNDING INFORMATION

Maximilian Riesenhuber, National Science Foundation (<https://dx.doi.org/10.13039/1000000001>), Award ID: BCS-1756313. Maximilian Riesenhuber, National Science Foundation (<https://dx.doi.org/10.13039/1000000001>), Award ID: ACI-1548562. Ashley VanMeter, Foundation for the National Institutes of Health (<https://dx.doi.org/10.13039/1000000009>), Award ID: 1S10OD023561.

AUTHOR CONTRIBUTIONS

Srikanth R. Damera: Conceptualization; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – reviewing & editing. **Lillian Chang:** Investigation; Writing – reviewing & editing. **Plamen P. Nikolov:** Investigation; Writing – reviewing & editing. **James A. Mattei:** Investigation; Writing – reviewing & editing. **Suneel Banerjee:** Investigation; Writing – reviewing & editing. **Laurie S. Glezer:** Conceptualization; Methodology. **Patrick H. Cox:** Conceptualization; Methodology. **Xiong Jiang:** Conceptualization; Methodology; Project administration; Writing – reviewing & editing. **Josef P. Rauschecker:** Conceptualization; Project administration; Writing – reviewing & editing. **Maximilian Riesenhuber:** Conceptualization; Methodology; Project administration; Writing – reviewing & editing.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at <https://osf.io/pbuh5/>.

REFERENCES

- Archakov, D., DeWitt, I., Kuśmierk, P., Ortiz-Rios, M., Cameron, D., Cui, D., Morin, E. L., VanMeter, J. W., Sams, M., Jääskeläinen, I. P., & Rauschecker, J. P. (2020). Auditory representation of learned sound sequences in motor regions of the macaque brain. *Proceedings of the National Academy of Sciences*, 117(26), 15242–15252. <https://doi.org/10.1073/pnas.1915610117>, PubMed: 32541016
- Ashburner, J., Barnes, G., Chen, C.-C., Daunizeau, J., Flandin, G., Friston, K., Gitelman, D., Glauche, V., Henson, R., Hutton, C.,

- Jafarian, A., Kiebel, S., Kilner, J., Litvak, V., Mattout, J., Moran, R., Penny, W., Phillips, C., Razi, A., ... Zeidman, P. (2021). *SPM12 manual*. Wellcome Centre for Human Neuroimaging.
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459. <https://doi.org/10.3758/BF03193014>, PubMed: 17958156
- Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1), 90–101. <https://doi.org/10.1016/j.neuroimage.2007.04.042>, PubMed: 17560126
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767), 309–312. <https://doi.org/10.1038/35002078>, PubMed: 10659849
- Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10(5), 512–528. <https://doi.org/10.1093/cercor/10.5.512>, PubMed: 10847601
- Bodin, C., Trapeau, R., Nazarian, B., Sein, J., Degiovanni, X., Baurberg, J., Rapha, E., Renaud, L., Giordano, B. L., & Belin, P. (2021). Functionally homologous representation of vocalizations in the auditory cortex of humans and macaques. *Current Biology*, 31(21), 4839–4844. <https://doi.org/10.1016/j.cub.2021.08.043>, PubMed: 34506729
- Bogen, J. E., & Bogen, G. M. (1976). Wernicke's region—Where is it? *Annals of the New York Academy of Sciences*, 280(1), 834–843. <https://doi.org/10.1111/j.1749-6632.1976.tb25546.x>, PubMed: 1070943
- Brown, J. C. (1991). Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, 89(1), 425–434. <https://doi.org/10.1121/1.400476>
- Buračas, G. T., & Boynton, G. M. (2002). Efficient design of event-related fMRI experiments using M-sequences. *NeuroImage*, 16(3, Pt. A), 801–813. <https://doi.org/10.1006/nimg.2002.1116>, PubMed: 12169264
- Chevillet, M. A., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2013). Automatic phoneme category selectivity in the dorsal auditory stream. *Journal of Neuroscience*, 33(12), 5208–5215. <https://doi.org/10.1523/JNEUROSCI.1870-12.2013>, PubMed: 23516286
- Cohen, L., Jobert, A., Bihan, D. L., & Dehaene, S. (2004). Distinct unimodal and multimodal regions for word processing in the left temporal cortex. *NeuroImage*, 23(4), 1256–1270. <https://doi.org/10.1016/j.neuroimage.2004.07.052>, PubMed: 15589091
- Coltheart, M. (2004). Are there lexicons? *Quarterly Journal of Experimental Psychology Section A*, 57(7), 1153–1172. <https://doi.org/10.1080/027249804430000007>, PubMed: 15513241
- Damera, S. R., Malone, P. S., Stevens, B. W., Klein, R., Eberhardt, S. P., Auer, E. T., Bernstein, L. E., & Riesenhuber, M. (2021). Metamodal coupling of vibrotactile and auditory speech processing systems through matched stimulus representations. *bioRxiv*. <https://doi.org/10.1101/2021.05.04.442660>
- Damera, S. R., Martin, J. G., Scholl, C., Kim, J. S., Glezer, L., Malone, P. S., & Riesenhuber, M. (2020). From shape to meaning: Evidence for multiple fast feedforward hierarchies of concept processing in the human brain. *NeuroImage*, 221, Article 117148. <https://doi.org/10.1016/j.neuroimage.2020.117148>, PubMed: 32659350
- Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6), 254–262. <https://doi.org/10.1016/j.tics.2011.04.003>, PubMed: 21592844
- Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, 9(7), 335–341. <https://doi.org/10.1016/j.tics.2005.05.004>, PubMed: 15951224
- DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, 109(8), E505–E514. <https://doi.org/10.1073/pnas.1113427109>, PubMed: 22308358
- DeWitt, I., & Rauschecker, J. P. (2013). Wernicke's area revisited: Parallel streams and word processing. *Brain and Language*, 127(2), 181–191. <https://doi.org/10.1016/j.bandl.2013.09.014>, PubMed: 24404576
- Evans, S., Kyong, J. S., Rosen, S., Golestani, N., Warren, J. E., McGettigan, C., Mourão-Miranda, J., Wise, R. J. S., & Scott, S. K. (2014). The pathways for intelligible speech: Multivariate and univariate perspectives. *Cerebral Cortex*, 24(9), 2350–2361. <https://doi.org/10.1093/cercor/bht083>, PubMed: 23585519
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47. <https://doi.org/10.1093/cercor/1.1.1-a>, PubMed: 1822724
- Geschwind, N. (1970). The organization of language and the brain: Language disorders after brain damage help in elucidating the neural basis of verbal behavior. *Science*, 170(3961), 940–944. <https://doi.org/10.1126/science.170.3961.940>, PubMed: 5475022
- Glezer, L. S., Eden, G., Jiang, X., Luetje, M., Napoliello, E., Kim, J., & Riesenhuber, M. (2016). Uncovering phonological and orthographic selectivity across the reading network using fMRI-RA. *NeuroImage*, 138, 248–256. <https://doi.org/10.1016/j.neuroimage.2016.05.072>, PubMed: 27252037
- Glezer, L. S., Jiang, X., & Riesenhuber, M. (2009). Evidence for highly selective neuronal tuning to whole words in the “visual word form area.” *Neuron*, 62(2), 199–204. <https://doi.org/10.1016/j.neuron.2009.03.017>, PubMed: 19409265
- Glezer, L. S., Kim, J., Rule, J., Jiang, X., & Riesenhuber, M. (2015). Adding words to the brain's visual dictionary: Novel word learning selectively sharpens orthographic representations in the VWFA. *Journal of Neuroscience*, 35(12), 4965–4972. <https://doi.org/10.1523/JNEUROSCI.4031-14.2015>, PubMed: 25810526
- Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5(11), 887–892. <https://doi.org/10.1038/nrn1538>, PubMed: 15496866
- Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, 107(1–3), 293–321. [https://doi.org/10.1016/S0001-6918\(01\)00019-1](https://doi.org/10.1016/S0001-6918(01)00019-1), PubMed: 11388140
- Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, 28(12), 1860–1871. <https://doi.org/10.1016/j.cub.2018.04.033>, PubMed: 29861132
- Henson, R., Shallice, T., & Dolan, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science*, 287(5456), 1269–1272. <https://doi.org/10.1126/science.287.5456.1269>, PubMed: 10678834
- Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron*, 69(3), 407–422. <https://doi.org/10.1016/j.neuron.2011.01.019>, PubMed: 21315253
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), 393–402. <https://doi.org/10.1038/nrn2113>, PubMed: 17431404
- Hirshorn, E. A., Li, Y., Ward, M. J., Richardson, M. R., Fiez, J. A., & Ghuman, A. (2016). Decoding and disrupting left midfusiform

- gyrus activity during word reading. *Proceedings of the National Academy of Sciences*, 113(29), 8162–8167. <https://doi.org/10.1073/pnas.1604126113>, PubMed: 27325763
- Hubel, D. H., & Wiesel, T. N. (1977). Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 198(1130), 1–59. <https://doi.org/10.1098/rspb.1977.0085>, PubMed: 20635
- Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, 36(6), 2014–2026. <https://doi.org/10.1523/JNEUROSCI.1779-15.2016>, PubMed: 26865624
- Jasmin, K., Lima, C. F., & Scott, S. K. (2019). Understanding rostral–caudal auditory cortex contributions to auditory perception. *Nature Reviews Neuroscience*, 20(7), 425–434. <https://doi.org/10.1038/s41583-019-0160-2>, PubMed: 30918365
- Jiang, X., Chevillet, M. A., Rauschecker, J. P., & Riesenhuber, M. (2018). Training humans to categorize monkey calls: Auditory feature- and category-selective neural tuning changes. *Neuron*, 98(2), 405–416. <https://doi.org/10.1016/j.neuron.2018.03.014>, PubMed: 29673483
- Kajikawa, Y., Frey, S., Ross, D., Falchier, A., Hackett, T. A., & Schroeder, C. E. (2015). Auditory properties in the parabelt regions of the superior temporal gyrus in the awake macaque monkey: An initial survey. *Journal of Neuroscience*, 35(10), 4140–4150. <https://doi.org/10.1523/JNEUROSCI.3556-14.2015>, PubMed: 25762661
- Kell, A., Yamins, D., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3), 630–644. <https://doi.org/10.1016/j.neuron.2018.03.044>, PubMed: 29681533
- Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17(1), 26–49. <https://doi.org/10.1016/j.tics.2012.10.011>, PubMed: 23265839
- Krekelberg, B., Boynton, G. M., & van Wezel, R. J. A. (2006). Adaptation: From single cells to BOLD signals. *Trends in Neurosciences*, 29(5), 250–256. <https://doi.org/10.1016/j.tins.2006.02.008>, PubMed: 16529826
- Kronbichler, M., Hutzler, F., Wimmer, H., Mair, A., Staffen, W., & Ladurner, G. (2004). The visual word form area and the frequency with which words are encountered: Evidence from a parametric fMRI study. *NeuroImage*, 21(3), 946–953. <https://doi.org/10.1016/j.neuroimage.2003.10.021>, PubMed: 15006661
- Laird, A. R., Fox, P. M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., Turkeltaub, P. E., Kochunov, P., & Fox, P. T. (2005). ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping*, 25(1), 155–164. <https://doi.org/10.1002/hbm.20136>, PubMed: 15846811
- Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1), 42–55. <https://doi.org/10.1038/nrn.2016.150>, PubMed: 27881854
- Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *Journal of Neuroscience*, 30(22), 7604–7612. <https://doi.org/10.1523/JNEUROSCI.0296-10.2010>, PubMed: 20519535
- Liuzzi, A. G., Bruffaerts, R., & Vandenberghe, R. (2019). The medial temporal written word processing system. *Cortex*, 119, 287–300. <https://doi.org/10.1016/j.cortex.2019.05.002>, PubMed: 31174078
- Lochy, A., Jacques, C., Maillard, L., Colnat-Coulbois, S., Rossion, B., & Jonas, J. (2018). Selective visual representation of letters and words in the left ventral occipito-temporal cortex with intracerebral recordings. *Proceedings of the National Academy of Sciences*, 115(32), E7595–E7604. <https://doi.org/10.1073/pnas.1718987115>, PubMed: 30038000
- Malone, P. S., Glezer, L. S., Kim, J., Jiang, X., & Riesenhuber, M. (2016). Multivariate pattern analysis reveals category-related organization of semantic representations in anterior temporal cortex. *Journal of Neuroscience*, 36(39), 10089–10096. <https://doi.org/10.1523/JNEUROSCI.1599-16.2016>, PubMed: 27683905
- Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1), 29–63. [https://doi.org/10.1016/0010-0285\(78\)90018-X](https://doi.org/10.1016/0010-0285(78)90018-X)
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86. [https://doi.org/10.1016/0010-0285\(86\)90015-0](https://doi.org/10.1016/0010-0285(86)90015-0)
- Medler, D. A., & Binder, J. R. (2005). *MCWord: An on-line orthographic database of the English language*. Language Imaging Laboratory, Medical College of Wisconsin. <https://www.neuro.mcw.edu/mcword/>
- Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76(2), 165–178. <https://doi.org/10.1037/h0027366>
- Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, 20(10), 2486–2495. <https://doi.org/10.1093/cercor/bhp318>, PubMed: 20100898
- Ortiz-Rios, M., Kušmírek, P., DeWitt, I., Archakov, D., Azevedo, F. A. C., Sams, M., Jääskeläinen, I. P., Keliris, G. A., & Rauschecker, J. P. (2015). Functional MRI of the vocalization-processing network in the macaque brain. *Frontiers in Neuroscience*, 9, Article 113. <https://doi.org/10.3389/fnins.2015.00113>, PubMed: 25883546
- Perrachione, T. K., & Ghosh, S. S. (2013). Optimized design and analysis of sparse-sampling fmri experiments. *Frontiers in Neuroscience*, 7, Article 55. <https://doi.org/10.3389/fnins.2013.00055>, PubMed: 23616742
- Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, 8(4), 516–521. [https://doi.org/10.1016/S0959-4388\(98\)80040-8](https://doi.org/10.1016/S0959-4388(98)80040-8), PubMed: 9751652
- Rauschecker, J. P. (2011). An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hearing Research*, 271(1–2), 16–25. <https://doi.org/10.1016/j.heares.2010.09.001>, PubMed: 20850511
- Rauschecker, J. P. (2018). Where, when, and how: Are they all sensorimotor? Towards a unified view of the dorsal pathway in vision and audition. *Cortex*, 98, 262–268. <https://doi.org/10.1016/j.cortex.2017.10.020>, PubMed: 29183630
- Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6), 718–724. <https://doi.org/10.1038/nn.2331>, PubMed: 19471271
- Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings*

- of the National Academy of Sciences, 97(22), 11800–11806. <https://doi.org/10.1073/pnas.97.22.11800>, PubMed: 11050212
- Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207), 111–114. <https://doi.org/10.1126/science.7701330>, PubMed: 7701330
- Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123(12), 2400–2406. <https://doi.org/10.1093/brain/123.12.2400>, PubMed: 11099443
- Staib, M., & Frühholz, S. (2023). Distinct functional levels of human voice processing in the auditory cortex. *Cerebral Cortex*, 33(4), 1170–1185. <https://doi.org/10.1093/cercor/bhac128>, PubMed: 35348635
- Tian, B., Reser, D., Durham, A., Kustov, A., & Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science*, 292(5515), 290–293. <https://doi.org/10.1126/science.1058911>, PubMed: 11303104
- Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (2011). Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, 72(2), 385–396. <https://doi.org/10.1016/j.neuron.2011.09.013>, PubMed: 22017995
- van der Heijden, K., Rauschecker, J. P., de Gelder, B., & Formisano, E. (2019). Cortical mechanisms of spatial hearing. *Nature Reviews Neuroscience*, 20(10), 609–623. <https://doi.org/10.1038/s41583-019-0206-5>, PubMed: 31467450
- Vinckier, F., Dehaene, S., Jobert, A., Dubus, J., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, 55(1), 143–156. <https://doi.org/10.1016/j.neuron.2007.05.031>, PubMed: 17610823
- Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, 2(3), 125–141. <https://doi.org/10.1089/brain.2012.0073>, PubMed: 22642651
- Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7(7), 701–702. <https://doi.org/10.1038/nn1263>, PubMed: 15184903
- Wollams, A. M. (2015). Lexical is as lexical does: Computational approaches to lexical representation. *Language, Cognition and Neuroscience*, 30(4), 395–408. <https://doi.org/10.1080/23273798.2015.1005637>, PubMed: 25893204
- Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, 5(3), 389–398. <https://doi.org/10.1038/s41562-020-00982-w>, PubMed: 33257877