JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Systems/Circuits

Metamodal Coupling of Vibrotactile and Auditory Speech Processing Systems Through Matched Stimulus Representations

https://doi.org/10.1523/JNEUROSCI.1710-22.2023

Cite as: J. Neurosci 2023; 10.1523/JNEUROSCI.1710-22.2023

Received: 8 September 2022
Revised: 10 March 2023
Accepted: 29 April 2023

Title: Metamodal Coupling of Vibrotactile and Auditory Speech Processing Systems Through Matched Stimulus Representations

Running Title: Metamodal Coupling Through Matched Representations

Srikanth R. Damera[1], Patrick S. Malone[1], Benson W. Stevens[1], Richard Klein[1], Silvio P. Eberhardt[2], Edward T. Auer Jr.[2], Lynne E. Bernstein[2], Maximilian Riesenhuber[1]

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

2 - Department of Speech Language & Hearing Sciences, George Washington University, Washington DC, USA

Correspondence should be addressed to M.R. (mr287@georgetown.edu)

Department of Neuroscience

Georgetown University Medical Center

Research Building Room WP-12

3970 Reservoir Rd. NW

Washington, DC 20007, USA

Figure Count

7 Figures

Word Count

Abstract: 250

Significance Statement: 119

Introduction: 648

Discussion: 1495

Acknowledgements: 73

Competing Interests

The authors declare no competing financial interests

Acknowledgments

This work was supported in part by Facebook, and National Science Foundation BCS-1439338 and BCS-1756313. We thank Ali Israr, Frances Lau, Keith Klumb, Robert Turcott, and Freddy Abnousi for their involvement in the early stages of the project, including the design and evaluation of the token-based algorithm; and Dr. Ella Striem-Amit for helpful feedback on earlier versions of this manuscript.

Abstract

It has been postulated that the brain is organized by “metamodal”, sensory-independent cortical modules capable of performing tasks (e.g., word recognition) in both “standard” and novel sensory modalities. Still, this theory has primarily been tested in sensory-deprived individuals, with mixed evidence in neurotypical subjects, thereby limiting its support as a general principle of brain organization. Critically, current theories of metamodal processing do not specify requirements for successful metamodal processing at the level of neural representations. Specification at this level may be particularly important in neurotypical individuals, where novel sensory modalities must interface with existing representations for the standard sense. Here we hypothesized that effective metamodal engagement of a cortical area requires congruence between stimulus representations in the standard and novel sensory modalities in that region. To test this, we first used fMRI to identify bilateral auditory speech representations. We then trained 20 human participants (12 female) to recognize vibrotactile versions of auditory words using one of two auditory-to-vibrotactile algorithms. The vocoded algorithm preserved the dynamics and representational similarities of auditory speech while the token-based algorithm did not. Crucially, using fMRI, we found that only in the vocoded group did trained-vibrotactile stimuli recruit speech representations in the superior temporal gyrus and lead to increased coupling between them and somatosensory areas. Our results advance our understanding of brain organization by providing new insight into unlocking the metamodal potential of the brain, thereby benefitting the design of novel sensory substitution devices that aim to tap into existing processing streams in the brain.

Significance Statement

It has been proposed that the brain is organized by “metamodal”, sensory-independent modules specialized for performing certain tasks. This idea has inspired therapeutic applications such as sensory substitution devices, e.g., enabling blind individuals “to see” by transforming visual input into soundscapes. Yet, other studies have failed to demonstrate metamodal engagement. Here, we tested the hypothesis that metamodal engagement in neurotypical individuals requires matching representational similarities between stimuli in novel and standard modalities. We trained two groups of subjects to recognize words generated by one of two auditory-to-vibrotactile transformations. Critically, only vibrotactile stimuli that preserved representational similarities of auditory speech engaged auditory speech areas after training. This suggests that matching representational similarities is critical to unlocking the brain’s metamodal potential.

Introduction

The dominant view of brain organization revolves around cortical areas dedicated for processing information from specific sensory modalities. However, emerging evidence over the past two decades has led to the idea that cortical areas are defined by task-specific computations that are invariant to sensory modality(Pascual-Leone and Hamilton, 2001). Evidence for this comes from studies in sensory-deprived populations (Sadato et al., 1996; Lomber et al., 2010; Bola et al., 2017) which show that areas that traditionally perform unisensory processing can be recruited by stimuli from another sensory modality to perform the same task. This ability of a novel sensory modality stimuli to engage a cortical area the same way as the standard sensory modality stimulus is called metamodal engagement. Importantly, there is evidence (Renier et al., 2005, 2010; Amedi et al., 2007; Siuda-Krzywicka et al., 2016) for metamodal engagement of traditionally unisensory areas even in neurotypical individuals – thereby opening the door for novel sensory modalities to recruit established sensory processing pathways. This idea has given rise to promising therapeutic applications like sensory substitution devices (SSDs). These devices can, for instance, enable blind individuals to process visual information by translating camera input to sounds (Meijer, 1992; Bach-y-Rita and Kercel, 2003). Still, other studies (Fairhall et al., 2017; Twomey et al., 2017; Benetti et al., 2020; Mattioni et al., 2020; Vetter et al., 2020) failed to find or found less robust evidence of cross-modal engagement in neurotypical subjects. This calls into question the conditions under which a cortical area can be successfully recruited by stimuli from a novel sensory modality.

Current theories emphasize that metamodal engagement of a cortical area, depends on a task-level correspondence irrespective of the stimulus modality and the presence of task-relevant connectivity (Heimler et al., 2015). Thus, metamodal theories are specified at the level of computation (i.e., shared task) and implementation (i.e., sufficient connectivity) – the first and third levels of Marr's levels of analysis (Marr, 1982). However, consideration of these two levels alone cannot explain the failure of certain studies to find metamodal engagement. We argue that metamodal engagement depends on not just an abstract correspondence between standard and novel modality stimuli, but also on a correspondence between their encoding in the target area. This correspondence at the level of encoding corresponds to Marr's second level, the algorithmic level. For instance, since auditory cortex in neurotypical adults is sensitive to the temporal dynamics of auditory speech (Yi et al., 2019; Penikis and Sanes, 2022), metamodal engagement of this area

by novel modality stimuli depends on their ability to match the temporal dynamics of spoken words. Failure to do so may favor alternate learning mechanisms such as paired associate learning (McClelland et al., 1995; Eichenbaum et al., 1996).

In the present study, we tested the hypothesis that metamodal engagement of a brain area in neurotypical individuals depends on matching the encoding schemes between stimuli from the novel and standard sensory modalities. We used functional MRI (fMRI) data from an independent auditory scan to identify target auditory speech areas for metamodal engagement in the bilateral STG. We then built on prior behavioral studies to train two groups of neurotypical adults to recognize words using one of two auditory-to-VT sensory substitution algorithms. Critically, while both algorithms preserved behavioral word similarities, one encoding ("vocoded") closely matched the temporal dynamics of auditory speech whereas the other ("token-based") did not. Our results show that while subjects in both algorithm groups learned to accomplish the word recognition task equally well, only those trained on the similarity-preserving vocoded VT representation exhibited metamodal engagement of the bilateral STG. Consistent with these findings, only subjects in the vocoded VT group exhibited increased functional connectivity between the auditory and somatosensory cortex after training. These findings suggest that metamodal engagement of a cortical area in neurotypical adults depends not only on a correspondence between standard and novel modality stimuli at the task-level but also at the neural representational level.

Materials and Methods

Participants

We recruited a total of 22, right-handed, healthy, native English speakers in this study (ages 18-27, 12 females). Georgetown University's Institutional Review Board approved all experimental procedures, and written informed consent was obtained from all subjects before the experiment. We excluded 4 subjects from the auditory scan due to excessive motion (>20% of volumes) resulting in a total of 18 subjects. In the vibrotactile (VT) scans, subjects were alternately assigned to one of the two VT algorithm groups (see below), resulting in 11 subjects per group. However, 2 of the 22 subjects, 1 from each VT algorithm group, were excluded because they failed to complete the training. Thus, a total of 20 subjects were analyzed for the VT scans (10 per group). An effect-size sensitivity analysis was performed using an α of $p = 0.05$, power of 0.8, and a two-tailed one-sample or two-sample t-test for auditory and VT scans respectively using G*Power (Faul et al., 2007). This calculation yielded a minimum detectable effect size of 0.7 and 0.99 for the auditory and VT scans respectively.

Stimuli and Materials

Stimulus Selection

A set of word stimuli (Tbl. 1) was developed according to the following criteria: 1) short monosyllabic stimuli (~4 phonemes); 2) only contain phonemes from a limited subset of English consonants (8 consonants and 6 vowels); 3) set containing items predicted to be perceptually unique and therefore learnable; and 4) words that span the VT vocoder perceptual space (see below). To develop the set meeting these criteria we utilized a computational modeling approach based on the methods described in (Auer and Bernstein, 1997). Existing tactile consonant and vowel perceptual identification data (Bernstein, unpublished) were used in combination with the PhLex lexical database (Seitz, Bernstein, Auer, & MacEachern, 1998) to model the lexical perceptual space. In outline, the modeling steps are: (1) Transform phoneme identification data into groupings of phonemes as a function of a set level of dissimilarity; (2) Re-transcribe a phonemically transcribed lexical database so that all the words are represented in terms of only the phonemic distinctions across groupings; and (3) Collect words that are identical under the re-transcription and count how many are in each collection. In

this study, the lexical equivalence class (LEC) size –the number of words in a collection—was set to three. Only words that were accompanied by three or fewer other words following re-transcription were considered candidates for the study. Words in smaller LECs are predicted to be perceptually easier (more unique) than words in larger LECs, which offer more opportunities for confusions.

The set of words meeting the first three criteria was further examined as a function of consonants and vowel patterns to identify the largest pool of potential stimulus words. Three consonant (C) and vowel (V) segment patterns (CVC, CCVC, and CVCC) were selected for the final stimulus set. The words with these segment patterns were then examined in relation to the predicted VT vocoder perceptual space. The tactile identification confusion matrices were transformed into phoneme distance matrices using a phi-square transform (Iverson et al., 1998). Within a segment pattern, all word-to-word distances were computed as the sum of the pairwise phoneme distances. The word distance matrix was then submitted to multidimensional scaling to facilitate two-dimensional visualization of the lexical space. Close pairs were selected with goal of achieving distributed coverage in each of the three lexical spaces (CVC, CVCC, and CCVC). For each close pair, a third more distant word was chosen that provided a bridge to other pairs in the space. Final selection was based on the word-to-word computed distances using phi-square distances rather than the multidimensional space as clear warping was present due to the reduction of dimensionality.

This resulted in 60 total words or 20 sets of triples. We trained subjects to associate 30 words (10 triplets) with their corresponding VT tokens. In the fMRI scans we used 15 (5 triplets) of these trained words of which 9 belonged to the CVCC, 3 to the CCVC and 3 to the CVC lexical classes (Tbl. 1).

Description of VT Stimulus System

Custom hardware and software was used to present the VT stimuli. The system that vocoded the acoustic speech for VT stimuli had been developed and used previously for real time speech analysis and stimulus presentation (Bernstein et al., 1991; Iverson et al., 1998; Eberhardt et al., 2014). The vocoder filters were as described as the “GULin” vocoder algorithm (Bernstein et al., 1991). Their vocoder implemented 15 sixth-order bandpass filters with frequencies centered at 260, 392, 525, 660, 791, 925, 1060, 1225, 1390, 1590, 1820, 2080, 2380, 2720, and 3115 Hz, with respective bandwidths of 115, 130, 130, 130, 130, 130, 145, 165,190, 220, 250, 290, 330, 375, and 435 Hz. The 16th channel was a high-pass filter with a 3565-Hz cutoff. Because

the energy in speech rolls off by 6 dB per octave, and because the skin has a narrow dynamic range, the energy passed by each of the filters was linearly scaled, resulting in a good representation of the speech formant patterns across a range of signal intensities (Bernstein et al., 1991). Because new MRI-compatible hardware was built for the current study, and therefore new driver software was needed, the vibrotactile drive signals of the old vocoder system were sampled to drive the new system (thus guaranteeing that the underlying acoustic analysis remained the same), to maintain timing and output channel information. Then at presentation time, prior to each stimulus, the stimulus timing record that specified the onset time of each pulse on each channel was uploaded to the VT control system.

The hardware transducer was an updated version of the one used in Malone et al., (Malone et al., 2019). The stimulator hardware comprised piezoelectric bimorph transducers (Fig. 1A). During operation, a constant +57-V applied to all stimulators retracted the contactors into the surround, and each applied -85-V pulse drove the contactor into the skin. The display's control system comprised the power supplies (-85V, +57V), high voltage switching circuits to apply these voltages to the piezoelectric bimorphs, and a digital control system that accepted from a controlling computer's serial COM port the digital records specifying a stimulus (comprising the times and channels to output pulses on), and a command to initiate stimulus output. All pulses were identical. The drive signal was a square wave, with a pulse time of 2 ms and a maximum pulse rate of 150 pulses per second.

The 16-channel (20cm x 11.0 cm) array was organized as 2 rows of 8 stimulators (Fig. 1A), with center-to-center stimulator spacing of 2.54 cm, which was worn on the volar forearm. This spacing is greater than the average distance on the volar forearm at which participants achieved at least 95% correct discrimination on a tactile spatial acuity task (Tong et al., 2013). Transducers were arranged so that similar frequencies were mapped to similar locations along the forearm. Specifically, low frequencies mapped to transducers near the wrist, and higher frequencies mapped to transducers near the elbow (Fig. 1C). To ensure that the stimulators would maintain contact with the volar forearm, the transducer array comprised four rigid modules connected with stiff plastic springs. Velcro straps were used to mount the device to the arm firmly while bending the array to conform to the arm's shape. With no applied voltage to the piezoelectric bimorphs, the contactors were flush with the circuit board surface facing the skin.

Token-based VT Speech Encoding

The same 16-channel VT device was used to present subjects with the token-based stimuli. Token-based stimuli were constructed based on prior work (Reed et al., 2018) and reflected the idea that spoken words can be described as a string of phonemes. Phonemes in turn can be uniquely described by a set of phonetic features. Therefore, each phonetic feature was assigned a unique VT pattern. In this study, we used place, manner, and voicing features to describe phonemes (Fig. 1B). Place was coded as patterns that occurred either proximal or distal to the wrist. Stop and fricative manner features were codded as patterns that occurred either medial or lateral to the body respectively. The nasal manner feature was distinguished by driving two channels instead of one for stops and fricatives. Voicing was coded as either driving high frequency vibrations (250Hz) or low frequency vibrations (100Hz). Vowels were coded in a similar feature-based manner, but were dynamic stimuli (e.g., swirls and sweeps) whereas consonants were static. Importantly, all consonant patterns lasted 120ms and all vowel stimuli lasted 220ms and there was a 100ms gap between each pattern. As a result, token-based stimuli were either 660ms or 880ms long. CVCC trained token-based stimuli used in fMRI analyses were 880ms long while their VT vocoded counterparts had a mean duration of 727ms and standard deviation of 91.6ms. A paired t-test revealed that token-based stimuli were significantly longer ($t_8 = 4.99$; $p = 0.001$) than their vocoded counterparts. Thus, not only did VT vocoded but not token based stimuli preserve the temporal dynamics found in auditory speech, but they also conveyed more information per unit time.

Experimental Design

In the current study, subjects participated in two pre-training fMRI sessions and one post-training session upon successful completion of 6 behavioral training sessions. The final post-training session was followed by a 10-AFC experiment to assess if subjects retained the trained associations between VT stimuli and the words. The two pre-training fMRI sessions consisted of an auditory scan followed by a VT scan and were done on separate days. After the pre-training VT scan subjects performed 6 sessions of behavioral training in which they learned to associate patterns of vibrotactile stimulation with words. Subjects could only perform one training session per day. A subject was considered to have successfully completed the training and thus was eligible for the post-training scan if he or she completed all 6 training sessions.

Behavioral Training

Subjects performed a total of 6 training sessions and could only perform 1 training session per day. Each session took place in a quiet room while the subject was seated and listened to an auditory white noise stimulus through over-the-ear headphones. Auditory white noise was presented to mask the mechanical sound of the VT stimulation. During a training session the subject performed 5 blocks of an N-alternative forced choice (N-AFC) task consisting of 60 trials with self-paced breaks between blocks. During the training sessions only the 15 stimuli to be trained were presented. At the beginning of each trial, the orthographic labels for the word choices were displayed on the screen, and a VT stimulus was played after a short delay. Participants then indicated which label corresponded to the VT stimulus using a numerical key. The keys '1' to '0' corresponded to the left-to-right progression of the word choices displayed on the screen. Feedback was given after each trial, as well as an opportunity to replay any of the word choices. To facilitate training progression, the training paradigm utilized a leveling system organized in sets of 3 levels. The level of the participant determined the similarity of the stimuli on each trial as well as the number of choices (N) in the N-AFC task. In a set of 3 levels, the number of choices (N) was kept constant, but the choices themselves were increasingly confusable. For example, in level 1 subjects may have to distinguish 'sand and 'meat' but in level 3 they may have to distinguish the more similar pair 'sand and 'tanned'. Subjects started on level 1 which utilized a 2-AFCs, and the number of choices N was increased by 1 when progressing between each set of 3 levels (e.g., level 3 to level 4). An accuracy of 80% was required to advance to the next level. After the completion of all 6 training sessions subjects were invited to perform a post-training fMRI scan. Then on a separate day from the post-training scan subjects were brought back to perform a 10-AFC task. Stimuli presented in the 10-AFC task consisted only of the 15 trained words and like the training sessions consisted of 5 self-paced blocks of 60 trials each.

FMRI Experimental Procedures

EPI images were collected from 9 event-related runs in the auditory scan and 6 runs in each of the VT scans. A sparse acquisition paradigm was used in the auditory scan. Each run contained either 30 auditory vocoded, 30 VT vocoded, or 30 VT token-based stimuli. The same words were used in all the scans, but subjects were only trained to recognize VT versions of 15 of them. In both scans, subjects performed a 1-back task that was

utilized to maintain attention: Subjects were asked to press a button in their left hand whenever the same stimulus was presented on two consecutive trials. These catch trials comprised ten percent of the trials in each run. Furthermore, an additional ten percent of trials were null trials.

In the auditory scan each trial was 3s long and started with 1.5s of volume acquisition followed by the auditory word (during the silent period, see below, "Data Acquisition"; Fig. 1D). There were 118 trials per run plus an additional 15s fixation at the start of the run for a total run length of 369s and session length of 43 min. In the VT scan, each trial was 4s long (Fig. 1D) and there was a total of 111 trials per run plus an additional 10s fixation at the start and end of the run for total run length of 464 s and a session length of 46 min.

FMRI Data Acquisition

MRI data were acquired at the Center for Functional and Molecular Imaging at Georgetown University on a 3.0 Tesla Siemens Trio Scanner for both the auditory and VT scans. We used whole-head echo-planar imaging sequences (flip angle = 90°, TE = 30 ms, FOV = 205, 64x64 matrix) with a 12-channel head coil. For the auditory scan we used a sparse acquisition paradigm (TR = 3000 ms, TA = 1500 ms) in which each image was followed by an equal duration of silence before the next image was acquired. 28 axial slices were acquired in descending order (thickness = 3.5 mm, 0.5 mm gap; in-plane resolution = 3.0x3.0 mm^2). This sequence was used in previous auditory studies from our lab (Chevillet et al., 2013). For the VT scan, we used a continuous acquisition paradigm (TR=2000 ms) and collected 33 interleaved descending slices at the same resolution as in the auditory scan. A T1-weighted MPRAGE image (resolution 1x1x1mm^3) was also acquired for each subject.

FMRI Data Preprocessing

Image preprocessing was performed in SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) and AFNI version 20.1.03 (Cox, 1996; Cox and Hyde, 1997). The first four acquisitions of each run were discarded to allow for T1 stabilization, and the remaining EPI images were slice-time corrected to the middle slice for the VT scans. No slice-time correction was performed for the auditory scans due to using a sparse acquisition paradigm due to temporal discontinuities between successive volumes (Perrachione and Ghosh, 2013). These images were then spatially realigned and submitted to the AFNI *align_epi_anat.py* function to co-register the

anatomical EPI images for each subject. This was used because, upon inspection, it provided better registration between the anatomical and functional scans than the corresponding SPM12 routine.

Anatomical Preprocessing

Freesurfer version 6.0 (Fischl et al., 1999) was used to reconstruct cortical surface models including an outer pial and inner white-matter surface using the standard *recon-all* function. These surfaces were then brought into the SUMA environment (Argall et al., 2006; Saad and Reynolds, 2011) and fit to a standardized mesh based on an icosahedron with 64 linear divisions using AFNI's MapIcosehedron command (Oosterhof et al., 2011; Saad and Reynolds, 2011). This procedure yielded 81,924 nodes for each participant's whole-brain cortical surface mesh. Each node on the standard mesh corresponds to the same location across subjects – thereby allowing node-wise group-level analysis. This improved the spatial resolution of our analyses since interpolation of the functional data is unnecessary (Oosterhof et al., 2011). Finally, we used the CoSMoMVPA toolbox (Oosterhof et al., 2016), and the Surfing Toolbox (Oosterhof et al., 2011) to construct searchlights around each surface node by selecting the 30 closest voxels measured by geodesic distance.

Univariate Analyses

We fit a general linear model (GLM) to each subject's pre-processed functional images. For both the auditory and VT studies, we specified 38 regressors in for each run: 30 regressors, 1 for each word, 1 regressor for button press, and 6 motion regressors of no interest, For the all scans, a "Stimuli-Baseline" contrast image was generated for each subject. The contrast maps were smoothed using a 8mm FWHM smoothing kernel and then mapped to the cortical surface using 3dVol2Surf. For each scan, a one-sample t-test was used to compare "Stimuli-Baseline" versus 0. Finally, a paired t-test was used to compare pre- versus post-training scans.

Defining Regions of Interest

In the current study we tested evidence for metamodal engagement in specific regions of interest (ROIs). These ROIs were defined either functionally or structurally. Functional ROIs were defined by applying whole-brain representational similarity analysis (RSA)(Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013) to

auditory fMRI data (see below). This revealed statistically significant bilateral STG clusters that were then used in the main analyses with VT data. We also hypothesized that learning to recognize VT stimuli as words might rely on routes other than metamodal recruitment in the STG, specifically through paired-associate learning in the hippocampus (Eichenbaum et al., 1996, 2007; Gilbert and Kesner, 2003; Treder et al., 2021). We therefore defined bilateral hippocampal ROI using the HCP-MMP1.0 (Glasser et al., 2016) atlas.

Representational Similarity Analysis (RSA)

Constructing Model Representational Dissimilarity Matrices (mRDMs)

Three mRDMs were generated: an auditory mRDM, a VT vocoded mRDM, and a VT token-based mRDM. Entries in these mRDMs corresponded to distances between words. The distance metric that was used was the edit distance between the words where the edits were weighted by the perceptual confusability of the phonemes to be substituted. Edit distances are frequently used with highly intelligible speech, for which there are no phoneme-to-phoneme dissimilarity data, and when more refined segment-to-segment distances are not available as was the case for the VT token-based algorithm. Furthermore, recent work (Kell et al., 2018) has shown that the representational format captured by the edit distance matched those found in both higher-order STG speech regions and speech recognition-specific representations learned in later layers of a deep neural network. Auditory phoneme confusability was derived from a behaviorally measured perceptual auditory vocoded phoneme identification task. For both VT algorithms, phoneme confusability was generated using the last training block of N-AFC training data collected in this study. This procedure involved constructing word confusion matrices and using it to extract phoneme-level confusion matrices. Vowel confusions were extracted directly from the monosyllabic word confusion data. Consonant confusions were extracted by collapsing over pre-and post-vocalic positions. In addition, a simplifying assumption was made for incorrect responses where single consonants were matched with consonant clusters. The implemented procedure resulted in credit for correct identifications of individual consonants in clusters while attributing incorrect responses to both consonants in a cluster. Once the phoneme-level confusability was computed for auditory and VT conditions, it was transformed into a distance measure using a phi-square transform (Iverson et al., 1998). Word-to-word distances were computed as the sum of the pairwise phoneme distances for all the position-specific phoneme pairs in each of the possible pairs of stimulus words. Given the difficulty of estimating a distance swap between

consonants and vowels as well as between segments of different lengths, we restricted our analyses to CVCC words which were our most common segmental class (Tbl. 1). This resulted in a 9-by-9 auditory vocoded, VT vocoded, and VT token-based mRDMs for the CVCC trained words (Fig. 2). These representational spaces are highly correlated (r = 0.94) and reflect the close representational congruence at the behavioral level between auditory and VT stimuli generated by both algorithms.

Whole-Brain Searchlight RSA Analysis

RSA (Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013) analyses were performed using the CoSMoMVPA toolbox (Oosterhof et al., 2016) and custom MATLAB scripts. Within a given searchlight, the activity (t-statistic) in the voxels for each condition constituted its pattern. A cocktail-blank removal was performed on this condition-by-voxel data matrix whereby the mean pattern of activity across conditions was removed for each voxel (Walther et al., 2016). A neural dissimilarity matrix (nRDM) was then computed in each searchlight by computing the pairwise Pearson correlation distance (1-Pearson Correlation) between the patterns of all pairs of conditions. To assess whether a given region represented stimuli in a hypothesized format, the nRDM was compared to the mRDM. This was done by taking the Spearman Correlation between the vectorized lower triangles of the nRDM and mRDM. This correlation was then Fischer z-transformed to render the correlations more amenable to parametric analyses (Kriegeskorte et al., 2008).

ROI-Based RSA Analysis

ROI-based RSA analyses were performed in the VT scans to test if, following training, VT stimuli engaged auditory speech representations. To do so, we averaged the Fischer z-transformed correlations of searchlights in each ROI for the four groups (pre/post x vocoded/token). We then fit these average ROI correlations with a linear mixed effects model in R using the Lme4 Package. This model included three main effects TrainingPhase (0 for pre-training, 1 for post-training), Algorithm (0 for token, 1 for vocoded), and Hemi (0 for right, 1 for left). It also included all interaction terms, as well as a random slope and intercept. The random effects terms allowed us to model the subject-specific variability in the pre-training and the training-related change in correlation. The final model is shown below:

$$Correlation \sim 1 + TrainingPhase * Algorithm * Hemi + (TrainingPhase \mid Subj)$$

The reference group corresponding to the intercept was specified as pre-training, token-based, right-hemisphere. All βs reported reflect deviations from this reference group given the other effects. The model was estimated using REML and degrees of freedom were adjusted using the Satterthwaite approximations. Post-hoc contrasts were computed using the *emmeans* package and all reported p-values were corrected for multiple comparisons using Sidak's method.

Task-Regressed Functional Connectivity

The metamodal theory critically hypothesizes that metamodal engagement consists in linking a brain area performing a particular computation (e.g., word representation) in a standard modality with an input stream from a novel modality. We therefore performed functional connectivity analyses to test for learning-induced changes in functional connectivity between the somatosensory and auditory ROI of interest. Specifically, we used the CONN-fMRI toolbox (Whitfield-Gabrieli & Nieto-Castanon, 2012) to perform smoothing, segmentation, and cleaning of the data as well as to compute seed-to-voxel correlation maps. Native-space functional data were smoothed using an 8mm FWHM smoothing kernel. Next, anatomical scans were segmented to identify regions of white matter and CSF. We then regressed out the signals from these regions using CompCor (Behzadi, Restom, Liau, & Liu, 2007) as well as the main effect of task. Whole-brain seed-to-voxel correlation maps were then computed within each subject. Finally, we mapped each subject's correlation maps to a standard cortical mesh using 3dVol2Surf to perform group analyses.

Whole-Brain Statistical Correction

We tested the group-level significance of whole-brain RSA analyses as well as functional connectivity differences by first computing a t-statistic at each node on the standard surface. To correct these t-statistic maps for multiple comparisons, we first estimated the smoothness of the data for each analysis in each hemisphere using the AFNI/SUMA *SURFFWHM* command. We then used this smoothness estimate to generate noise surface maps using the AFNI/SUMA *slow_surf_clustsim.py* command. This then allowed us to generate an expected cluster size distribution at various thresholds that we compared clusters in our actual data to. For the whole-brain analyses a two-tailed cluster-defining threshold of $\alpha = .005$ was used. Since the

auditory RSA scan was used as an independent localizer scan in which to investigate neural effects of VT training a stricter cluster-defining threshold (α = .001) was applied in the auditory RSA scan to isolate more spatially restrictive clusters. All resulting clusters were corrected at the $p \leq .05$ level. Tables report the coordinates of the center of mass of clusters in MNI space and their location as defined by the HCP-MMP1.0 (Glasser et al., 2016) and Talairach-Tournoux Atlases.

Results

Analysis Overview

We first examined univariate engagement of cortical areas by VT and auditory stimuli. Next, we used RSA to identify areas encoding auditory speech, and then tested whether VT stimuli were encoded like the auditory speech stimuli in those areas following training. We also examined training-related changes in functional connectivity between somatosensory and auditory areas to provide complementary evidence for learning-related differences. Finally, we examined if token-based stimuli, which failed to show metamodal engagement in auditory areas, were encoded in the hippocampus which is known to play a role in associative learning.

Behavior

Subjects (n=22) were trained to recognize stimuli derived from either a token-based of vocoded auditory-to-VT sensory substitution algorithm, but 2 subjects were excluded due to a failure to complete the training paradigm. Importantly, the performance for all subjects was markedly above chance on the 10-AFC session performed after the post-training fMRI scan. Participants in both vocoded and token-based groups achieved progressively higher levels in the behavioral training paradigm across training sessions (Fig. 3A). The median final levels achieved were 8 and 7 for the token-based and vocoded VT groups respectively. After the final post-training fMRI scan, subjects completed a 10-AFC test on the trained words (Fig. 3B). All subjects performed better than chance (10%) and the median accuracies were 35.3% and 48.5% for the token-based and vocoded VT groups respectively. A two-sample t-test revealed no significant difference in accuracy between algorithm groups (t_{18} = 0.386, p = 0.704).

Univariate fMRI Analysis

In the auditory scan, the contrast of “All Words>Baseline” revealed bilateral Superior Temporal Gyrus (STG) activation (Table 2 and Fig. 4A). In the VT scans, unpaired two-sample t-tests revealed no significant differences between the vocoded and token-based groups in either the pre-training or post-training phase. Therefore, subjects were combined within training-phase to test for the cortical common response to VT stimulation. The contrast “All Vibrotactile Words>Baseline” revealed several regions, including bilateral supplementary motor area (SMA), precentral gyri (Table 2 and Fig. 4B-C). No significant clusters were identified for the post- vs pre- training contrast. To gain a better picture of the neuronal representations underlying these responses, we performed a series of RSA analyses.

Whole-brain searchlight analysis reveals bilateral STG regions are engaged in the perception of spoken vocoded words

We conducted a whole-brain searchlight RSA analysis to identify regions engaged by auditory vocoded words (Fig. 5). This revealed left (x = -58, y = -18, z = 5; α = 0.001; p = 0.001) and right mid-STG (x = 58, y = -14, z = 3; α = 0.001; p = 0.016) clusters. There is strong evidence (Hamilton et al., 2018, 2021) that these regions are involved in processing complex temporal patterns found in auditory speech.

Vocoded but not token-based VT stimuli are encoded similarly to auditory spoken words in the mid-STG following VT speech training

Next, we conducted ROI-based RSA analyses to test the prediction that trained VT stimuli would be encoded similarly to auditory words in the mSTG. To do so, we used a linear mixed-effects model (see Methods) to test the effects of training phase, algorithm, hemisphere, as well as the interaction among them on the correlations between neural and model RDMs (Fig. 6).

This revealed a significant interaction effect between training phase and algorithm (β = 0.240, $t_{31.09}$ = 2.679, p = 0.012). Post-hoc tests revealed a significant training effect in the right mSTG for the vocoded ($t_{31.1}$ = 3.380, p = 0.008 Sidak-adjusted; *d* = 2.09, 95% CI = [0.78, 3.40]) but not the token-based group ($t_{31.1}$ = -0.408, p = 0.990 Sidak-adjusted; *d* = -0.25, 95% CI = [-1.51, 1.00]). Furthermore, post-hoc tests did not reveal a significant increase between the pre- and post-training correlations in the left mSTG for either the vocoded ($t_{31.1}$ = 1.781, p = 0.298 Sidak-adjusted; *d* = 1.10, 95% CI = [-0.17, 2.38]) or the token-based ($t_{31.1}$ = 0.250, p = 0.999 Sidak-adjusted; *d* = 0.15, 95% CI = [-1.11, 1.42]) group. Analyses were repeated for the VT token-based group using its corresponding behavioral mRDM (Fig. 2) which still showed a non-significant training effect in both the left (t_{30} = 0.025, p = 0.989 Sidak-adjusted; *d* = 0.267, 95% CI = [-1.042, 1.58]) and right mSTG (t_{30} = -0.008, p = 0.999 Sidak-adjusted; *d* = -0.091, 95% CI = [-1.399, 1.217]).

Although there was no significant three-way interaction, we performed exploratory analyses to compare the correlation between the left vs. right mid-STG. This revealed significantly (t_{36} = 2.396, p = 0.011 uncorrected; *d* = 1.07, 95% CI = [0.15, 2.00]) higher correlations post-training in the right than left mSTG. In addition, there was a non-significant (t_9 = 2.185, p = 0.057 uncorrected; *d* = 0.61, 95% CI = [-0.03, 1.25]) difference when the difference between pre- and post-training correlations were compared between the right and left mid-STG. Finally, since training on VT stimuli may induce changes in high-level somatosensory areas, we also examined training-related changes in S2 representations using the OPI/SII HCP-MMP1.0 ROI (Glasser et al., 2016). However, there were no significant interactions or main effects. These results indicate that trained VT stimuli based on vocoded speech were encoded similarly to auditory speech in the mid-STG while token-based VT stimuli were not.

Training with Vocoded VT Speech Stimuli Increases Functional Connectivity Between Somatosensory and Auditory Regions

Previous studies showed that learning is accompanied by increased functional connectivity between cortical areas (Lewis et al., 2009; Urner et al., 2013; Siuda-Krzywicka et al., 2016). Therefore, we tested the hypothesis that training on the vocoded VT word stimuli was associated with increased functional connectivity of somatosensory regions and the auditory word encoding right mid-STG ROI (Fig. 5). To do so, we computed the training-related changes in the right mid-STG seed-to-voxel functional connectivity in the vocoded group (Fig. 7A, Table 3). This revealed two clusters, one in the left STG (α = 0.005; p = 0.044) and another in the left secondary somatosensory (SII) (α = 0.005; p = 0.026). Furthermore, reasoning that VT stimulation on the right arm would engage the left SII region, we performed an additional seed-to-voxel analysis using the left SII seed defined by the HCP-MMP1.0 atlas (Glasser et al., 2016). This complementary analysis (Fig. 7B) revealed two clusters, one in the right insula and Heschl's Gyrus (α = 0.005; p = 0.001) and another in the right STG (α = 0.005; p = 0.001). The left SII also showed an increase in connectivity to the left central sulcus (α = 0.005; p = 0.001). (Table 3). Similar seed-to-voxel analyses also using the left hippocampus or the bilateral mid-STG ROIs as seeds revealed no significant training-related differences in the token-based group. This pattern of training-related functional connectivity between somatosensory and auditory areas for VT vocoded but not token based stimuli was also found when calculating ROI-to-ROI functional connectivity (Fig. 7C-D). These results support a model in which vocoded VT speech training leads to increased functional connectivity between somatosensory areas and auditory speech areas.

Training Increases Encoding of the VT Token-Based Stimuli in the Left Hippocampus

The noteworthy difference in the encoding of VT vocoded versus token-based speech in the mid-STG raised the question: what other pathway underlies subjects' ability to learn the token-based VT stimuli as words (see Fig. 3). As mentioned in the Introduction, it is possible that a poor match between the temporal dynamics of VT token-based stimuli and auditory speech precludes metamodal engagement in the mSTG and instead favors alternate strategies to learn associations between arbitrary pairs of stimuli. A key region involved in learning such associations is the hippocampus (McClelland et al., 1995; Eichenbaum et al., 1996, 2007; O'Reilly and Rudy, 2001).

We therefore used a linear mixed effect model to test whether the hippocampus encoded token-based stimuli after training (Fig. 8). This analysis revealed a significant two-way interaction between training phase and hemisphere ($\beta = 0.095$, $t_{36} = 2.696$, $p = 0.011$; Fig. 8) as well as a significant three-way interaction effect between training phase, algorithm, and hemisphere ($\beta = -0.151$, $t_{36} = -3.027$, $p = 0.005$). The three-way interaction suggests that the relationship between training phase and hemisphere varied depending on the algorithm. In the left hemisphere, post-hoc tests revealed a significant ($t_{30.7} = 3.232$, $p = 0.012$ Sidak-adjusted; $d = 2.022$, 95% CI = [0.70, 3.35]) training-related increase in correlations for the token-based but not vocoded ($t_{30.7} = -0.785$, $p = 0.901$ Sidak Adjusted; $d = 0.49$, 95% CI = [-0.79, 1.77]) VT group. In the right hemisphere, there was a trending increase in correlation for the vocoded group ($t_{30.7} = 2.387$, $p = 0.0902$ Sidak Adjusted; $d = 1.49$, 95% CI = [0.19, 2.80]) but not the token-based ($t_{30.7} = 0.506$, $p = 0.9783$ Sidak Adjusted; $d = 0.32$, 95% CI = [-0.96, 1.60]) VT group. Of note, using the vocoded and token-based mRDMs for the corresponding groups also resulted in the same significant two-($\beta = 0.071$, $t_{36} = 2.139$, $p = 0.039$) and three-way ($\beta = 0.108$, $t_{36} = 2.287$, $p = 0.028$) interaction effects. Furthermore, there was also a significant training effect for VT token-based stimuli in the left hippocampus ($t_{28.2} = 2.598$, $p = 0.015$); $d = 1.76$, 95% CI = [0.359, 3.172]).

Discussion

Metamodal theories of brain organization (Pascual-Leone and Hamilton, 2001; Heimler et al., 2015) propose that cortical areas are best described by their task-specific sensory modality-invariant function. However, mixed evidence for metamodal brain organization in neurotypical individuals (Sadato et al., 1996; Ptito et al., 2005; Amedi et al., 2007; Siuda-Krzywicka et al., 2016; Bola et al., 2017) has called into question the conditions under which metamodal engagement occurs. We argue, that metamodal engagement in neurotypical individuals requires not just correspondence at the task level(Marr, 1982) but also between stimuli at the level of neural encoding. In the current study, we investigated this hypothesis by training subjects on the same word recognition task using one of two auditory-to-VT transformation algorithms. One algorithm (vocoded) preserved the temporal dynamics of auditory speech while the other algorithm (token-based) did not. First, using whole brain RSA and an independent auditory scan we identified auditory speech areas in the bilateral mSTG that served as putative targets for metamodal engagement by VT stimuli. We then showed that, after training only VT vocoded stimuli engaged this area like auditory vocoded words. Importantly, subjects in both groups achieved comparable levels of proficiency on the post-training recognition task and had similar behavioral confusions. This eliminates performance differences as a reason for the different training effects at the neural level. We then showed that only VT vocoded but not token-based stimuli were associated with a significant training-related increase in functional connectivity between the mid-STG and secondary somatosensory areas. Finally, both algorithms, to different degrees, engaged hippocampal areas previously implicated in paired-associate learning.

In this study, we show that adequately capturing (and eventually harnessing) the metamodal potential of cortex requires not only the right task and sensory modalities but also an understanding of the information representation in these regions. Prior work has primarily investigated metamodal engagement in congenitally sensory-deprived individuals (Lomber et al., 2010; Reich et al., 2011; Bola et al., 2017). In such cortical areas, given the right task-relevant connectivity, bottom-up input from another sensory modality can conceivably drive the *de novo* learning of task-relevant representations even for encoding schemes very different from those in neurotypical individuals (Striem-Amit et al., 2012). However, in neurotypical adults, existing representations in traditionally unisensory areas reflect the task-relevant features of the typical sensory input (Simoncelli and Olshausen, 2001; Lewicki, 2002). Therefore, for metamodal engagement to occur, information partially

processed in one sensory hierarchy needs to interface with pre-existing representations derived from the typical modality. The lack of evidence for metamodal engagement of the mid-STG by token-based VT stimuli in our study and the mixed evidence in prior studies of neurotypical individuals may reflect a failure to successfully perform this interface mapping.

The ability to map between sensory hierarchies likely depends on both anatomical and functional convergence. Anatomical (Schroeder et al., 2003; Mothe et al., 2006a; Smiley et al., 2007) and functional studies in humans and non-human primates (Schroeder et al., 2001; Foxe et al., 2002; Kayser et al., 2009; Ro et al., 2013) have established convergence points between somatosensory and auditory cortices such as the belt and parabelt areas. Given this connectivity, prior computational studies have shown that the mapping between different representational formats can be learnt through simple biologically plausible learning rules (Pouget and Snyder, 2000; Davison and Frégnac, 2006). Still, while it is simple to learn the mapping between static features, it is non-trivial to match the temporal dynamics between functional hierarchies (Pouget and Snyder, 2000; Davison and Frégnac, 2006). In the auditory cortex studies (Overath et al., 2015; Moore and Woolley, 2019) have shown that auditory stimuli that do not preserve the same temporal modulations found in conspecific communication signals sub-optimally drive higher-order auditory cortex and preclude learning. This is supported by our current results, that only VT vocoded stimuli that preserve these fast temporal dynamics can drive auditory perceptual speech representations in the mid-STG.

The token-based algorithm was based on a previously published algorithm (Reed at al., 2018) where stimulus durations were chosen to optimize recognizability. In this study, the token-based algorithm generated longer stimuli than those generated by the vocoded algorithm. Given this difference, it is remarkable that words were recognized equally well in both algorithms, even though the vocoded stimuli were shorter (therefore requiring more information processed per unit time). Thus, although both algorithms were similarly learnable, effective metamodal engagement may facilitate more efficient learning. Yet, differences in stimulus length between the two algorithms could lead to a trivial difference in BOLD contrast responses, making our within-subject before/after experimental design essential for controlling stimulus differences and isolating the neural effects of training.

Intriguingly, we find stronger evidence of metamodal engagement by VT vocoded stimuli in the right rather than left mid-STG. A significant body of work (Boemio et al., 2005; Obleser et al., 2008; Giraud and Poeppel, 2012;

Flinker et al., 2019; Albouy et al., 2020) suggests that the left and right STG are differentially sensitive to spectrotemporal content of auditory stimuli. Specifically, it has been proposed (Flinker et al., 2019) that the left STG samples auditory information on fast and slow timescales while the right preferentially does the latter. In the current study, our VT vocoded stimuli preserve the coarse temporal dynamics of auditory speech, but due to hardware limitations have a lower temporal resolution than the auditory source signal. Also, the temporal resolution of vibrotactile perception is lower than that of auditory processing – since receptors in the skin act as additional low pass filters (Bensmaïa and Hollins, 2005). Thus, the observed metamodal engagement of the right more than the left STG provides support for the asymmetric spectrotemporal modulation theory of hemispheric processing (Flinker et al., 2019).

Given that subjects were able to learn token-based and vocoded VT stimuli as words with roughly equal proficiency, how does the former group accomplish this task? We initially hypothesized that the slower temporal dynamics of token-based stimuli would engage more anterior STG areas that are thought to integrate information on longer timescales (Overath et al., 2015; Hullett et al., 2016). However, we did not find evidence for this in the current study. This may be due to insufficient connectivity between somatosensory and anterior STG (Mothe et al., 2006b). However, we did find evidence that token-based stimuli engage neural representations in the left hippocampus. This fits with previous proposals that learned associations can be retrieved using paired-associate recall circuits in the medial temporal lobe (Miyashita, 2019). A more thorough understanding of this process through future studies will shed additional insight into which pathways and mechanisms are leveraged to learn different types of associations.

The present study has some limitations. For instance, token-based stimuli may be encoded in the mSTG in a format that may be captured by an alternative mRDM. Still, we did not find a significant training effect in the mSTG even when using a mRDM derived specifically for token-based stimuli. Likewise, our hypothesis that only the vocoded encoding led to multimodal engagement was supported by the functional connectivity analyses that revealed significant training-related changes in connectivity between somatosensory and auditory areas only for the VT vocoded but not the token-based group. This acts as complementary evidence that VT token-based stimuli are unable to “engage” the mSTG. Next, despite evidence for a training-related effect for VT token-based stimuli in the hippocampus, this result should be interpreted with caution since the post-training correlation was not significantly greater than 0. Finally, a limitation of the present study is the

modest sample size used. This is especially true in the VT scans, in which only large effect sizes could be detected. In the current study, the crucial hypothesis was that VT vocoded stimuli engaged auditory word representations in the STG after training better than token-based VT stimuli. Mixed-effects analysis was able to detect this significant predicted interaction. Yet, it is possible that token-based stimuli might also exhibit small training-related changes that may have been missed due to the small sample size.

In summary, ours is the first study to use two different sensory substitution algorithms to demonstrate that metamodal engagement in neurotypical individuals relies on a correspondence between the encoding schemes of novel and standard sensory modality stimuli. This extends metamodal theories (Heimler et al., 2015) that only emphasize a correspondence at the task level (Heimler et al., 2015). Consideration of these correspondences may provide insight into how the brain maps between various levels of different functional hierarchies like sub-lexical and lexical orthography and phonology (Share, 1999). It also suggests that therapeutic sensory substitution devices might benefit from different algorithms for patients with acquired rather than congenital sensory deprivation. For the former, careful consideration should be given to the type of sensory substitution algorithm to best interfaces with spared sensory representations. The ability to "piggyback" onto an existing processing hierarchy (e.g., auditory speech recognition) may facilitate the rapid learning of novel stimuli presented through a spared sensory modality (e.g., VT). Future work should explore whether this observed integration into existing processing streams leads to improved generalization and transfer of learning.

References

Albouy P, Benjamin L, Morillon B, Zatorre RJ (2020) Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. Science 367:1043–1047.

Amedi A, Stern WM, Camprodon JA, Bermpohl F, Merabet L, Rotman S, Hemond C, Meijer P, Pascual-Leone A (2007) Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. Nat Neurosci 10:687–689.

Argall BD, Saad ZS, Beauchamp MS (2006) Simplified intersubject averaging on the cortical surface using SUMA. Hum Brain Mapp 27:14–27.

Atteveldt N van, Formisano E, Goebel R, Blomert L (2004) Integration of Letters and Speech Sounds in the Human Brain. Neuron 43:271–282.

Atteveldt N van, Roebroeck A, Goebel R (2009) Interaction of speech and script in human auditory cortex: Insights from neuro-imaging and effective connectivity. Hearing Research 258:152–164.

Auer ET, Bernstein LE (1997) Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. J Acoust Soc Am 102:3704–3710.

Bach-y-Rita P, Kercel SW (2003) Sensory substitution and the human–machine interface. Trends Cogn Sci 7:541–546.

Benetti S, Zonca J, Ferrari A, Rezk M, Rabini G, Collignon O (2020) Visual motion processing recruits regions selective for auditory motion in early deaf individuals. Biorxiv:2020.11.27.401489.

Bensmaïa S, Hollins M (2005) Pacinian representations of fine surface texture. Percept Psychophys 67:842–854.

Bernstein LE, Demorest ME, Coulter DC, O'Connell MP (1991) Lipreading sentences with vibrotactile vocoders: performance of normal-hearing and hearing-impaired subjects. J Acoust Soc Am 90:2971–2984.

Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci 8:389–395.

Bola Ł, Zimmermann M, Mostowski P, Jednoróg K, Marchewka A, Rutkowski P, Szwed M (2017) Task-specific reorganization of the auditory cortex in deaf humans. Proc National Acad Sci 114:E600–E609.

Chevillet MA, Jiang X, Rauschecker JP, Riesenhuber M (2013) Automatic Phoneme Category Selectivity in the Dorsal Auditory Stream. The Journal of Neuroscience 33:5208–5215.

Cox RW (1996) AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. Comput Biomed Res 29:162–173.

Cox RW, Hyde JS (1997) Software tools for analysis and visualization of fMRI data. Nmr Biomed 10:171–178.

Davison AP, Frégnac Y (2006) Learning Cross-Modal Spatial Transformations through Spike Timing-Dependent Plasticity. J Neurosci 26:5604–5615.

Eberhardt SP, Jr. ETA, Bernstein LE (2014) Multisensory training can promote or impede visual perceptual learning of speech stimuli: visual-tactile vs. visual-auditory training. Front Hum Neurosci 8:829.

Eichenbaum H, Schoenbaum G, Young B, Bunsey M (1996) Functional organization of the hippocampal memory system. Proc National Acad Sci 93:13500–13507.

Eichenbaum H, Yonelinas AP, Ranganath C (2007) The Medial Temporal Lobe and Recognition Memory. Annu Rev Neurosci 30:123–152.

Fairhall SL, Porter KB, Bellucci C, Mazzetti M, Cipolli C, Gobbini MI (2017) Plastic reorganization of neural systems for perception of others in the congenitally blind. Neuroimage 158:126–135.

Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191.

Fischl B, Sereno MI, Tootell RBH, Dale AM (1999) High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum Brain Mapp 8:272–284.

Flinker A, Doyle WK, Mehta AD, Devinsky O, Poeppel D (2019) Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. Nat Hum Behav 3:393–405.

Foxe JJ, Wylie GR, Martinez A, Schroeder CE, Javitt DC, Guilfoyle D, Ritter W, Murray MM (2002) Auditory-Somatosensory Multisensory Processing in Auditory Association Cortex: An fMRI Study. J Neurophysiol 88:540–543.

Gilbert PE, Kesner RP (2003) Localization of Function Within the Dorsal Hippocampus: The Role of the CA3 Subregion in Paired-Associate Learning. Behav Neurosci 117:1385–1394.

Giraud A-L, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511–517.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Essen DCV (2016) A multi-modal parcellation of human cerebral cortex. Nature 536:171–178.

Glezer LS, 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.

Hamilton LS, Edwards E, Chang EF (2018) A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. Curr Biol 28:1860-1871.e4.

Hamilton LS, Oganian Y, Chang EF (2020) Topography of speech-related acoustic and phonological feature encoding throughout the human core and parabelt auditory cortex. Biorxiv:2020.06.08.121624.

Hamilton LS, Oganian Y, Hall J, Chang EF (2021) Parallel and distributed encoding of speech across human auditory cortex. Cell 184:4626-4639.e13.

Heimler B, Striem-Amit E, Amedi A (2015) Origins of task-specific sensory-independent organization in the visual and auditory brain: neuroscience evidence, open questions and clinical implications. Curr Opin Neurobiol 35:169–177.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nature Reviews Neuroscience 8:nrn2113.

632 Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF (2016) Human Superior Temporal Gyrus
633 Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. The Journal of
634 Neuroscience 36:2014–2026.

635 Iverson P, Bernstein LE, Jr ETA (1998) Modeling the interaction of phonemic intelligibility and lexical structure
636 in audiovisual word recognition. Speech Commun 26:45–63.

637 Kayser C, Petkov CI, Logothetis NK (2009) Multisensory interactions in primate auditory cortex: fMRI and
638 electrophysiology. Hearing Res 258:80–88.

639 Kell A, Yamins D, Shook EN, Norman-Haignere SV, McDermott JH (2018) A Task-Optimized Neural Network
640 Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing
641 Hierarchy. Neuron 98:630-644.e16.

642 Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain.
643 Trends in Cognitive Sciences 17:401–412.

644 Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis - connecting the branches of
645 systems neuroscience. Frontiers in Systems Neuroscience 2:4.

646 Lewicki MS (2002) Efficient coding of natural sounds. Nature Neuroscience 5:356.

647 Lewis CM, Baldassarre A, Committeri G, Romani GL, Corbetta M (2009) Learning sculpts the spontaneous
648 activity of the resting human brain. P Natl Acad Sci Usa 106:17558–17563.

649 Lomber SG, Meredith MA, Kral A (2010) Cross-modal plasticity in specific auditory cortices underlies visual
650 compensations in the deaf. Nat Neurosci 13:1421–1427.

651 Malone PS, Eberhardt SP, Wimmer K, Sprouse C, Klein R, Glomb K, Scholl CA, Bokeria L, Cho P, Deco G,
652 Jiang X, Bernstein LE, Riesenhuber M (2019) Neural mechanisms of vibrotactile categorization. Hum Brain
653 Mapp 40:3078–3090.

654 Marr D (1982) Vision: A Computational Investigation into the Human Representation and Processing of Visual
655 Information. Henry Holt and Co., Inc.

656 Mattioni S, Rezk M, Battal C, Bottini R, Mendoza KEC, Oosterhof NN, Collignon O (2020) Categorical
657 representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. Elife
658 9:e50732.

659 McClelland JL, McNaughton BL, O'Reilly RC (1995) Why there are complementary learning systems in the
660 hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning
661 and memory. Psychol Rev 102:419–457.

662 Meijer PBL (1992) An experimental system for auditory image representations. Ieee T Bio-med Eng 39:112–
663 121.

664 Miyashita Y (2019) Perirhinal circuits for memory processing. Nat Rev Neurosci 20:577–592.

665 Moore JM, Woolley SMN (2019) Emergent tuning for learned vocalizations in auditory cortex. Nat Neurosci
666 22:1469–1476.

667 Mothe LA de la, Blumell S, Kajikawa Y, Hackett TA (2006a) Cortical connections of the auditory cortex in
668 marmoset monkeys: Core and medial belt regions. J Comp Neurol 496:27–71.

669 Mothe LA de la, Blumell S, Kajikawa Y, Hackett TA (2006b) Cortical connections of the auditory cortex in
670 marmoset monkeys: Core and medial belt regions. J Comp Neurol 496:27–71.

671 Obleser J, Eisner F, Kotz SA (2008) Bilateral Speech Comprehension Reflects Differential Sensitivity to
672 Spectral and Temporal Features. J Neurosci 28:8116–8123.

673 Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of
674 Neuroimaging Data in Matlab/GNU Octave. Frontiers in Neuroinformatics 10:27.

675 Oosterhof NN, Wiestler T, Downing PE, Diedrichsen J (2011) A comparison of volume-based and surface-
676 based multi-voxel pattern analysis. Neuroimage 56:593–600.

677 O'Reilly RC, Rudy JW (2001) Conjunctive representations in learning and memory: Principles of cortical and
678 hippocampal function. Psychol Rev 108:311–345.

679 Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal
680 structure revealed by responses to sound quilts. Nat Neurosci 18:903–911.

681 Pascual-Leone A, Hamilton R (2001) The metamodal organization of the brain. Prog Brain Res 134:427–445.

682 Penikis KB, Sanes DH (2022) A Redundant Cortical Code for Speech Envelope. J Neurosci 43:93–112.

683 Perrachione TK, Ghosh SS (2013) Optimized Design and Analysis of Sparse-Sampling fMRI Experiments.
684 Front Neurosci-switz 7:55.

685 Pouget A, Snyder LH (2000) Computational approaches to sensorimotor transformations. Nat Neurosci
686 3:1192–1198.

687 Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate
688 human speech processing. Nature Neuroscience 12:718–724.

689 Reed CM, Tan HZ, Perez ZD, Wilson EC, Severgnini FM, Jung J, Martinez JS, Jiao Y, Israr A, Lau F, Klumb K,
690 Turcott R, Abnousi F (2018) A Phonemic-Based Tactile Display for Speech Communication. Ieee T Haptics
691 12:2–17.

692 Reich L, Szwed M, Cohen L, Amedi A (2011) A Ventral Visual Stream Reading Center Independent of Visual
693 Experience. Curr Biol 21:363–368.

694 Renier L, Collignon O, Poirier C, Tranduy D, Vanlierde A, Bol A, Veraart C, Volder AGD (2005) Cross-modal
695 activation of visual cortex during depth perception using auditory substitution of vision. Neuroimage 26:573–
696 580.

697 Renier LA, Anurova I, Volder AGD, Carlson S, VanMeter J, Rauschecker JP (2010) Preserved Functional
698 Specialization for Spatial Processing in the Middle Occipital Gyrus of the Early Blind. Neuron 68:138–148.

699 Ro T, Ellmore TM, Beauchamp MS (2013) A Neural Link Between Feeling and Hearing. Cereb Cortex
700 23:1724–1730.

701 Saad ZS, Reynolds RC (2011) SUMA. Neuroimage 62:768–773.

702 Sadato N, Pascual-Leone A, Grafman J, Ibañez V, Deiber M-P, Dold G, Hallett M (1996) Activation of the
703 primary visual cortex by Braille reading in blind subjects. Nature 380:526–528.

704 Schroeder CE, Lindsley RW, Specht C, Marcovici A, Smiley JF, Javitt DC (2001) Somatosensory Input to
705 Auditory Association Cortex in the Macaque Monkey. J Neurophysiol 85:1322–1327.

706 Schroeder CE, Smiley J, Fu KG, McGinnis T, O'Connell MN, Hackett TA (2003) Anatomical mechanisms and
707 functional implications of multisensory convergence in early cortical processing. Int J Psychophysiol 50:5–
708 17.

709 Share DL (1999) Phonological Recoding and Orthographic Learning: A Direct Test of the Self-Teaching
710 Hypothesis. Journal of Experimental Child Psychology 72:95–129.

711 Simoncelli EP, Olshausen BA (2001) Natural Image Statistics and Neural Representation. Annu Rev Neurosci
712 24:1193–1216.

713 Siuda-Krzywicka K, Bola Ł, Paplińska M, Sumera E, Jednoróg K, Marchewka A, Śliwińska MW, Amedi A,
714 Szwed M (2016) Massive cortical reorganization in sighted Braille readers. Elife 5:e10762.

715 Smiley JF, Hackett TA, Ulbert I, Karmas G, Lakatos P, Javitt DC, Schroeder CE (2007) Multisensory
716 convergence in auditory cortex, I. Cortical connections of the caudal superior temporal plane in macaque
717 monkeys. J Comp Neurol 502:894–923.

718 Striem-Amit E, Cohen L, Dehaene S, Amedi A (2012) Reading with Sounds: Sensory Substitution Selectively
719 Activates the Visual Word Form Area in the Blind. Neuron 76:640–652.

720 Tong J, Mao O, Goldreich D (2013) Two-Point Orientation Discrimination Versus the Traditional Two-Point
721 Test for Tactile Spatial Acuity Assessment. Front Hum Neurosci 7:579.

722 Treder MS, Charest I, Michelmann S, Martín-Buro MC, Roux F, Carceller-Benito F, Ugalde-Canitrot A, Rollings
723 DT, Sawlani V, Chelvarajah R, Wimber M, Hanslmayr S, Staresina BP (2021) The hippocampus as the
724 switchboard between perception and memory. Proc National Acad Sci 118:e2114171118.

725 Twomey T, Waters D, Price CJ, Evans S, MacSweeney M (2017) How Auditory Experience Differentially
726 Influences the Function of Left and Right Superior Temporal Cortices. J Neurosci 37:9564–9573.

727 Urner M, Schwarzkopf DS, Friston K, Rees G (2013) Early visual learning induces long-lasting connectivity
728 changes during rest in the human brain. Neuroimage 77:148–156.

729 Vetter P, Bola Ł, Reich L, Bennett M, Muckli L, Amedi A (2020) Decoding Natural Sounds in Early "Visual"
730 Cortex of Congenitally Blind Individuals. Curr Biol 30:3039-3044.e2.

731 Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J (2016) Reliability of dissimilarity measures for
732 multi-voxel pattern analysis. NeuroImage 137:188–200.

733 Yi H, Leonard MK, Chang EF (2019) The Encoding of Speech Sounds in the Superior Temporal Gyrus. Neuron
734 102:1096–1110.

735

736 **Figure Legends and Tables**

737 **Figure 1: VT hardware, speech-to-tactile transformation algorithms, stimuli, fMRI experimental design,**
738 **and model dissimilarity matrix.** (A) 16-channel MRI-compatible VT stimulator. (B) Shows the token-based
739 algorithm for transforming spoken words into VT patterns. It assigns each phoneme a distinct VT pattern (see
740 Methods section for more details). (C) Shows the vocoding algorithm which focuses on preserving the
741 temporal dynamics between the auditory and VT stimuli. (D) Shows the auditory (top) and VT (bottom) fMRI
742 one-back paradigms used in the study. In both paradigms, subjects focused on a central fixation cross, and
743 pressed a button in their left hand if they heard or felt the same stimulus twice in a row. Green frame (not
744 shown in task) indicates such a one-back trial. Abbreviations: TR – repetition time, ITI – intertrial interval, TA –
745 acquisition time.
746
747 **Figure 2: Behavioral level correspondence between auditory and VT stimuli.** The auditory and the two VT
748 perceptual model representational dissimilarity matrices (mRDMs) for the 9 CVCC trained words are highly
749 correlated (r = 0.94) demonstrating a correspondence of stimulus similarities at the behavioral level.
750
751 **Figure 3: Progression of learning VT stimuli as speech.** (A) Shows the performance of individuals on the
752 behavioral training paradigm across sessions. The left and right plots show the training progression for
753 subjects in the token-based and vocoded VT groups, respectively. Data for the final training sessions for two
754 subjects, one per group, are missing due to technical error. Shaded lines connect the same individual across
755 sessions. Data for the final session of two subjects was lost due to technical error. (B) Shows the performance
756 of subjects by algorithm group on 10-AFC task completed after the final post-training fMRI scan. A two-sample
757 t-test reveals no significant difference in performance between the groups ($t_{18} = 0.386$, p = 0.704). Dashed red
758 line indicates chance performance. Horizontal lines in the violin plots reflect the median.
759
760 **Figure 4: Univariate activity for "Stimuli-Baseline" in the auditory and VT scans.** (A) Shows the group-
761 level speech perception network revealed by the contrast of all auditory words > baseline. (B) Shows the pre-
762 training group-level VT perception network revealed by the contrast of all vibrotactile words > baseline. (C)
763 Same as (B), but for post-training scans. Results are rendered on a SUMA-derived standard surface. All

results are presented at a cluster-defining two-tailed $\alpha = 0.005$ and $p \leq 0.05$. LH and RH refer to left and right hemisphere, respectively.

Figure 5: Auditory Scan – Representational similarity analysis (RSA) of vocoded auditory words. RSA revealed that neural RDMs in bilateral STG regions significantly correlated with the predicted auditory perceptual mRDM (Fig. 2) (n=18; $\alpha = 0.001$; $p \leq .05$). The center of mass of the left STG cluster was centered on MNI: -58, -18, 5. The center of mass of the right STG cluster was centered on MNI: 58, -14, 3. Colors reflect across-subject t-statistics.

Figure 6: Vocoded but not token-based VT stimuli are encoded similarly to auditory spoken words in the mid-STG following VT speech training. Linear mixed-effects analysis revealed a significant two-way interaction between Training Phase and Algorithm ($\beta = 0.240$, $t_{31.1} = 2.679$, $p = 0.012$). To investigate this interaction, we created interaction effects plots. (A) The mean Fisher-transformed Pearson correlation between neural and model RDMs estimated from the mixed-effects model for the vocoded group are represented by the opaque lines. For the VT vocoded group, post-hoc tests show a significant difference between pre- and post-training in the right ($t_{31.1} = 3.380$, $p = 0.008$ Sidak-adjusted) but not the left STG ($t_{31.1} = 1.781$, $p = 0.298$ Sidak-adjusted). (B) The same as (A) but for the token-based group. Post-hoc tests show no significant difference in the right ($t_{31.1} = -0.408$, $p = 0.990$ Sidak-adjusted) or left STG ($t_{31.1} = 0.250$, $p = 0.999$ Sidak-adjusted). Values above each violin reflect the *uncorrected* p-value from a one-sample t-test against 0. Semi-transparent lines reflect raw individual subject correlations from either the left (teal) or right (orange) STG. Horizontal lines in the violin plots reflect the median. Green asterisk marks significant ($p \leq .05$) differences after multiple comparisons correction.

Figure 7: Training with Vocoded VT Speech Stimuli Increases Functional Connectivity Between Somatosensory and Auditory Regions. (A) Using the right mid-STG ROI (Fig. 5) as a seed revealed two significant clusters of increased functional connectivity after training in the left STG (MNI: -50, -19, 7) and in the left supramarginal gyrus (MNI: -55, -28, 21). (B) Using the left SII seed derived from the HCP-MMP1.0 atlas

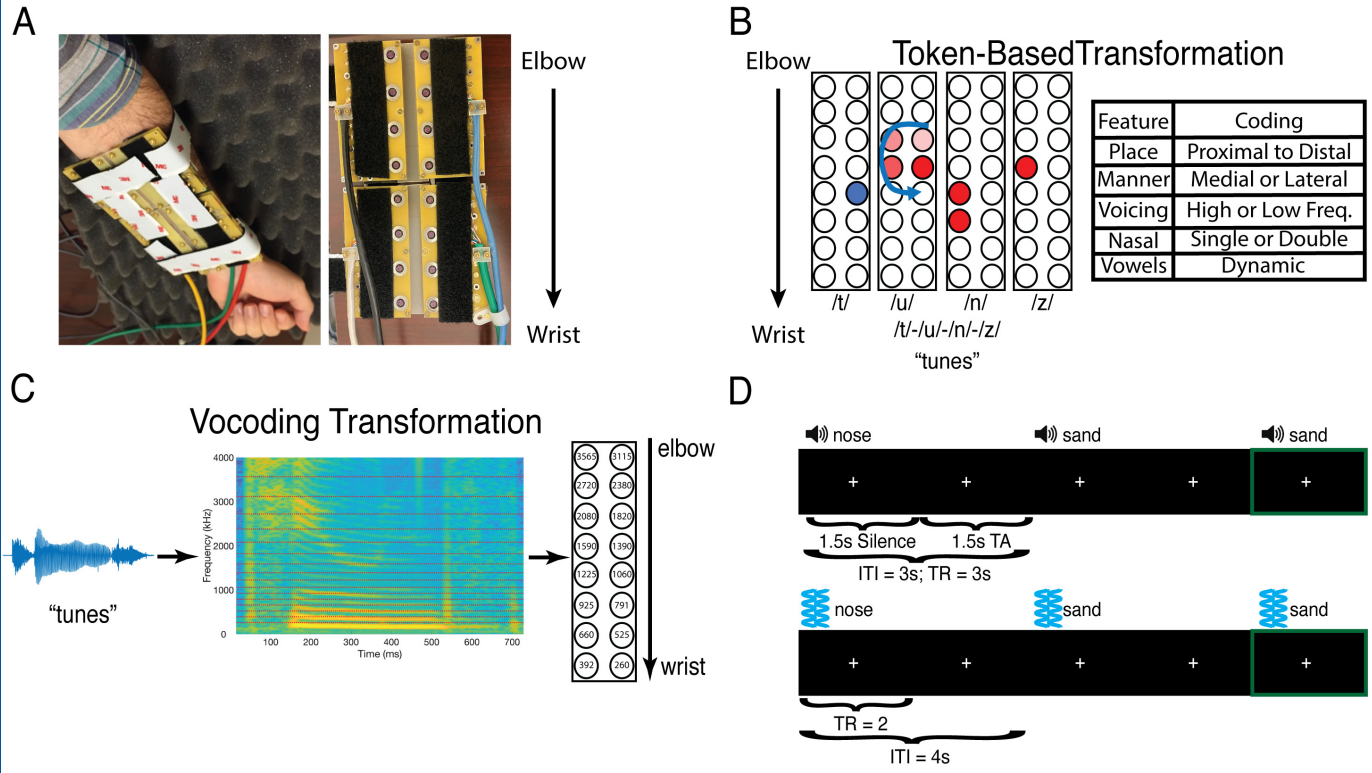
(Glasser et al., 2016) revealed a significant cluster in the left central sulcus (MNI: -40, -19, 42). It also identified two significant clusters in the right hemisphere. The first encompassed right insula and Heschl's gyrus (MNI: 40, -17, 11). The other is on the right STG (MNI: 63, -22, 7). All whole-brain results shown are corrected at two-tailed voxel-wise $\alpha = 0.005$ and cluster-$p \leq 0.05$. Colors reflect across-subject t-statistics. (C-D) Shows the post-pre training correlations for the VT vocoded and token-based groups respectively using an ROI-to-ROI functional connectivity. Color bar reflects the post-pre training difference in functional connectivity between ROIs. A paired t-test was performed to compare changes in functional connectivity post-pre training. Green asterisks mark $p \leq 0.05$ FDR corrected.

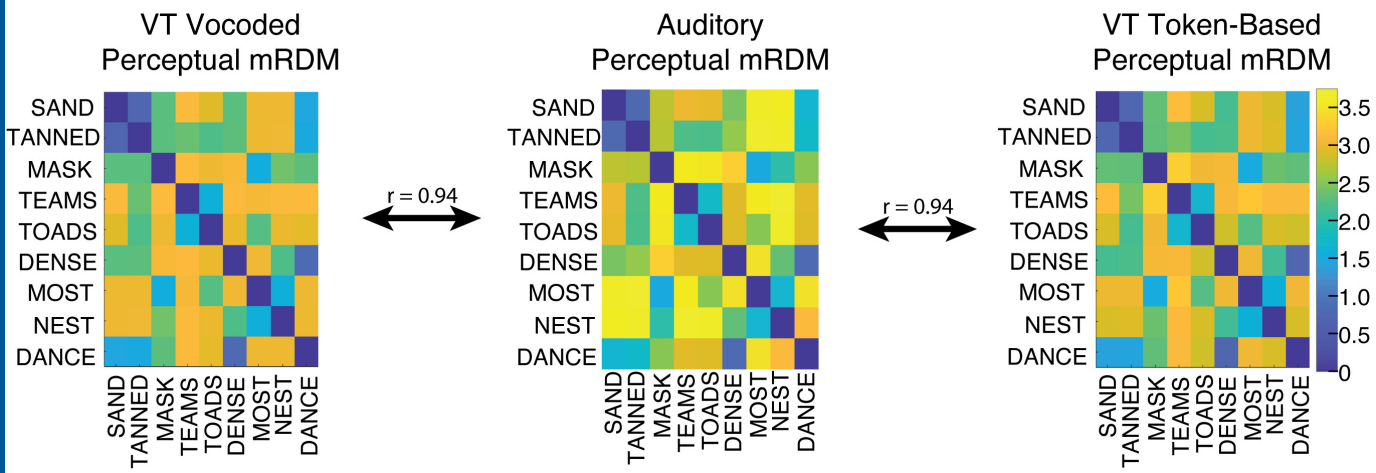
Figure 8: Training Increases Encoding of the VT Token-Based Stimuli in the Left Hippocampus. Linear mixed-effects analysis revealed a significant three-way interaction between Training Phase, Algorithm, and Hemisphere ($\beta = -0.151$, $t_{36} = -3.027$, $p = 0.005$). To investigate this interaction, we created interaction effects plots. (A) The mean Fisher-transformed Pearson correlation between neural and model RDMs estimated from the mixed-effects model for the vocoded group are represented by the opaque lines. For the VT vocoded group, post-hoc tests show a trending difference between pre- and post-training in the right ($t_{30.7} = 2.387$, $p = 0.0902$ Sidak-adjusted) but not the left hippocampus ($t_{30.7} = 0.785$, $p = 0.901$ Sidak-adjusted). (B) The same as (A) but for the token-based group. Post-hoc tests show no significant difference in the right ($t_{30.7} = 0.506$, $p = 0.978$ Sidak-adjusted), but do show a significant difference in the left hippocampus ($t_{30.7} = 3.232$, $p = 0.012$ Sidak-adjusted). Values above each violin reflect the uncorrected p-value from a one-sample t-test against 0. Semi-transparent lines reflect raw individual subject correlations from either the left (teal) or right (orange) hippocampus. Horizontal lines in the violin plots reflect the median. Green asterisk and orange tilde mark significant ($p \leq .05$) and trending ($p \leq .1$) differences, respectively, after multiple comparisons correction.

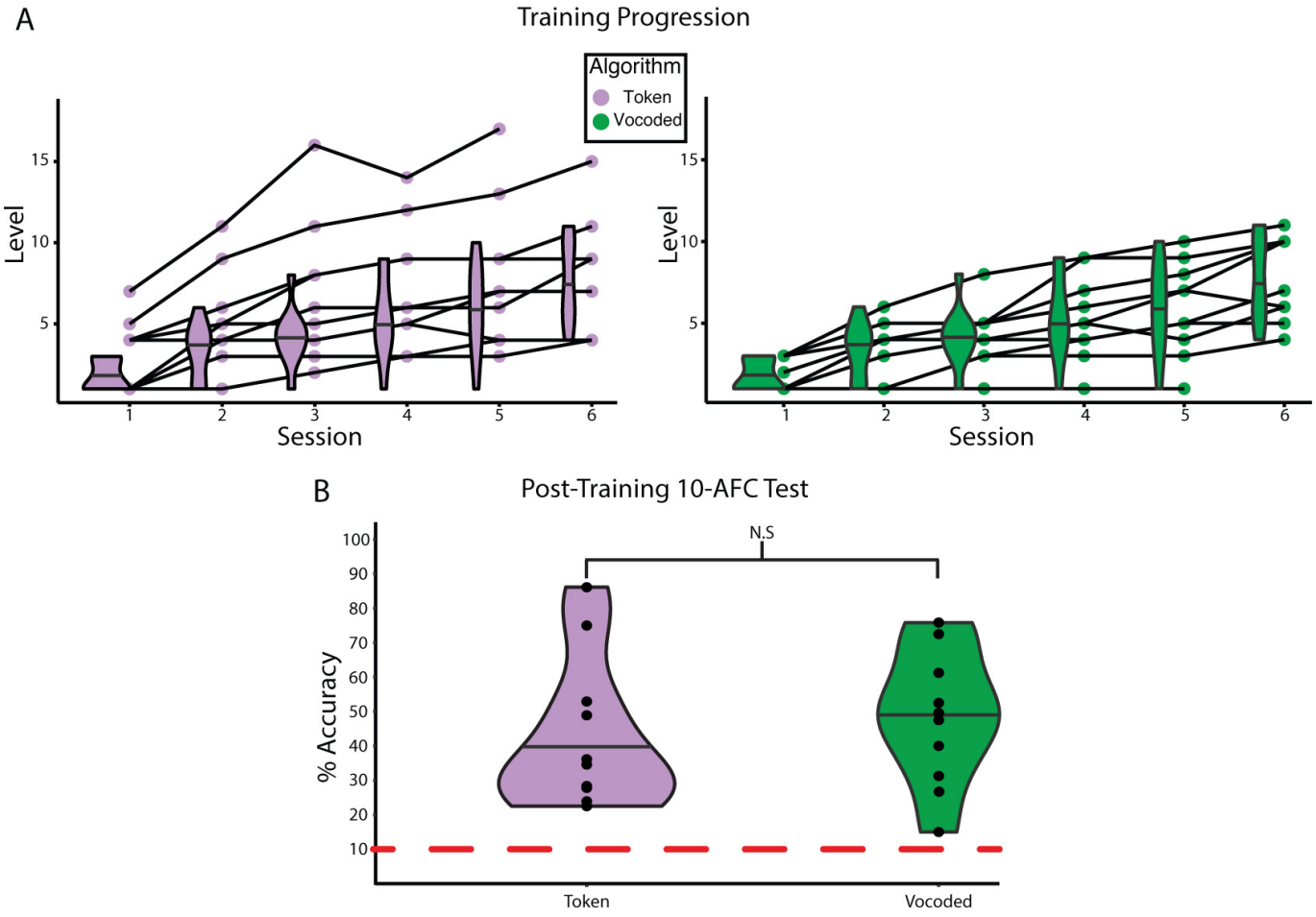
Table 1: Breakdown of word stimuli presented to participants

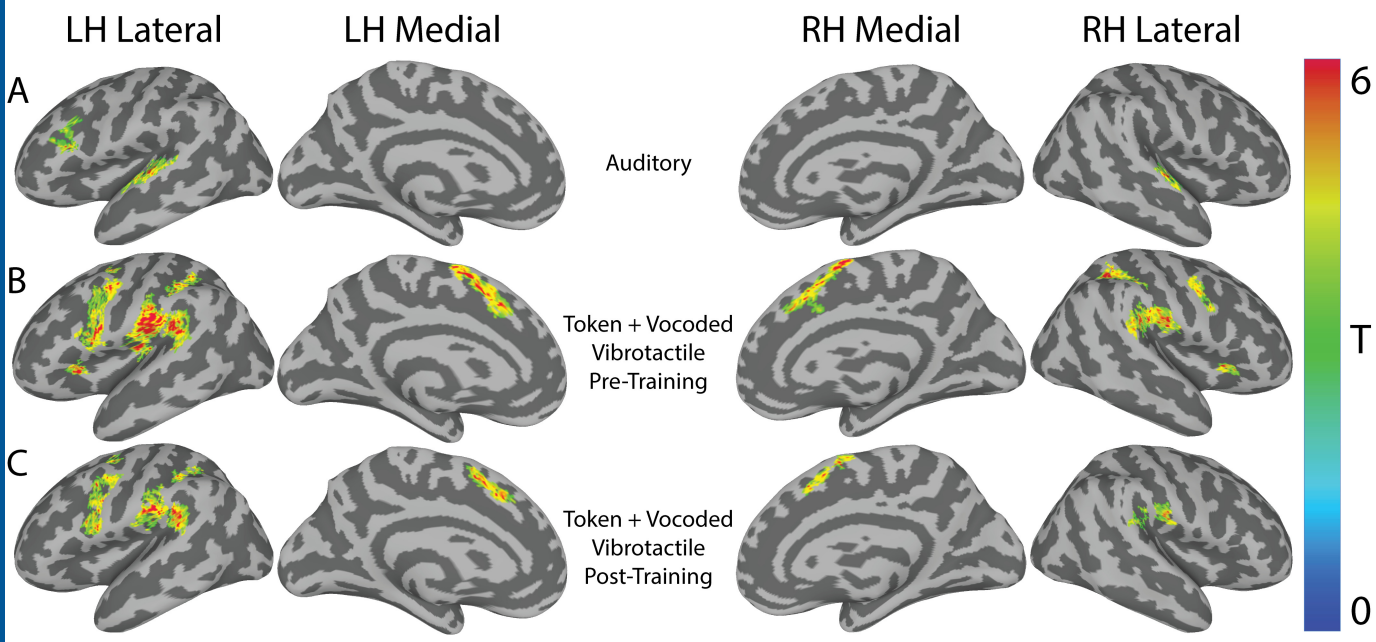
Table 2: Location for all regions with significant activation vs baseline. Clusters are thresholded at a voxel-wise $\alpha < 0.001$ and cluster-level $p < 0.05$, FWE corrected.

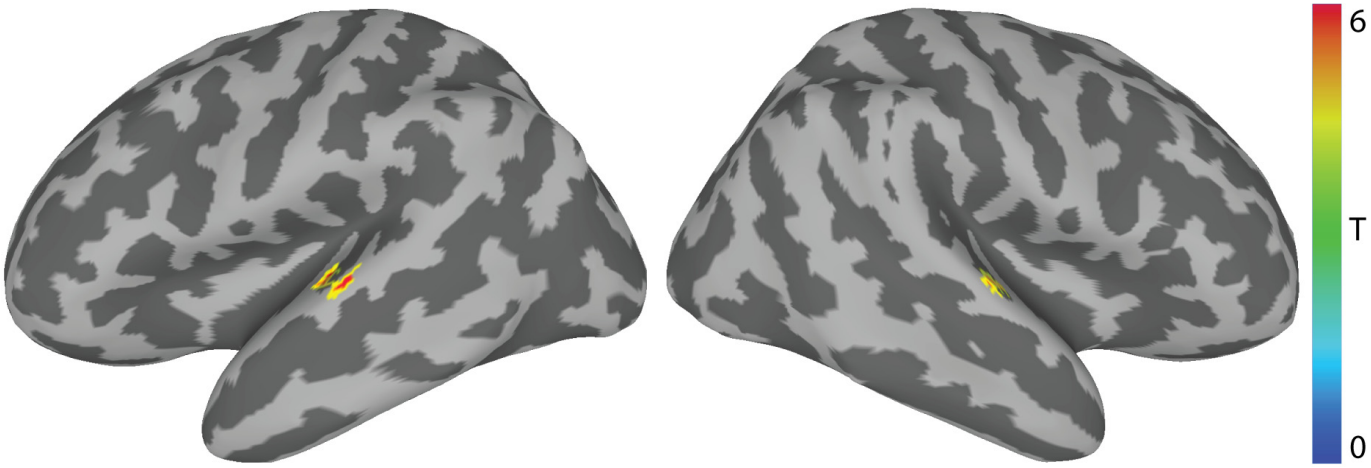
Table 3: Location for all regions with significant training-related changes in seed-to-voxel functional connectivity in the VT vocoded group. Clusters are thresholded at a voxel-wise $\alpha < 0.001$ and cluster-level $p < 0.05$, FWE corrected.

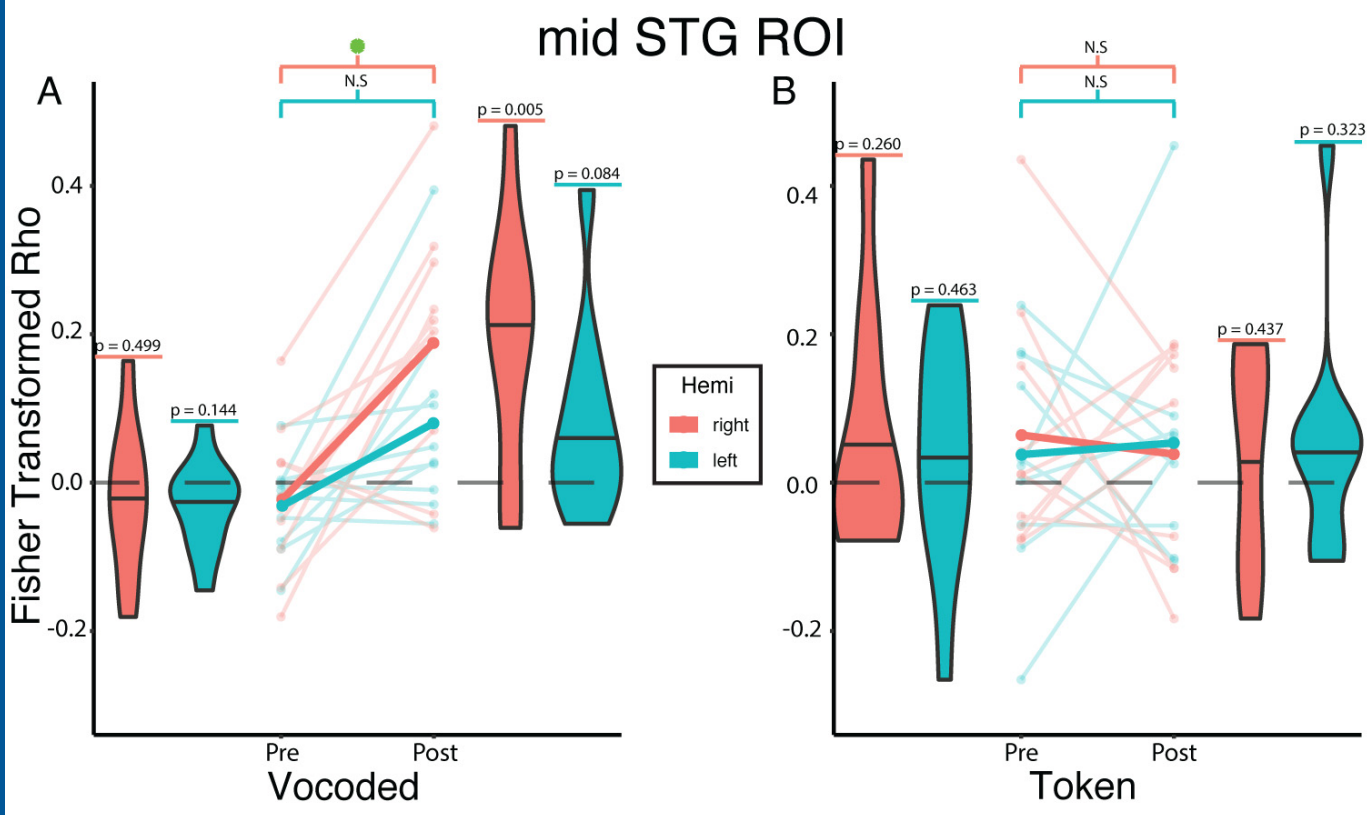


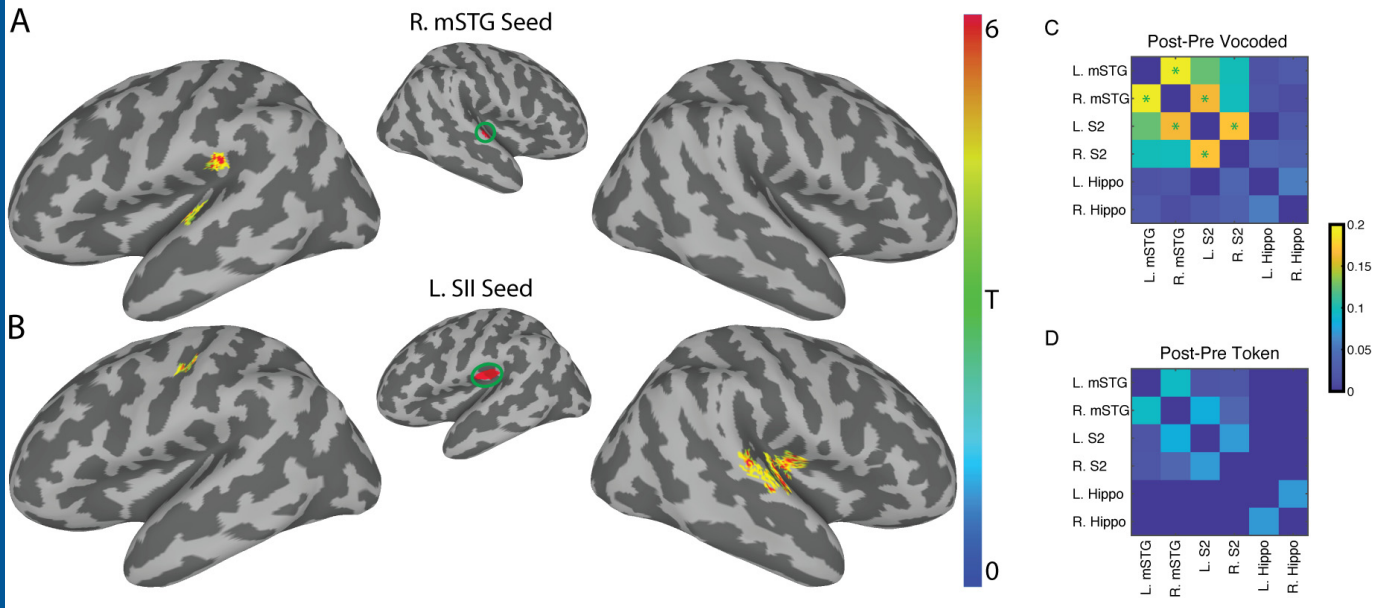












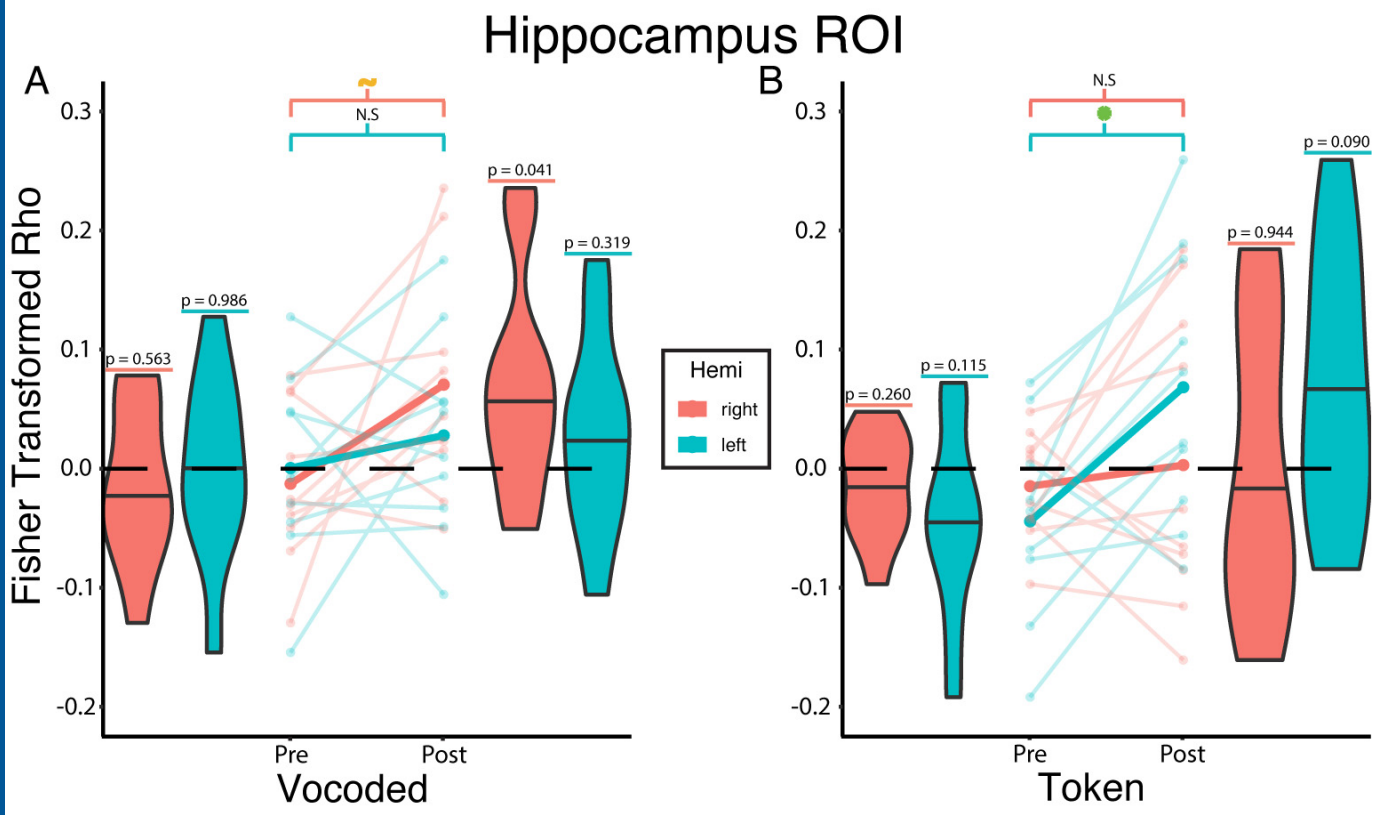


Table 1

	All Stimuli	
Trained	CVCC	Sand, tanned, mask, teams, toads, dense, most, nest, dance
	CCVC	Spit, spin, stoop
	CVC	Meat, peace, nose
Untrained	CVCC	Send, tend, max, seems, zones, nets, meant, mist, maps
	CCVC	Snip, skin, stoke
	CVC	Peat, knees, soak

Table 2

Scan	Hemi	**HCP-MMP1.0 ROI (Talairach-Tournoux Atlas)**	Cluster p-Value	Center of Mass Coordinates (MNI)		
				x	y	z
Auditory	RH	Parabelt Complex (Superior Temporal Gyrus)	0.001	57	-13	3
	LH	Parabelt Complex (Superior Temporal Gyrus)	0.001	-56	-19	5
		Auditory 5 Complex (Superior Temporal Gyrus)	0.001	-62	-36	7
Pre-Training	RH	Area PF Complex (Inferior Parietal Lobule)	0.001	55	-25	24
		Anterior Intraparietal Area (Inferior Parietal Lobule)	0.001	39	-39	42
		Supplementary and Cingulate Eye Field (Medial Frontal Gyrus)	0.001	8	13	52
		Premotor Eye Fields (Middle Frontal Gyrus)	0.001	51	2	41
		Anterior Ventral Insular Area (Insula)	0.001	30	25	3
	LH	Area OP1/SII (Inferior Parietal Lobule)	0.001	-52	-27	23
		Rostral Area 6 (Precentral Gyrus)	0.001	-50	2	28
		Supplementary and Cingulate Eye Field (Superior Frontal Gyrus)	0.001	-8	9	54
		Anterior Intraparietal Area (Inferior Parietal Lobule)	0.001	-45	-38	42
		Anterior Ventral Insular Area (Insula)	0.001	-30	25	7
		Frontal Eye Fields (Middle Frontal Gyrus)	0.002	-30	-3	48
Post-Training	RH	Retroinsular Cotex (Inferior Parietal Lobule)	0.001	53	-32	25
		Supplementary and Cingulate Eye Field (Medial Frontal Gyrus)	0.001	7	15	49
		Area PF Opercular (Post Central Gyrus)	0.003	57	-16	22
		Area Posterior 24 Prime (Medial Frontal Gyrus)	0.019	7	2	65
	LH	Rostral Area 6 (Precentral Gyrus)	0.001	-48	2	29
		Area PF Opercular (Post Central Gyrus)	0.001	-59	-22	25
		Area PF Complex (Inferior Parietal Lobule)	0.001	-50	-40	26
		Supplementary and Cingulate Eye Field	0.001	-9	14	49

		(Medial Frontal Gyrus)				
		Area 6 Anterior (Middle Frontal Gyrus)	0.001	-29	-5	48
		Anterior Intraparietal Area (Inferior Parietal Lobule)	0.002	-47	-35	42
		Anterior Intraparietal Area (Inferior Parietal Lobule)	0.002	-35	-44	40

Table 3

Seed ROI	Hemi	HCP-MMP1.0 ROI (Talairach-Tournoux Atlas)	Cluster p-Value	Center of Mass Coordinates (MNI)		
				x	y	z
lS2	RH	Insular Granular Complex (Insula)	0.001	40	-17	11
		Auditory 5 Complex (Superior Temporal Gyrus)	0.001	63	-22	7
	LH	Primary Motor Cortex (Precentral Gyrus)	0.012	-40	-19	42
lSTG	RH	Lateral Belt Complex (Superior Temporal Gyrus)	0.001	53	-18	6
rS2	RH	Posterior Insular Area 2 (Insula)	0.017	37	-8	6
	LH	Area OP2-3/VS (Insula)	0.026	-42	-16	20
rSTG	LH	Area PF_{cm} (Postcentral Gyrus)	0.026	-55	-28	21
		Lateral Belt Complex (Superior Temporal Gyrus)	0.044	-50	-19	7