

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

3
4 Srikanth R. Damera¹, Patrick S. Malone¹, Benson W. Stevens¹, Richard Klein¹, Silvio P. Eberhardt², Edward T.
5 Auer Jr.², Lynne E. Bernstein², Maximilian Riesenhuber¹

6
7 1 - Department of Neuroscience, Georgetown University Medical Center, Washington DC, USA

8 2 - Department of Speech Language & Hearing Sciences, George Washington University, Washington DC,
9 USA

10
11 Correspondence should be addressed to M.R. (mr287@georgetown.edu).

12 Department of Neuroscience

13 Georgetown University Medical Center

14 Research Building Room WP-12

15 3970 Reservoir Rd. NW

16 Washington, DC 20007, USA

17
18 **Figure Count**

19 7 Figures

20 2 Supplementary figures, 4 Supplementary tables

21
22 **Word Count**

23 Summary: 211

24 Introduction: 1183

25 Discussion: 1820

26
27 **Competing Interests**

28 The authors declare no competing financial interests

29 **Summary**

30 It has been postulated that the brain is organized by “metamodal”, sensory-independent cortical modules
31 implementing particular computations, leading to the intriguing hypothesis that brain areas can perform tasks
32 (such as word recognition) not just in “standard” sensory modalities but also in novel sensory modalities. Yet,
33 evidence for this theory, especially in neurotypical subjects, has been variable. We hypothesized that effective
34 metamodal engagement of a brain area requires congruence between the novel and standard sensory
35 modalities not only at the task level (e.g., “word recognition”) but critically also a match at the algorithmic level
36 (in Marr’s terminology), i.e., at the level of neural representation of the information of interest. To test this
37 hypothesis, we trained participants to recognize vibrotactile versions of auditory words using two encoding
38 schemes. The vocoded approach preserved the dynamics and representational similarities of auditory speech
39 while the token-based approach used an abstract phoneme-based code. Although both groups learned the
40 vibrotactile word recognition task, only in the vocoded group did trained vibrotactile stimuli recruit the auditory
41 speech network and lead to increased coupling between somatosensory and auditory speech areas. In
42 contrast, the token-based encoding appeared to rely on paired-associate learning. Thus, matching neural input
43 representations is a critical factor for assessing and leveraging the metamodal potential of cortical modules.

14 Introduction

15 The dominant view of brain organization revolves around cortical areas dedicated for processing information
16 from specific sensory modalities. However, emerging evidence over the past two decades has led to the idea
17 that many sensory cortical areas are “metamodal”, i.e., characterized by computations that are task-specific,
18 yet invariant to sensory modality (Heimler et al., 2015; Pascual-Leone and Hamilton, 2001). Early evidence for
19 sensory modality-invariant processing in cortical areas comes from cross-modal plasticity studies in sensory-
20 deprived populations (Rauschecker, 1995; Rauschecker et al., 1992; Sadato et al., 1996; Théoret et al., 2004).
21 These studies showed that cortical areas traditionally considered to be dedicated to unisensory processing
22 could be recruited by stimuli from another, non-preferred sensory modality. Interestingly, although broad
23 swaths of cortex no longer received the normal sensory input in these populations, non-preferred modality
24 stimuli activated the specific cortical areas normally relevant for a particular task in the preferred sensory
25 modality, such as localization and recognition (Amedi et al., 2007; Bi et al., 2016; Bola et al., 2017, 2020;
26 Lomber et al., 2010; Meredith et al., 2011; Ptito et al., 2005; Reich et al., 2011; Renier et al., 2014; Striem-Amit
27 et al., 2012). This task-specific cross-modal engagement is thought to reflect a functional unmasking of existing
28 anatomical connections. Importantly, there is evidence for cross-modal engagement of traditionally unisensory
29 areas even in neurotypical individuals (Amedi et al., 2007; Renier et al., 2005, 2010; Siuda-Krzywicka et al.,
30 2016) – thereby opening the door to recruiting previously established sensory processing pathways for novel
31 sensory modalities. A prime example of this process is reading, which is initially thought to recruit auditory
32 speech processing pathways through grapheme-to-phoneme conversion (Pugh et al., 2001), and the same
33 idea has given rise to promising therapeutic applications such as sensory substitution devices (SSDs, which,
34 for instance enable processing of visual information in blind individuals by translating camera input to acoustic
35 stimuli (Bach-y-Rita and Kercel, 2003; Meijer, 1992). Yet, other studies (Benetti et al., 2017, 2020; Bola et al.,
36 2017; Fairhall et al., 2017; Mattioni et al., 2020; Pietrini et al., 2004; Twomey et al., 2017; Vetter et al., 2020)
37 have failed to find or have found far less robust evidence of cross-modal engagement in neurotypical subjects,
38 raising the critical question of the conditions under which a particular sensory area can be successfully
39 recruited for “metamodal” processing.

70 Current theories for the “metamodal” brain differ on the necessary requirements for the engagement of a
71 sensory area. Prior studies (Amedi et al., 2002, 2007; Striem-Amit et al., 2012) have emphasized that
72 metamodal engagement of a cortical area, in addition to the presence of task-relevant connectivity (Hannagan
73 et al., 2015; Mahon and Caramazza, 2011; Saygin et al., 2012, 2016), depends on a correspondence of
74 preferred and non-preferred modality stimuli. Yet, it has been challenging to elucidate the nature of this
75 necessary “correspondence” for metamodal engagement. For example, the fact that (visual) shape-selective
76 ventral occipito-temporal cortex (VOTC) is engaged in subjects trained on a visual-to-auditory SSD designed to
77 transform visual images to sounds has been used to argue that this region responds to any modality that
78 conveys shape information (Amedi et al., 2007; Hannagan et al., 2015; Heimler et al., 2015; Striem-Amit et al.,
79 2012). Consequentially, failures to find metamodal engagement are due to an absence of shape information in
30 the stimuli encoded in the non-standard modality. Yet, in the absence of a clear definition of auditory “shape”

31 such theories face reverse inference challenges and are difficult to test, let alone falsify. In terms of Marr's
32 levels of analysis (Computational, Algorithmic, Implementational), what is needed is a consideration of the
33 Algorithmic level that serves as a crucial bridge between the Computational level of task goals, and the
34 Implementational level of neural responses (Barsalou, 2017; Hagoort, 2020). Here we propose that successful
35 metamodal coupling requires not just correspondence at the Computational/task level, but also a match of the
36 novel modality representation to the standard modality representation at the Algorithmic level. This theory
37 concretizes and makes testable the helpful yet vague notion of "correspondence" between preferred and non-
38 preferred modality representations by tying it to a correspondence of representational spaces. A failure to
39 achieve this *representational* correspondence would then be predicted to impede or even preclude metamodal
30 engagement.

31 In the present study, we test the hypothesis that cross-sensory recruitment of existing learned sensory
32 processing pathways critically depends on this representational match. Specifically, we focused on auditory-to-
33 tactile sensory substitution. This field has a long history dating back to the invention of the Tadoma method
34 (Alcorn, 1945) – a method whereby deaf individuals learn to perceive auditory speech received via vibrotactile
35 (VT) input from their fingers which are placed over the articulators of a speaker. Over a century of work on
36 auditory-to-tactile sensory substitution has led to the development of VT speech aids (Gault, 1924, 1926).
37 These devices have been used successfully to teach both deaf and hearing individuals to recognize auditory
38 speech through touch (Bernstein et al., 1991; Brooks and Frost, 1983; Cieřla et al., 2019). Here we used such
39 a device to train two neurotypical groups of adult subjects on the same word recognition task, with each group
30 being trained with one of two auditory-to-VT sensory substitution algorithms. One algorithm was designed to
31 preserve as much of the temporal dynamics of auditory speech as possible ("vocoded speech"), aiming to
32 achieve a neural congruence between vibrotactile speech stimuli and auditory speech representations in brain
33 areas that are part of the auditory speech system. The other algorithm ("token") used a code in which specific
34 VT patterns corresponded to specific phonetic features (Chomsky and Halle, 1968; Reed et al., 2018).
35 Interestingly, at the behavioral level, subjects in both algorithm groups learned to associate VT patterns with
36 spoken words at an equivalent level after training. However, fMRI analyses revealed critical differences in the
37 cross-modal recruitment of brain areas between the two groups, with only the vocoded encoding group
38 showing metamodal engagement of auditory speech processing areas, specifically the areas whose neural
39 representations of auditory speech representation well matched the representational similarity of the
30 vibrotactile word stimuli. Consistent with these findings, functional connectivity analyses showed that increased
31 coupling between the auditory and somatosensory cortex after training also depended on the nature of the
32 input representations produced by the different VT algorithms. These findings suggest that metamodal
33 engagement of a cortical area is dependent not only on its task-relevant anatomical connectivity but also on
34 the match at the level of representational encoding between the standard and novel modalities. Adopting the
35 nomenclature of David Marr's levels (Marr, 1982), our data show that a mere congruence at the highest,
36 computational level (e.g., VT stimuli corresponding to auditory words) is insufficient for metamodal
37 engagement. Rather, metamodal coupling requires a congruence at the algorithmic level (e.g., a match in
38
39
40
41
42
43
44
45
46
47

18 neural representations). Thus, our study not only critically advances our understanding of metamodal
19 engagement and thus general principles of brain organization, but also opens the door to designing more
20 efficient sensory substitution algorithms that better interface with existing cortical processing pathways (as in
21 the present study, where the algorithmically matched vocoded speech representation conveyed ~1.2 times as
22 much information per unit time than the non-matched one).

23 **Materials and Methods**

24 **Participants**

25 We recruited a total of 22, right-handed, healthy, native English speakers in this study (ages 18-27, 12
26 females). Georgetown University's Institutional Review Board approved all experimental procedures, and
27 written informed consent was obtained from all subjects before the experiment. We excluded 4 subjects from
28 the auditory scan due to excessive motion (>20% of volumes) and 2 subjects from the vibrotactile (VT) scans
29 because they failed to complete the training. As a result, we analyzed a total of 18 subjects for the auditory
30 scans and 20 for the VT scans.

31 **Stimulus Selection**

32 A set of word stimuli was developed according to the following criteria: 1) short monosyllabic stimuli (~4
33 phonemes); 2) only contain phonemes from a limited subset of English consonants (8 consonants and 6
34 vowels); 3) set containing items predicted to be perceptually unique and therefore learnable; and 4) words that
35 span the VT vocoder perceptual space (see below). To develop the set meeting these criteria we utilized a
36 computational modeling approach based on the methods described in (Auer and Bernstein, 1997). Existing
37 tactile consonant and vowel perceptual identification data (Bernstein, unpublished) were used in combination
38 with the PhLex lexical database (Seitz, Bernstein, Auer, & MacEachern, 1998) to model the lexical perceptual
39 space. In outline, the modeling steps are: (1) Transform phoneme identification data into groupings of
40 phonemes as a function of a set level of dissimilarity; (2) Re-transcribe a phonemically transcribed lexical
41 database so that all of the words are represented in terms of only the phonemic distinctions across groupings;
42 and (3) Collect words that are identical under the re-transcription and count how many are in each collection. In
43 this study, the lexical equivalence class (LEC) size –the number of words in a collection—was set to three.
44 Only words that were accompanied by three or fewer other words following re-transcription were considered
45 candidates for the study. Words in smaller LECs are predicted to be perceptually easier (more unique) than
46 words in larger LECs, which offer more opportunities for confusions.

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

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

62 **Behavioral Training**

63 The training paradigm used an N-alternative forced choice (N-AFC) task and a leveling system organized in
64 sets of 3 to facilitate training progression. In each set of 3 levels, the number of choices (N) in the N-AFC task
65 was kept constant, but the choices themselves were increasingly confusable. The number of choices N was
66 increased by 1 when progressing between each set of 3 levels. The first level utilized a 2-AFC task, and the
67 final level (level 15) utilized an 8-AFC task. An accuracy of 80% was required to advance to the next level.
68 Subjects performed each training session in a quiet room while listening to an auditory white noise stimulus
69 through over-the-ear headphones. Auditory white noise was presented in order to mask the mechanical sound
70 of the VT stimulation. At the beginning of each trial, the orthographic labels for the word choices were
71 displayed on the screen, and a VT stimulus was played after a short delay. Participants then indicated which
72 label corresponded to the VT stimulus. Feedback was given after each trial, as well as an opportunity to replay
73 any of the word choices. Subjects completed a total of 6 training sessions, followed by a post-training fMRI
74 scan. After their post-training fMRI scan, subjects performed a final 10-AFC task.

75 **Description of VT Device**

76 A (20cm x 11.0 cm) 16-channel MRI-compatible vibrotactile stimulator array was organized as 2 rows of 8
77 stimulators (Fig. 1A), with center-to-center stimulator spacing of 2.54 cm. To ensure that the stimulators would
78 maintain contact with the volar forearm, the array comprised four rigid modules connected with stiff plastic
79 springs. Velcro straps were used to mount the device to the arm firmly while bending the array to conform to
80 the arm's shape. With no applied voltage to the piezoelectric bimorphs, the contactors were flush with the
81 circuit board surface facing the skin. During operation, a constant +57-V voltage applied to all stimulators
82 retracted the contactors into the surround, and each applied -85-V pulse drove the contactor into the skin. All
83 pulses were identical. The drive signal was a square wave, with a pulse time of 2 ms, and with unpowered
84 intervals of 1ms between power reversals to protect the switching circuitry. The display's control system
85 comprised the power supplies (-85V, +57V), high voltage switching circuits to apply these voltages to the
86 piezoelectric bimorphs, and a digital control system that accepted from a controlling computer's serial COM
87 port the digital records specifying a stimulus (comprising the times and channels to output pulses on), and a
88 command to initiate stimulus output.

89 **VT Vocoder Speech Encoding**

90 This real-time vocoder was used to convert acoustic speech signals into VT stimuli. The initial stage of the
91 vocoder comprised a bank of filters whose output power was used to control the output of VT pulses. The VT
92 display (Fig. 1A) used a frequency-to-place mapping algorithm: The energy passed by each filter of the
93 vocoder was used to modulate the vibration of a specific MRI-compatible transducer on the 16-channel VT
94 device (Fig. 1A and 1C) placed on the volar forearm (Malone et al., 2019). Low frequencies mapped to

35 transducers near the wrist, and higher frequencies mapped to transducers near the elbow. If the energy within
36 a given filter exceeded a fixed threshold at a given time point, a VT pulse was emitted from the corresponding
37 transducer. The basic hardware design and software algorithms for the vocoder are referred to in (Bernstein et
38 al., 1991) as the “GULin” vocoder algorithm. Briefly, 16 bandpass filters with frequencies centered at 260, 392,
39 525, 660, 791, 925, 1060, 1225, 1390, 1590, 1820, 2080, 2380, 2720, and 3115 Hz, with respective
40 bandwidths of 115, 130, 130, 130, 130, 130, 145, 165, 190, 220, 250, 290, 330, 375, and 435 Hz. An additional
41 high-pass filter with cutoff 3565 Hz is also used. The energy detected in each band is used to amplitude-
42 modulate a fixed-frequency sinewave at the center frequency of that band (and at 3565 Hz in the case of the
43 high-pass filter). The combination of the 16 sinewaves comprises the vocoded acoustic signal, and the
44 resulting activation pattern over the 16 transducers constituted its vibrotactile instantiation.

35 **Token-based VT Speech Encoding**

36 The same 16-channel VT device was used to present subjects with the token-based stimuli. Token-based
37 stimuli were constructed based on prior work (Reed et al., 2018) and reflect the idea that spoken words can be
38 described as a string of phonemes. Phonemes in turn can be uniquely described by a set of phonetic features.
39 Therefore, each phonetic feature was assigned a unique VT pattern. In this study, we used place, manner, and
40 voicing features to describe phonemes (Fig. 1C). Place was coded as patterns that occurred either proximal or
41 distal to the wrist. Stop and fricative manner features were coded as patterns that occurred either medial or
42 lateral to the body respectively. The nasal manner feature was distinguished by driving two channels instead of
43 one for stops and fricatives. Voicing was coded as either driving high frequency vibrations (250Hz) or low
44 frequency vibrations (100Hz). Vowels were coded in a similar feature-based manner, but were dynamic stimuli
45 (e.g. swirls and sweeps) whereas consonants were static. Importantly, all consonant patterns lasted 120ms
46 and all vowel stimuli lasted 220ms and there was a 100ms gap between each pattern. As a result, token-based
47 stimuli were either 660ms or 880ms long. CVCC trained token-based stimuli used in fMRI analyses were
48 880ms long while their VT vocoded counterparts had a mean duration of 727ms and standard deviation of
49 91.6ms. A paired t-test revealed that token-based stimuli were significantly longer ($t(8) = 4.99$; $p = 0.001$) than
50 their vocoded counterparts. Thus, not only did VT vocoded but not token based stimuli preserve the temporal
51 dynamics found in auditory speech, but they also conveyed more information per unit time.

22 **Auditory Scan**

23 *fMRI Experimental Procedures*

24 EPI images from nine event-related runs were collected using a clustered acquisition paradigm. Within each
25 run, 30 words were presented three times in random order for a total of 90 trials. Each trial was 3s long and
26 started with 1.5s of volume acquisition followed by the auditory word (during the silent period, see below, “Data
27 Acquisition”; Fig. 1D). To maintain attention, subjects performed a 1-back task in the scanner: Subjects were
28 asked to press a button in their left hand whenever the same word was presented on two consecutive trials.
29 These catch trials comprised ten percent of the trials in each run. Furthermore, an additional ten percent of
30 trials were null trials. During these trials, which lasted for 3s, no words were presented. In total, there were 118

31 trials per run, with each trial lasting 3s for a total of 354s, plus an additional 15s fixation at the start of the run.
32 Thus, in total each run lasted 369s and the session lasted 43min.

33 *Data Acquisition*

34 MRI data were acquired at the Center for Functional and Molecular Imaging at Georgetown University on a 3.0
35 Tesla Siemens Trio Scanner. We used whole-head echo-planar imaging sequences (flip angle = 90°, TE = 30
36 ms, FOV = 205, 64x64 matrix) with a 12-channel head coil. A clustered acquisition paradigm (TR = 3000 ms,
37 TA = 1500 ms) was used such that each image was followed by an equal duration of silence before the next
38 image was acquired. 28 descending axial slices were acquired in descending order (thickness = 3.5 mm, 0.5
39 mm gap; in-plane resolution = 3.0x3.0 mm²). This sequence was used in previous auditory studies from our lab
40 (Chevillet et al., 2013). A T1-weighted MPRAGE image (resolution 1x1x1mm³) was also acquired for each
41 subject.

42 **VT Scan**

43 *fMRI Experimental Procedures*

44 EPI images from six event-related runs were collected. Within each run 30 stimuli (15 from the training set and
45 15 additional words) were presented three times in random order for a total of 90 trials. A 4 second intertrial
46 interval was used (Fig. 1D). As in the auditory scan, to maintain attention, subjects performed a 1-back task in
47 the scanner: Subjects were asked to press a button in their left hand whenever the same stimulus was
48 presented on two consecutive trials. These catch trials comprised ten percent of the trials in each run.
49 Furthermore, an additional ten percent of trials were null trials during which subjects were presented with a
50 blank screen for 3s. In total, there were 111 trials per run with each trial lasting 4s for a total of 444s plus an
51 additional 10s fixation at the start and end of the run. Thus, in total each run lasted 464s and the session lasted
52 46min.

53 *Data Acquisition*

54 MRI data were acquired at the Center for Functional and Molecular Imaging at Georgetown University on a 3.0
55 Tesla Siemens Trio Scanner. We used whole-head echo-planar imaging sequences (TR = 2000ms, flip angle =
56 90°, TE = 30 ms, FOV = 205, 64x64 matrix) with a 12-channel head coil. 33 interleaved descending axial slices
57 were acquired (thickness = 3.5 mm, 0.5 mm gap; in-plane resolution = 3.0x3.0 mm²). A T1-weighted MPRAGE
58 image (resolution 1x1x1mm³) was also acquired for each subject.

59 *fMRI Data Preprocessing*

60 Image preprocessing was performed in SPM12 (<http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>) and AFNI.
61 The first four acquisitions of each run were discarded to allow for T1 stabilization, and the remaining EPI
62 images were slice-time corrected to the middle slice for the VT scans. No slice-time correction was performed
63 for the auditory scans due to using a clustered acquisition paradigm due to temporal discontinuities between
64 successive volumes (Perrachione and Ghosh, 2013). These images were then spatially realigned and
65 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 (Fischl et al., 1999) was used to reconstruct cortical surface models including an outer pial and inner white-matter surface. These surfaces were then brought into the SUMA environment and fit to a standardized mesh based on an icosahedron with 64 linear divisions using AFNI's Maplcoshedron 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).

Representational Similarity Analysis (RSA)

Constructing Model Representational Dissimilarity Matrices (mRDMs)

Two candidate mRDMs were generated: an auditory perceptual mRDM, and a VT vocoded perceptual mRDM. These mRDMs were generated by modifying an edit mRDM which was generated using an edit distance metric between word pairs in the stimulus set. Here, 1 edit was considered a substitution, insertion, or deletion of a single phoneme. 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 matches those found in both higher order STG speech regions and speech recognition-specific representations learned in later layers of a deep neural network. The auditory and VT vocoded perceptual mRDMs were similarly created using an edit distance but now weighting phoneme edit by either its auditory or VT vocoded perceptual confusability. Auditory and VT vocoded perceptual phoneme confusability was derived from a behaviorally measured perceptual auditory and VT vocoded phoneme identification task. This confusability 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 (Fig. 1B). This resulted in a 9-by-9 auditory and VT perceptual mRDM for the CVCC trained words (Fig. 1E). These representational spaces are highly correlated ($r = 0.94$) and reflect the close representational congruence between auditory and VT vocoded stimuli.

Whole-Brain Searchlight RSA Analysis

RSA (Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008) analyses were performed using the CoSMoMvpa toolbox (Oosterhof, Connolly, & Haxby, 2016), Surfing Toolbox (Oosterhof et al., 2011) and custom MATLAB code. Searchlights were constructed around each surface node by selecting the 30 closest voxels measured by geodesic distance. 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 normally distributed (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 in functionally defined ROIs identified in the auditory scans. To do so, we averaged the Fischer z-transformed correlations of searchlights in a given 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. For both ROIs, we fit the maximal model that included three main effects, 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. (Glasser et al., 2016) The final model is shown below:

$$\text{Correlation} \sim 1 + \text{TrainingPhase} + \text{Algorithm} + \text{Hemi} + \text{TrainPhase: Algorithm} + \text{Algorithm: Hemi} + \text{TrainPhase: Hemi} + \text{TrainPhase: Algorithm: Hemi} + (1 + \text{TrainingPhase} | \text{Subj})$$

The reference group corresponding to the intercept was specified as pre-training, token-based, right-hemisphere. All β s 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 Tukey's method.

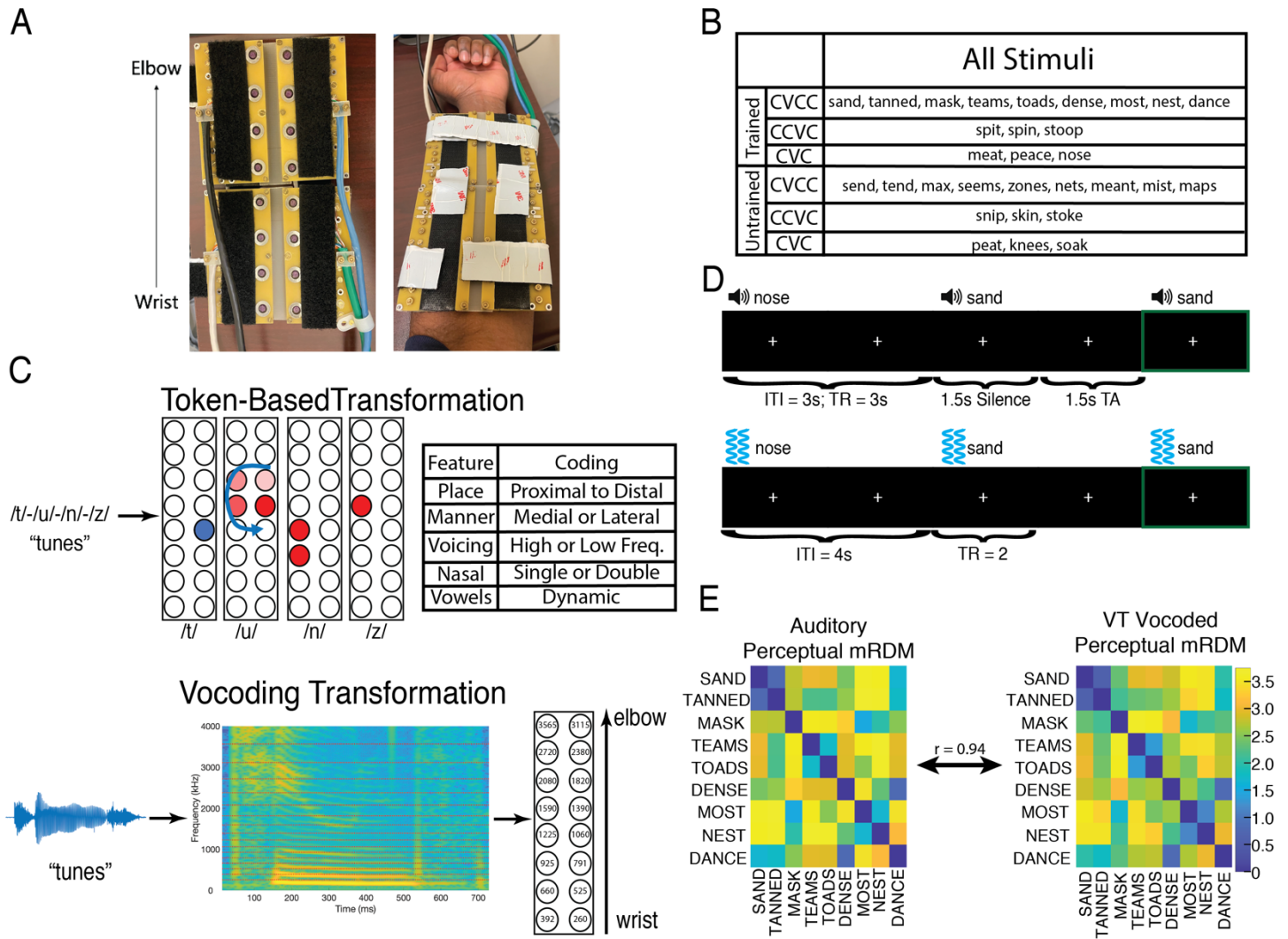
Task-Related Functional Connectivity

Functional connectivity analyses were performed using the CONN-fMRI toolbox (Whitfield-Gabrieli & Nieto-Castanon, 2012). To do so, 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, Liu, & 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 in order to perform group analyses.

32 **Whole-Brain Statistical Correction**

33 We tested the group-level significance of whole-brain RSA analyses as well as functional connectivity
34 differences by first computing a t-statistic at each node on the standard surface. To correct these t-statistic
35 maps for multiple comparisons, we first estimated the smoothness of the data for each analysis in each
36 hemisphere using the AFNI/SUMA *SURFFWHM* command. We then used this smoothness estimate to
37 generate noise surface maps using the AFNI/SUMA *slow_surf_clustsim.py* command. This then allowed us to
38 generate an expected cluster size distribution at various thresholds that we compared clusters in our actual
39 data to. For the auditory scan, we performed a one-sample t-test against 0 and applied a two-tailed cluster-
40 defining threshold of $\alpha = .001$. For the functional connectivity analyses in the VT scan, we performed a two-
41 sample paired t-test to seed-to-voxel functional connectivity in subjects pre- and post-training. We applied a
42 two-tailed cluster-defining threshold of $\alpha = .005$. All resulting clusters were corrected at the $p \leq .05$ level.
43 Tables report the coordinates of the center of mass of clusters in MNI space and their location as defined by
44 the Glasser Atlas (Glasser et al., 2016).

45



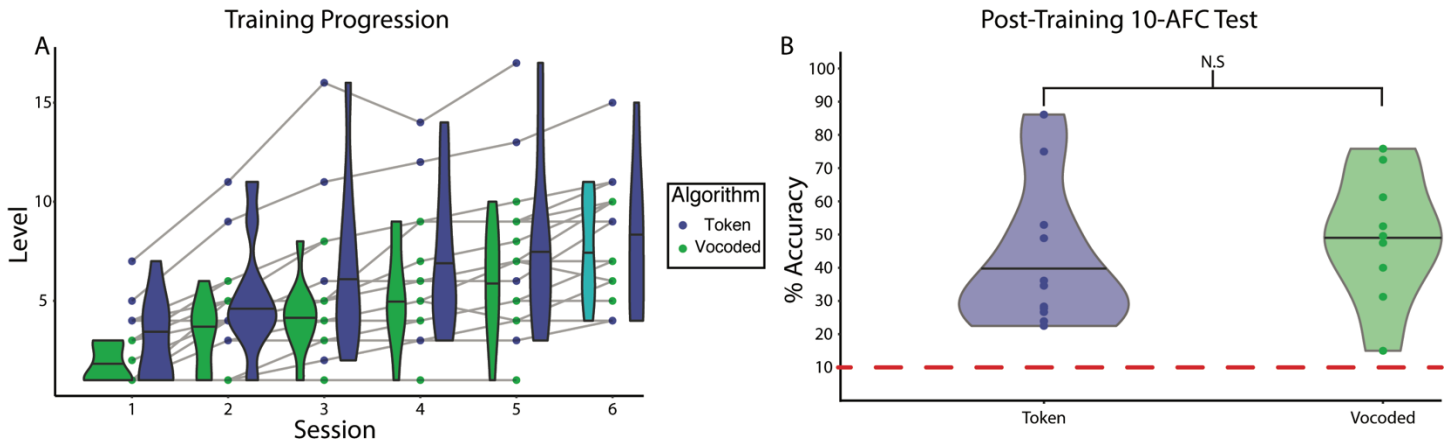
46
47 **Figure 1: VT hardware, speech-to-tactile transformation algorithms, stimuli, fMRI experimental design,**
48 **and model dissimilarity matrix.** (A) Fourteen-channel MRI-compatible VT stimulator. (B) Shows the
49 breakdown of the 30 words used in the study. The auditory scan used all the words, and subjects were trained
50 on half of the words (“trained” set). Words were further broken down by their syllable structure (9 CVCC, 3
51 CCVC, and 3 CVC words). (C) Shows the two transformations used to convert spoken words into tactile
52 stimulation patterns. The token-based approach (top) assigns each phoneme a distinct VT pattern (see
53 Methods section for more details). The vocoding approach (bottom) focuses on preserving the temporal
54 dynamics between the auditory and VT stimuli. (D) Shows the auditory (top) and VT (bottom) fMRI one-back
55 paradigms used in the study. In both paradigms, subjects focused on a central fixation cross, and pressed a
56 button in their left hand if they heard or felt the same stimulus twice in a row. (E) The auditory and VT vocoded
57 perceptual model representational dissimilarity matrix (mRDM) for the 9 CVCC trained words. The high
58 correlation ($r = 0.94$) between mRDMs provide evidence for the targeted close representational congruence
59 between auditory and VT vocoded stimuli.

30 Results

31 Behavior

32 Subjects (n=20) were trained to recognize stimuli derived from either a token-based or vocoded auditory-to-VT
33 sensory substitution algorithm (Fig. 1C). Subjects completed 6 behavioral training sessions in which they
34 performed a N-AFC task on each level (see Material and Methods). Only a single session was performed per
35 day. To progress to the next level, subjects had to achieve at least 80% accuracy on the current level. Both
36 vocoded and token-based achieved progressively higher levels in the behavioral training paradigm across
37 training sessions (Fig. 2A). The median final levels achieved were 8 and 7 for the token-based and vocoded
38 VT groups respectively. After the final post-training fMRI scan, subjects completed a 10-AFC test on the
39 trained words (Fig. 2B). All subjects performed better than chance (10%) and the median accuracies were
40 35.3% and 48.5% for the token-based and vocoded VT groups respectively. A two-sample t-test revealed no
41 significant difference in accuracy between algorithm groups ($t(18) = 0.386$, $p = 0.704$).

73

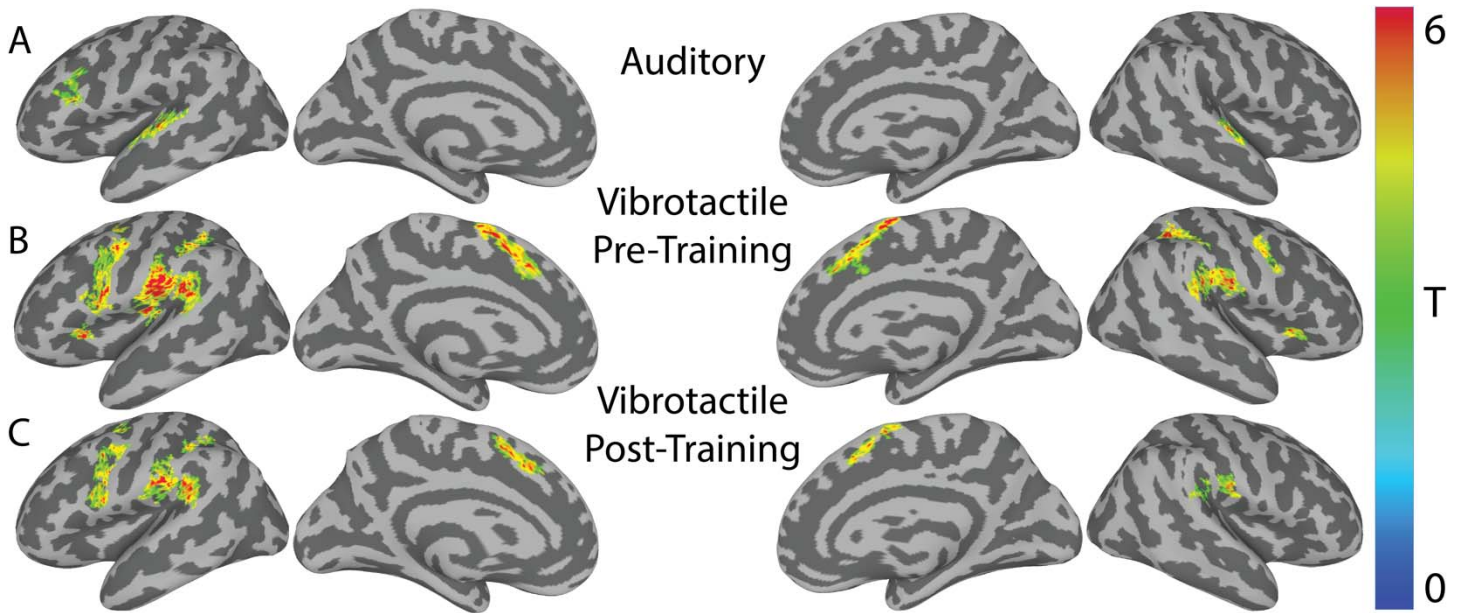


74

75 **Figure 2: Progression of learning VT stimuli as speech.** (A) Shows the leveling up of individuals on the
76 behavioral training paradigm across sessions. Shaded lines connect the same individual across sessions. Data
77 for the final session of two subjects was lost due to technical error. (B) Shows the performance of subjects by
78 algorithm group on 10-AFC task completed after the final post-training fMRI scan. A two-sample t-test reveals
79 no significant difference in performance between the groups ($t(18) = 0.386$, $p = 0.704$). Dashed red line
80 indicates chance performance. Horizontal lines in the violin plots reflect the median.

31 **Univariate fMRI Analysis**

32 Univariate analyses were conducted to examine the activation in response to the auditory and VT stimuli. In
33 the auditory scan, the contrast of “All Words>Baseline” revealed bilateral Superior Temporal Gyrus (STG)
34 activation (Table S1 and Fig. 3A). In the VT scans, unpaired two-sample t-tests revealed no significant
35 differences between the vocoded and token-based groups in either the pre-training or post-training phase.
36 Therefore, subjects were combined within training-phase to test for the cortical common response to VT
37 stimulation. The contrast “All Vibrotactile Words>Baseline” revealed several regions, including bilateral
38 supplementary motor area (SMA), precentral gyri (Table S1 and Fig. 3B-C). No significant clusters were
39 identified for the post- vs pre- training contrast. To gain a better picture of the neuronal selectivity underlying
30 these responses, we performed a series of RSA analyses.



32
33 **Figure 3: Univariate activity for "Stimuli-Baseline" in the auditory and VT scans.** (A) Shows the group-
34 level speech perception network revealed by the contrast of all auditory words > baseline. (B) Shows the pre-
35 training group-level VT perception network revealed by the contrast of all vibrotactile words > baseline. (C)
36 Same as (B), but for post-training scans. Results are rendered on a SUMA-derived standard surface. All
37 results are presented at a cluster-defining two-tailed $\alpha = 0.005$ and $p \leq 0.05$.

38

Supplementary Table 1: Univariate activity for all stimuli > baseline in the different scans

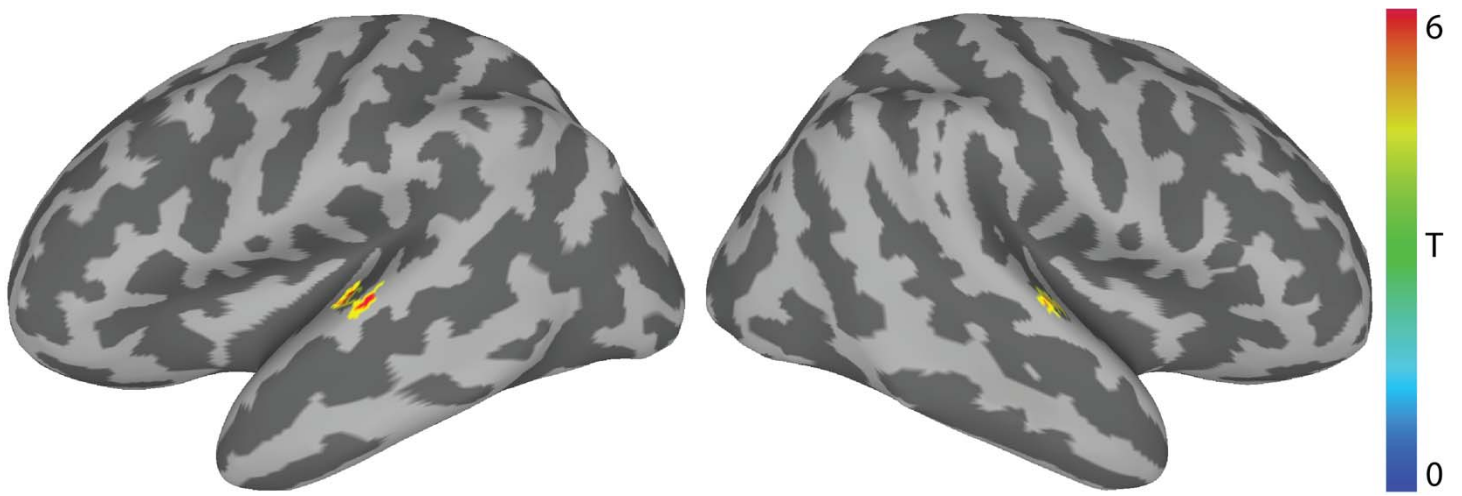
Scan	Hemi	Cluster Name (Glasser ROIs)	T _{max}	Cluster p-Value	Center of Mass Coordinates (MNI)		
					x	y	z
Auditory	RH	Parabelt Complex	6.68	0.001	57	-13	3
	LH	Parabelt Complex	6.79	0.001	-56	-19	5
		Auditory 5 Complex	7.46	0.001	-62	-36	7
Pre- Training	RH	Area PF Complex	7.57	0.001	55	-25	24
		Anterior Intraparietal Area	7.84	0.001	39	-39	42
		Supplementary and Cingulate Eye Field	8.97	0.001	8	13	52
		Premotor Eye Fields	5.75	0.001	51	2	41
		Anterior Ventral Insular Area	6.45	0.001	30	25	3
	LH	Area OP1/SII	10.78	0.001	-52	-27	23
		Rostral Area 6	8.04	0.001	-50	2	28
		Supplementary and Cingulate Eye Field	8.40	0.001	-8	9	54
		Anterior Intraparietal Area	6.59	0.001	-45	-38	42
		Anterior Ventral Insular Area	7.64	0.001	-30	25	7
		Frontal Eye Fields	6.62	0.002	-30	-3	48
Post- Training	RH	Retroinsular Cotex	4.58	0.001	53	-32	25
		Supplementary and Cingulate Eye Field	6.64	0.001	7	15	49
		Area PF Opercular	5.81	0.003	57	-16	22
		Area Posterior 24 Prime	7.17	0.019	7	2	65
	LH	Rostral Area 6	6.60	0.001	-48	2	29
		Area PF Opercular	8.99	0.001	-59	-22	25
		Area PF Complex	7.12	0.001	-50	-40	26
		Supplementary and Cingulate Eye Field	6.83	0.001	-9	14	49
		Area 6 Anterior	6.07	0.001	-29	-5	48
		Anterior Intraparietal Area	5.85	0.002	-47	-35	42
Anterior Intraparietal Area	5.71	0.002	-35	-44	40		

39

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

02 We conducted a whole-brain searchlight RSA analysis to identify regions showing selectivity for auditory
03 vocoded words. In each searchlight we constructed a neural RDM that was correlated to the auditory
04 perceptual mRDM (see Methods). The group-level t-statistic map was thresholded at a two-tailed $\alpha = .001$ and
05 the resulting clusters were corrected at two-tailed $p \leq 0.05$ (Fig. 4). This revealed left ($x = -58, y = -18, z = 5; \alpha$
06 $= 0.001; p = 0.001$) and right mid-STG ($x = 58, y = -14, z = 3; \alpha = 0.001; p = 0.016$) clusters. Of the 75 nodes in
07 the left mid-STG cluster, 8 are in left A1, 21 are in the lateral belt, 28 are in the parabelt, and 31 are in A4 as
08 defined by the Glasser Atlas. Of the 44 nodes in the right mid-STG cluster, 0 are in right A1, 12 are in the
09 lateral belt, 25 are in the parabelt, and 16 are in A4. Thus, the regions identified in this analysis are non-
10 primary auditory cortical regions that are likely selective for complex auditory spectrotemporal patterns involved
11 in speech perception (Hamilton et al., 2020).

13



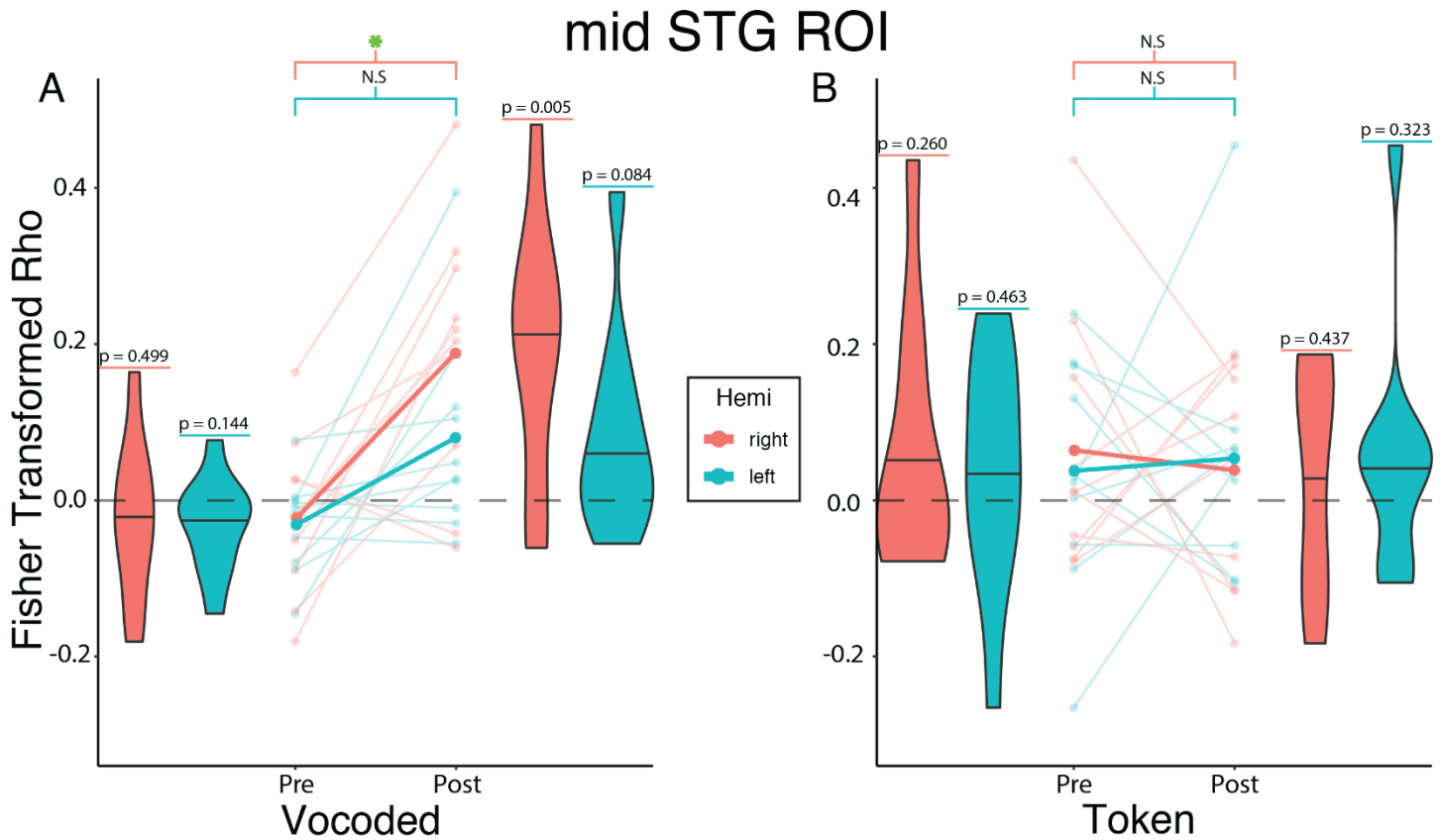
14

15 **Figure 4: Auditory Scan – Representational similarity analysis (RSA) of vocoded auditory words.** RSA
16 revealed that neural RDMs in bilateral STG regions significantly correlated with the predicted auditory
17 perceptual mRDM (Fig. 1E) ($n=18$; $\alpha = 0.001$; $p \leq .05$). The center of mass of the left STG cluster was centered
18 on MNI: -58, -18, 5. The center of mass of the right STG cluster was centered on MNI: 58, -14, 3. Colors reflect
19 across-subject t-statistics.

ROI-based analysis reveals that the right auditory word-selective region shows selectivity for VT vocoded, but not token-based words following VT speech training

Next, we conducted ROI-based RSA analyses to test the prediction that trained VT stimuli would engage the same representations as auditory words in the mid-STG. To do so, we first computed the average Fisher transformed correlation between the vibrotactile nRDMs and the auditory perceptual mRDM for the 9 trained CVCC words in the VT scans. A linear mixed-effects model was then constructed (see Methods) to test the effects of training phase, algorithm, hemisphere, as well as the interaction among them. This analysis revealed a significant interaction effect between training phase and algorithm ($\beta = 0.240$, $t(31.09) = 2.679$, $p = 0.012$; Table S2). Post-hoc tests revealed a significant ($t(31.1) = 3.380$, $p = 0.010$ Tukey-adjusted) increase between the pre- and post-training correlations in the right mSTG for the vocoded but not ($t(31.1) = -0.408$, $p = 0.977$ Tukey-adjusted) the token-based group. 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 ($t(31.1) = 1.781$, $p = 0.302$ Tukey-adjusted) or the token-based ($t(31.1) = 0.250$, $p = 0.994$ Tukey-adjusted) group. 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 ($t(9) = 2.783$, $p = 0.021$) higher correlations post-training in the right than left mSTG. In addition, there was a marginally significant ($t(9) = 2.185$, $p = 0.057$) difference when the difference between pre- and post-training correlations were compared between the right and left mid-STG. These results indicate that trained VT stimuli based on vocoded speech engaged auditory speech representations in the mid-STG and did so more strongly than token-based VT stimuli, and there was no evidence that token-based VT stimuli engaged these auditory speech representations. Furthermore, there is evidence that this effect may be stronger in the right hemisphere than the left.

The noteworthy difference in the engagement of mid-STG auditory speech representations for the vocoded but not token-based VT stimuli raised the question what other brain areas might underlie subjects' ability to learn the token-based VT stimuli as words (see Fig. 2). A possible explanation of the results is that because the token-based representation is not well matched to auditory speech representations (e.g., in its temporal dynamics), to learn the association between the two, the brain must rely on alternate strategies such as those used to learn arbitrary associations between pairs of stimuli. A key region involved in learning such associations is the hippocampus (McClelland et al., 1995; O'Reilly and Rudy, 2001). Therefore, we tested whether the hippocampus encoded token-based stimuli after training.



49
 50 **Figure 5: Vocoded but not token-based VT stimuli are represented in mid-STG auditory speech region**
 51 **following VT speech training.** Linear mixed-effects analysis revealed a significant two-way interaction
 52 between Training Phase and Algorithm ($\beta = 0.240$, $t(31.1) = 2.679$, $p = 0.012$). To investigate this interaction,
 53 we created interaction effects plots. (A) The mean Fisher-transformed Pearson correlation between neural and
 54 model RDMs estimated from the mixed-effects model for the vocoded group are represented by the opaque
 55 lines. For the VT vocoded group, post-hoc tests show a significant difference between pre- and post-training in
 56 the right ($t(31.1) = 3.380$, $p = 0.010$ Tukey-adjusted) but not the left STG ($t(31.1) = 1.781$, $p = 0.302$ Tukey-
 57 adjusted). (B) The same as (A) but for the token-based group. Post-hoc tests show no significant difference in
 58 the right ($t(31.1) = 0.408$, $p = 0.977$ Tukey-adjusted) or left STG ($t(31.1) = 0.250$, $p = 0.994$ Tukey-
 59 adjusted). Values above each violin reflect the *uncorrected* p -value from a one-sample t -test against 0. Semi-transparent
 60 lines reflect raw individual subject correlations from either the left (teal) or right (orange) STG. Horizontal lines
 61 in the violin plots reflect the median. Green asterisk marks significant ($p \leq 0.05$) differences after multiple
 62 comparisons correction.

33 **ROI-based analysis reveals that the Left Hippocampus is engaged during perception of VT token-**
34 **based, but not vocoded stimuli**

35 We therefore next tested the hypothesis that VT speech perception training led to an encoding of the VT stimuli
36 in the hippocampus. If trained VT speech stimuli were stored in a representation that reflected the associated
37 auditory speech stimuli, then we would expect neural activation pattern similarity for the VT stimuli to correlate
38 with the perceptual similarity of the auditory speech stimuli post- but not pre-training. To test this hypothesis,
39 we correlated neural activation patterns in response to VT speech stimuli in the two different encoding
40 schemes with the auditory perceptual mRDM before and after training. These correlations were then fit with a
41 linear mixed effects model. This analysis revealed a significant two-way interaction between training phase and
42 hemisphere ($\beta = 0.095$, $t(36) = 2.696$, $p = 0.011$; Fig. 6; Table S3) as well as a significant three-way interaction
43 effect between training phase, algorithm, and hemisphere ($\beta = -0.151$, $t(36) = -3.027$, $p = 0.005$; Table S3).
44 The three-way interaction suggests that the relationship between training phase and hemisphere varied
45 depending on the algorithm. Post-hoc tests revealed a significant ($t(30.7) = 3.232$, $p = 0.0148$ Tukey-adjusted)
46 training-related increase in correlations for the token-based but not vocoded ($t(30.7) = 0.785$, $p = 0.861$ Tukey
47 Adjusted) VT group in the left hemisphere. In the right hemisphere, there was a trending increase in correlation
48 for the vocoded group ($t(30.7) = 2.387$, $p = 0.101$ Tukey Adjusted) but not the token-based ($t(30.7) = .506$, $p =$
49 0.957 Tukey Adjusted) VT group.

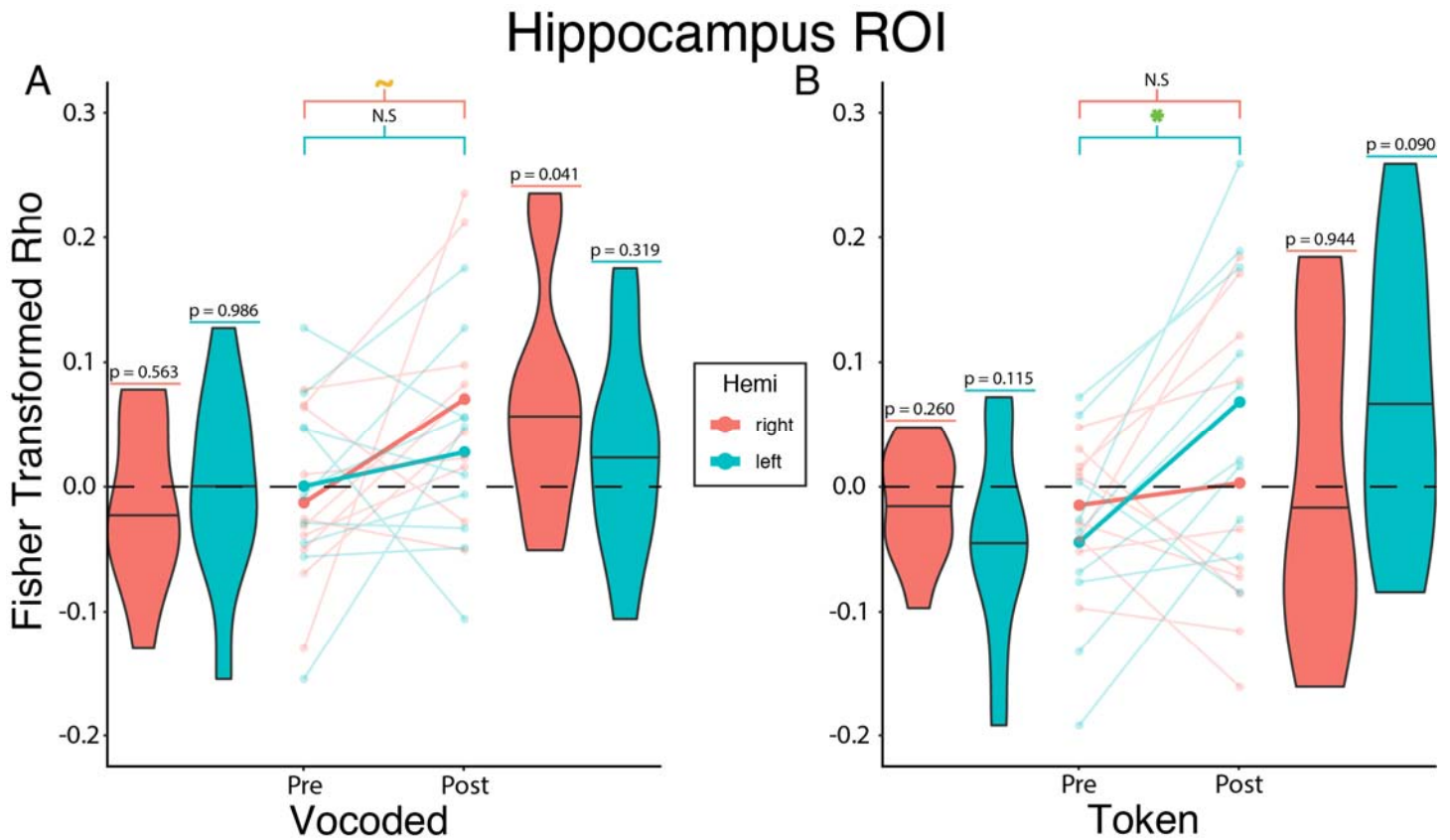
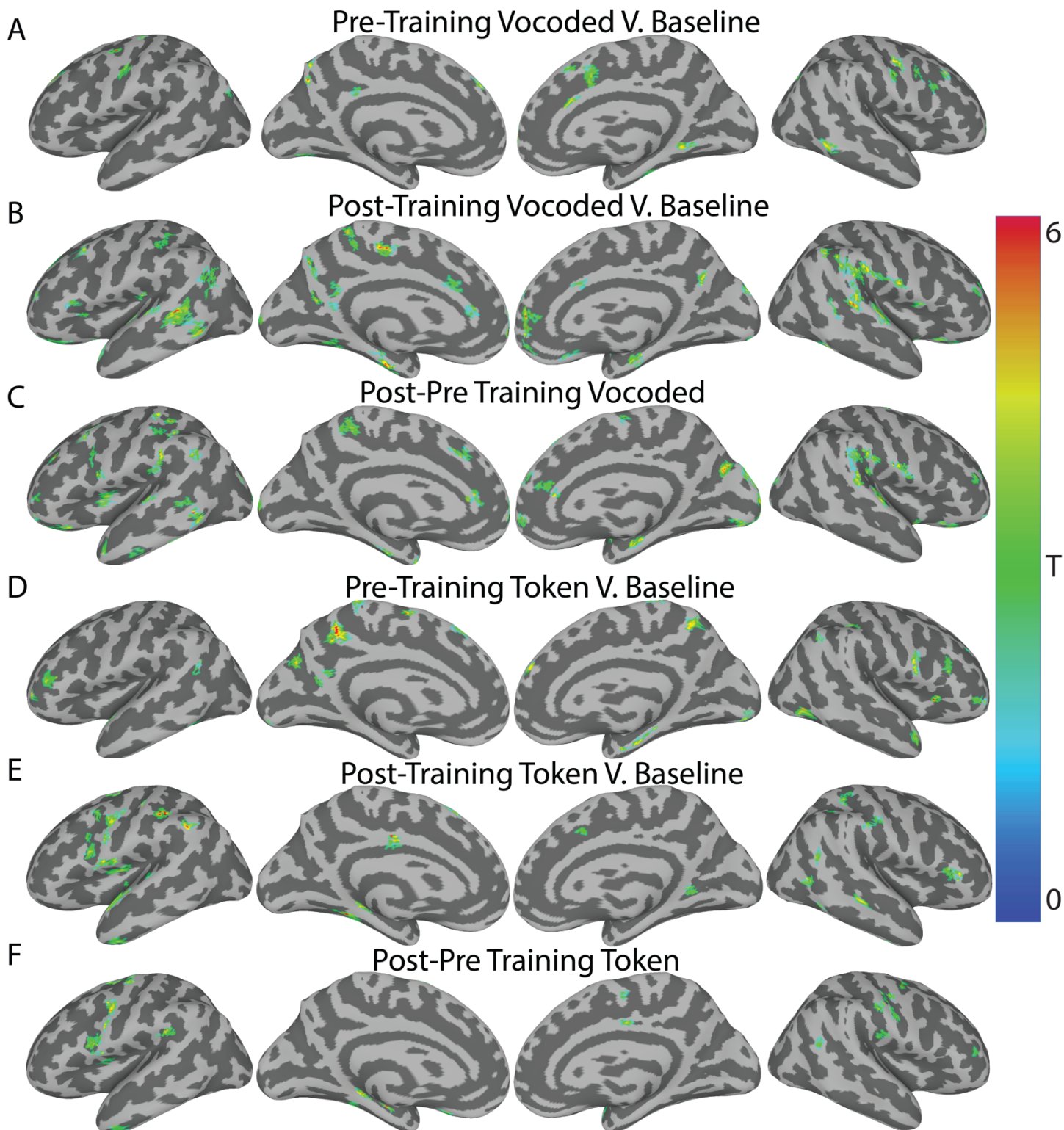


Figure 6: Token-based but not vocoded VT speech stimuli are represented in the left hippocampus following training. Linear mixed-effects analysis revealed a significant three-way interaction between Training Phase, Algorithm, and Hemisphere ($\beta = -0.151$, $t(36) = -3.027$, $p = 0.005$). To investigate 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 ($t(30.7) = 2.387$, $p = 0.101$ Tukey-adjusted) but not the left STG ($t(30.7) = 0.785$, $p = 0.861$ Tukey-adjusted). (B) The same as (A) but for the token-based group. Post-hoc tests show no significant difference in the right ($t(30.7) = 0.506$, $p = 0.957$ Tukey-adjusted), but do show a significant difference in the left STG ($t(30.7) = 3.232$, $p = 0.015$ Tukey-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 and orange tilde mark significant ($p \leq 0.05$) and trending ($p \leq 0.1$) differences, respectively, after multiple comparisons correction.



36
37 **Supplementary Figure 1: Whole-brain Searchlight RSA of Trained CVCC VT Stimuli with the Auditory**
38 **Perceptual mRDM.** (A-C) Whole-brain RSA results for the VT vocoded stimuli. (A) Pre-training scan: Fischer
39 transformed correlation against 0. (B) Post-training scan: Fischer transformed correlation against 0. (C) Post
40 minus Pre-Training change in the Fischer transformed correlations. (D-F) same as (A-C) but for the token-
41 based stimuli. All results are at an uncorrected two-tailed voxel-wise threshold $\alpha = 0.05$ with a extent threshold
42 of 50mm^2 . Colors reflect across-subject t-statistics.

03

Supplementary Table 2: Linear Mixed-Effects Model Summary for the mid-STG ROIs

Summary of Linear Mixed Effects Model: mid-STG ROIs					
Fixed Effects					
Predictors	β Estimate	Confidence Interval	T-Statistic	DOF	p-value
Intercept	0.063	-0.02 – 0.14	1.566	35.34	0.126
Training Phase	-0.026	-0.15 – 0.10	-0.408	31.09	0.686
Algorithm	-0.086	-0.20 – 0.03	-1.507	35.34	0.141
Hemisphere	-0.026	-0.12 – 0.06	-0.578	36	0.567
Training Phase:Algorithm	0.240	0.06 – 0.41	2.679	31.09	0.012
Algorithm:Hemisphere	0.018	-0.11 – 0.14	0.277	36	0.783
Training Phase:Hemisphere	0.042	-0.09 – 0.17	0.643	36	0.524
Training Phase: Algorithm:Hemisphere	-0.143	-0.32 – 0.04	-1.559	36	0.128
Random Effects					
Groups	Effect Name	σ (std. deviation)	Variance	Correlation Structure	
Subj	Intercept	0.075	0.006	N/A	-0.60
	Training Phase	0.138	0.019	-0.60	N/A
Residual		0.102	0.010		

04

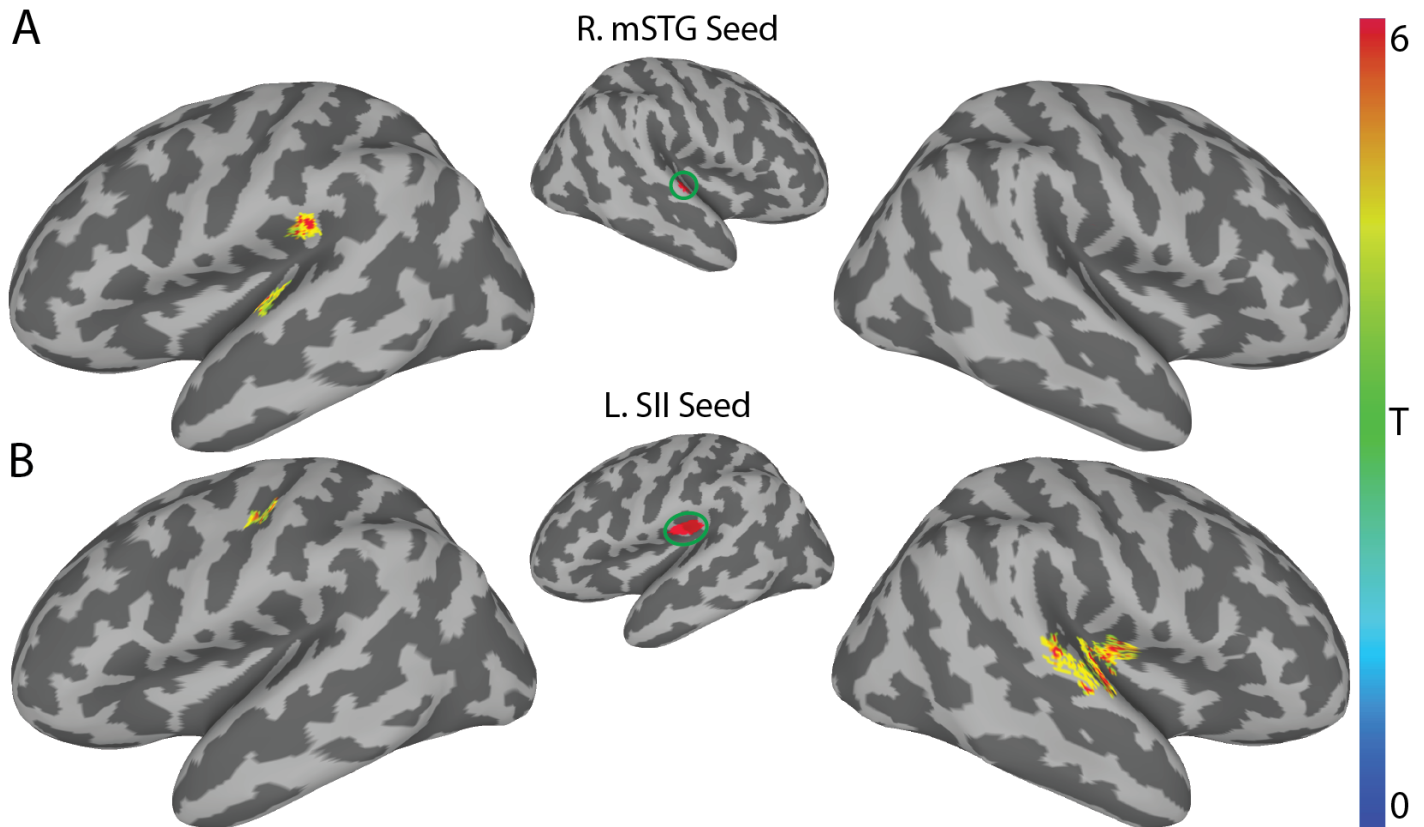
05 **Supplementary Table 3: Linear Mixed-Effects Model Summary for the Hippocampus ROIs**

Summary of Linear Mixed Effects Model: Hippocampus ROIs					
Fixed Effects					
Predictors	β Estimate	Confidence Interval	T-Statistic	DOF	p-value
Intercept	-0.015	-0.06 – 0.03	-0.668	34.94	0.508
Training Phase	0.018	-0.05 – 0.09	0.506	30.67	0.616
Algorithm	0.002	-0.06 – 0.07	0.066	34.94	0.948
Hemisphere	-0.029	-0.08 – 0.02	-1.185	36	0.244
Training Phase:Algorithm	0.066	-0.04 – 0.17	1.330	30.67	0.193
Algorithm:Hemisphere	0.043	-0.03 – 0.11	1.211	36	0.234
Training Phase:Hemisphere	0.095	0.02 – 0.17	2.696	36	0.011
Training Phase: Algorithm:Hemisphere	-0.151	-0.25 – -0.05	-3.027	36	0.005
Random Effects					
Groups	Effect Name	σ (std. deviation)	Variance	Correlation Structure	
Subj	Intercept	0.042	0.002	N/A	0.03
	Training Phase	0.077	0.006	0.03	N/A
Residual		0.056	0.003		

06

07 **Training with Vocoded VT Speech Stimuli Increases Functional Connectivity Between Somatosensory** 08 **and Auditory Regions**

09 Previous studies showed that learning is accompanied by increased functional connectivity between cortical
10 areas (Lewis et al., 2009; Siuda-Krzywicka et al., 2016; Uner et al., 2013). Therefore, we tested the
11 hypothesis that training on the vocoded VT word stimuli was associated with increased functional connectivity
12 of somatosensory regions and the auditory word-selective right mid-STG ROI (Fig. 4). To do so, we computed
13 the training-related changes in the right mid-STG seed-to-voxel functional connectivity in the vocoded group
14 (Fig. 7A, Table S4). This revealed two clusters, one in the left STG ($x = -50, y = -19, z = 7; \alpha = 0.005; p =$
15 0.044) and another in the left secondary somatosensory (SII) ($x = -55, y = -28, z = 21; \alpha = 0.005; p = 0.026$).
16 Furthermore, reasoning that VT stimulation on the right arm would engage the left SII region, we performed an
17 additional seed-to-voxel analysis using the left SII seed defined by the Glasser atlas (Glasser et al., 2016). This
18 complementary analysis revealed two clusters, one in the right insula and Heschl's Gyrus ($x = 40, y = -17, z =$
19 $11; \alpha = 0.005; p = 0.001$) and another in the right STG ($x = 63, y = -22, z = 7; \alpha = 0.005; p = 0.001$). The left SII
20 also showed an increase in connectivity to the left central sulcus ($x = -40, y = -19, z = 42; \alpha = 0.005; p =$
21 0.001). Using the left mid-STG region as a seed revealed significantly increased connectivity with the right
22 STG while using the right SII revealed significant training-related changes confined to bilateral SII. (Fig. S2,
23 Table S4). Similar seed-to-voxel analyses also using the left hippocampus or the bilateral mid-STG ROIs as
24 seeds revealed no significant training-related differences in the token-based group. This pattern of training-
25 related functional connectivity between somatosensory and auditory areas for VT vocoded but not token based
26 stimuli was also found when calculating ROI-to-ROI functional connectivity (Fig. S3). These results support a
27 model in which vocoded VT speech training leads to increased functional connectivity between somatosensory
28 areas and auditory speech areas.



30 **Figure 7: Training related differences in seed-to-voxel functional connectivity for vocoded VT stimuli.**

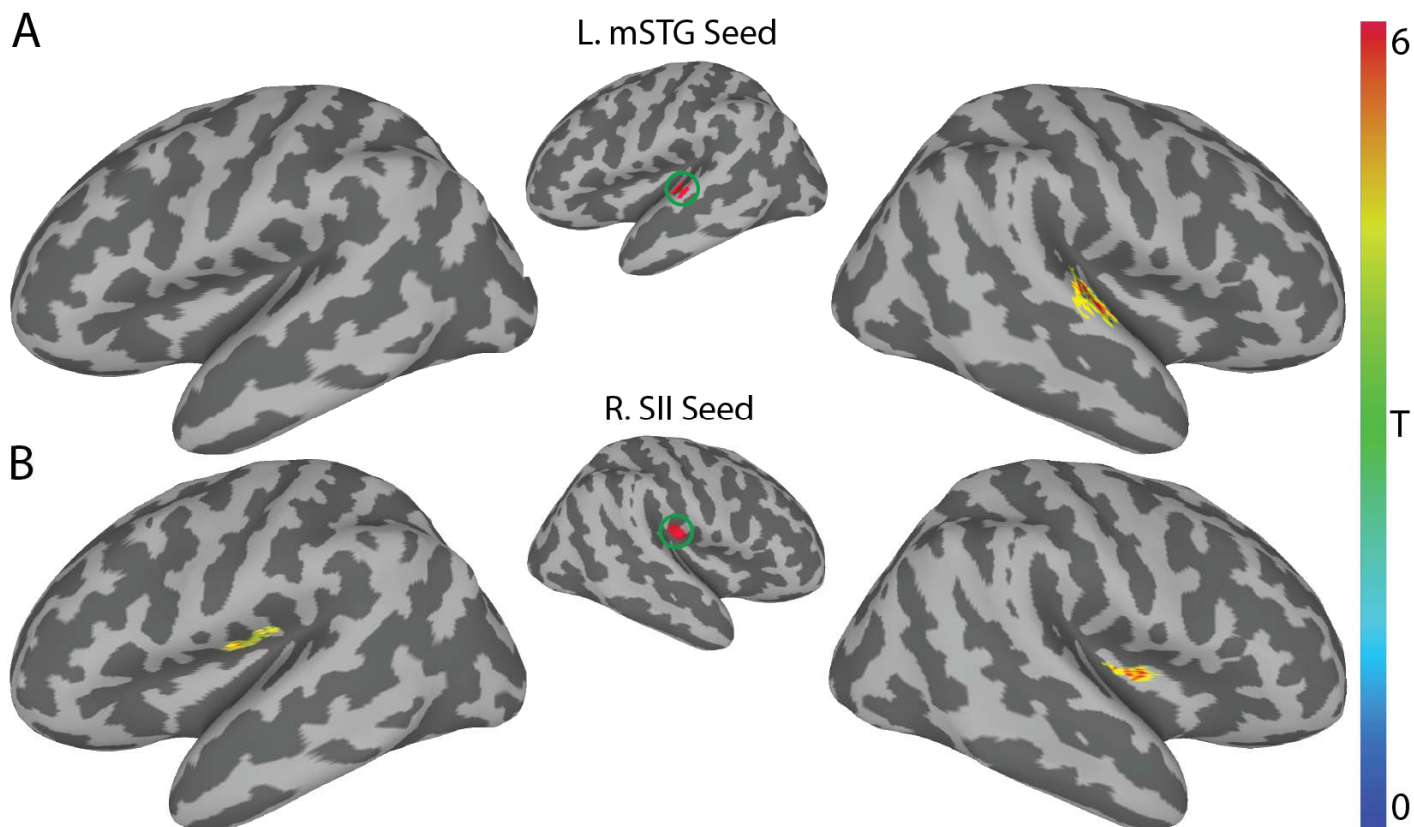
31 (A) Using the right mid-STG ROI (Fig. 4) as a seed revealed two significant clusters of increased functional
32 connectivity after training in the left STG (MNI: -50, -19, 7) and in the left supramarginal gyrus (MNI: -55, -28,
33 21). (B) Using the left SII seed derived from the Glasser atlas revealed a significant cluster in the left central
34 sulcus (MNI: -40, -19, 42). It also identified two significant clusters in the right hemisphere. The first
35 encompassed right insula and Heschl's gyrus (MNI: 40, -17, 11). The other is on the right STG (MNI: 63, -22,
36 7). All results shown are corrected at two-tailed voxel-wise $\alpha = 0.005$ and cluster- $p \leq 0.05$. Colors reflect
37 cross-subject t-statistics.
38
39

40

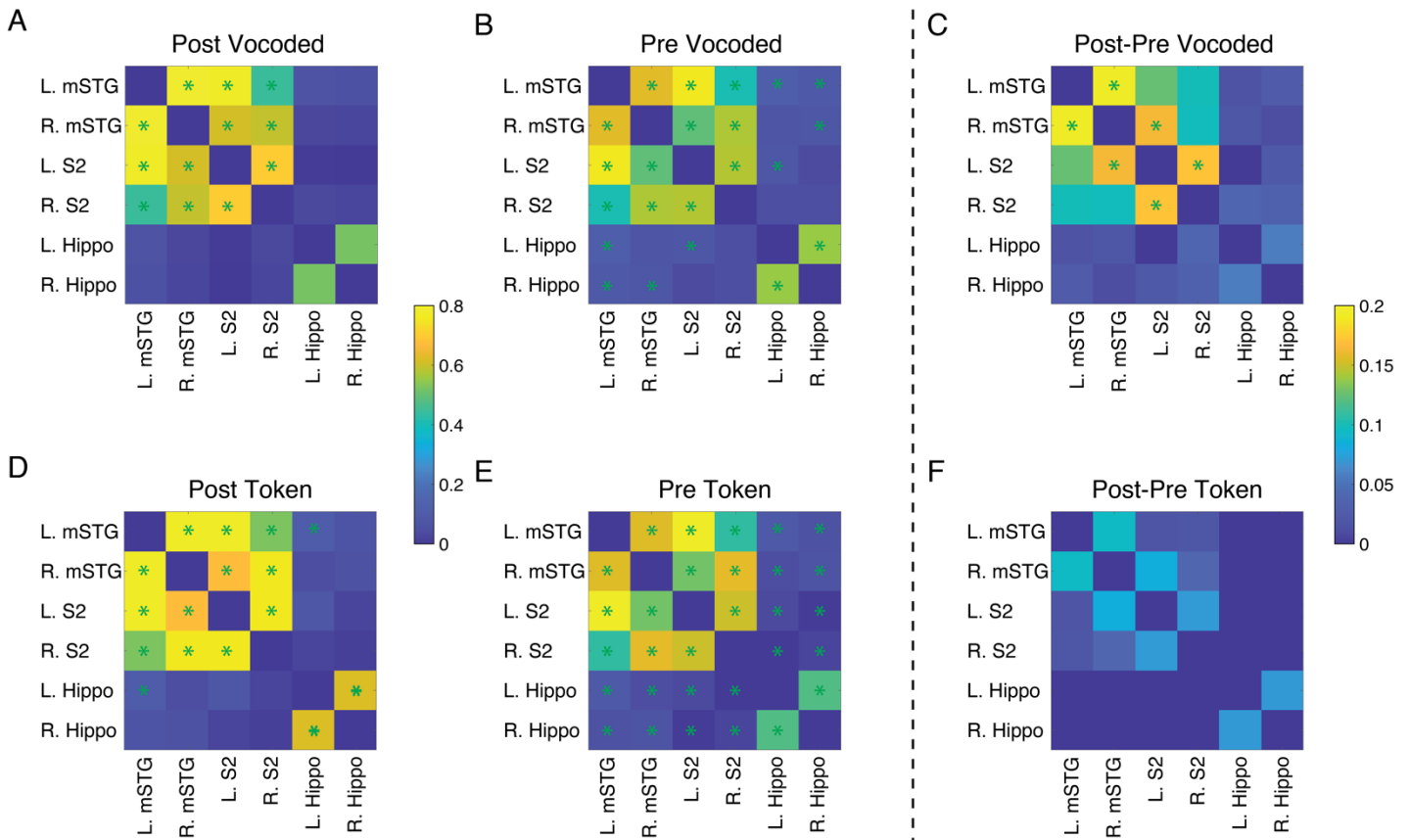
Supplementary Table 4: Training-related changes in functional connectivity in the vocoded group.

Seed ROI	Hemi	Cluster Location (Glasser ROIs)	T_{\max}	Cluster p-Value	Center of Mass Coordinates (MNI)		
					x	y	z
IS2	RH	Insular Granular Complex	8.04	0.001	40	-17	11
		Auditory 5 Complex	8.44	0.001	63	-22	7
	LH	Primary Motor Cortex	7.73	0.012	-40	-19	42
ISTG	RH	Lateral Belt Complex	6.54	0.001	53	-18	6
rS2	RH	Posterior Insular Area 2	7.41	0.017	37	-8	6
	LH	Area OP2-3/VS	5.73	0.026	-42	-16	20
rSTG	LH	Area PF _{cm}	8.20	0.026	-55	-28	21
		Lateral Belt Complex	6.75	0.044	-50	-19	7

41



42
43 **Supplementary Figure 2: Training-related differences in seed-to-voxel functional connectivity for**
44 **vocoded VT stimuli using the left STG and right SII seeds.** (A) Using the left mid-STG ROI (Fig. 4) as a
45 seed revealed one significant cluster of increased functional connectivity after training in the right mid-STG
46 (MNI: 55, -16, 3). (B) Using the right SII seed derived from the Glasser atlas revealed a significant cluster in
47 the left opercular region (MNI: -42, -15, 20) and right posterior Insula (MNI: 37, -3, 7). All results shown are
48 corrected at two-tailed voxel-wise $\alpha = 0.005$ and cluster- $p \leq 0.05$. Colors reflect across-subject t-statistics.



49
50 **Supplementary Figure 3: ROI-to-ROI based functional connectivity reveals significantly increased**
51 **coupling between the auditory and somatosensory system after training on VT vocoded stimuli.** (A-B)
52 Shows the ROI-to-ROI functional connectivity for the VT vocoded-based group during post (A) and pre (B)
53 training scans. (D-E) Same as (A-B) but for the VT token-based group. Color bar reflects the Fischer-
54 transformed Pearson correlation between ROIs. A paired t-test was performed to compare changes in
55 functional connectivity relative to baseline. Green asterisks mark $p \leq 0.05$ FDR corrected. (C, F) Shows the
56 post-pre training correlations for the VT vocoded and token-based groups respectively. Color bar reflects the
57 Post-Pre training difference between ROIs. A paired t-test was performed to compare changes in functional
58 connectivity post-pre training. Green asterisks mark $p \leq 0.05$ FDR corrected.

59 Discussion

60 Metamodal theories of brain organization (Heimler et al., 2015; Pascual-Leone and Hamilton, 2001) propose
61 that cortical areas are best described by their task-specific sensory modality-invariant function. However,
62 mixed evidence for metamodal brain organization in neurotypical individuals (Amedi et al., 2007; Bola et al.,
63 2017; Ptito et al., 2005; Sadato et al., 1996; Siuda-Krzywicka et al., 2016) has raised the question of if and
64 under what conditions metamodal engagement occurs. We argue, based on theoretical considerations, that
65 testing the metamodal hypothesis requires not just a consideration of high-level tasks (Marr's (Marr, 1982) top
66 level of "computational theory") but also and critically their algorithmic implementation (Marr's second level). In
67 the current study, we investigated this hypothesis by training subjects on the same task (recognition of
68 vibrotactile stimuli derived from auditory words) using one of two different auditory-to-VT sensory substitution
69 algorithms. One algorithm (vocoded) preserved the temporal modulations of auditory speech while the other
70 algorithm (token) attempted to establish an abstract congruence between VT patterns and the phonetic
71 features found in speech. First, using whole-brain searchlight RSA we identified auditory perceptual speech
72 representations whose locations along the superior temporal gyrus are compatible with models of the auditory
73 ventral speech recognition stream (DeWitt and Rauschecker, 2012; Hickok and Poeppel, 2007; Rauschecker
74 and Scott, 2009). Notably, this speech selectivity was found bilaterally, in agreement with other models of
75 speech processing in the brain (Hickok and Poeppel, 2007). We then showed that, before training, neither the
76 vocoded nor the token-based VT stimuli selectively engaged these auditory speech areas, as expected. Next,
77 over the course of six behavioral sessions, we trained two groups of subjects to recognize the VT-encoded
78 word stimuli, with each group trained on a different encoding scheme. Both groups of subjects achieved
79 comparable levels of proficiency, eliminating performance differences as a reason for the different training
80 effects at the neural level. Crucially, RSA revealed that after training, only the vocoded but not the token-based
81 VT stimuli engaged an auditory-speech selective region in the mid-STG (Hamilton et al., 2020). In addition,
82 both encoding schemes (to different degrees) appeared to engage hippocampal areas previously implicated in
83 paired-associate learning. Finally, we found evidence that metamodal engagement of the mid-STG by vocoded
84 VT stimuli was associated with a training-related increase in functional coupling between the mid-STG and
85 secondary somatosensory areas. Evidence of training-related increases in functional coupling was not found
86 for token-based stimuli.

87 In this study, we show that adequately capturing (and eventually harnessing) the metamodal potential of cortex
88 requires not only the right task and sensory modalities but also an understanding of the information
89 representation in these regions. Prior work has primarily investigated metamodal engagement in congenitally
90 sensory-deprived individuals (Arno et al., 2001; Bola et al., 2017; Lomber et al., 2010; Ptito et al., 2005; Reich
91 et al., 2011; Sadato et al., 1996). In such cortical areas, given the right task-relevant connectivity, bottom-up
92 input from another sensory modality can conceivably drive the *de novo* learning of task-relevant
93 representations even for encoding schemes very different from those in neurotypical individuals (Striem-Amit et
94 al., 2012). However, in neurotypical adults, existing representations in traditionally unisensory areas reflect the
95 task-relevant features of the typical sensory input (Lewicki, 2002; Simoncelli and Olshausen, 2001). Therefore,

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

40 The ability to map between representational formats in different sensory hierarchies likely depends on both
41 anatomical and functional convergence. Anatomical tracer (Mothe et al., 2006a; Schroeder et al., 2003; Smiley
42 et al., 2007) and studies in non-human primates (Kayser et al., 2009; Schroeder et al., 2001) as well as
43 neuroimaging studies in humans (Fuxe et al., 2002; Ro et al., 2013) have established convergence points
44 between somatosensory and auditory cortices including belt and parabelt areas. Given this connectivity, prior
45 computational studies have shown that the mapping between different representational formats can be learnt
46 through simple biologically plausible learning rules (Davison and Frégnac, 2006; Pouget and Sejnowski, 1997;
47 Pouget and Snyder, 2000). Still, while it is simple to learn the mapping between static features, it is non-trivial
48 to match the temporal dynamics between functional hierarchies. For example, Davison and Frégnac (2006)
49 computationally demonstrated the importance of temporally coherent activity between representational formats
50 when learning the mapping between cross-modal temporal sequences using spike-timing-dependent plasticity
51 mechanisms. In the auditory cortex specifically, studies (Moore and Woolley, 2019; Overath et al., 2015) have
52 shown that auditory stimuli that do not preserve the same temporal modulations found in conspecific
53 communication signals (e.g., speech, birdsong, etc.) sub-optimally drive higher-order auditory cortex and
54 preclude learning. Recent human intracranial EEG studies (Hamilton et al., 2018; Hullett et al., 2016) have
55 demonstrated that middle superior temporal cortex is characterized by very short temporal receptive fields
56 necessitating relatively rapid changes in the somatosensory signal. Accordingly, we find, in the current study,
57 that only vocoded stimuli that preserve these fast temporal dynamics are able to drive auditory perceptual
58 speech representations in the mid-STG. Conversely, the different dynamics (see Materials and Methods) of
59 token-based VT stimuli relative to auditory speech may explain why these stimuli were unable to interface with
60 mid-STG speech representations.

61 Intriguingly, we find stronger evidence of metamodal engagement by VT vocoded stimuli in the right rather than
62 left mid-STG. A significant body of work (Albouy et al., 2020; Boemio et al., 2005; Flinker et al., 2019; Giraud
63 and Poeppel, 2012; Obleser et al., 2008; Zatorre and Belin, 2001) suggests that the left and right STG are
64 differentially sensitive to spectrotemporal content of auditory stimuli. Specifically, it has been proposed (Flinker
65 et al., 2019) that the left STG tends to sample auditory information on fast and slow timescales while the right
66 preferentially does the latter. In the current study, our VT vocoded stimuli preserve the coarse temporal
67 dynamics of auditory speech, but due to hardware limitations have a lower temporal resolution than the
68 auditory source signal. In addition, the temporal resolution of vibrotactile perception is lower than that of
69 auditory processing, with receptors in the skin acting as an additional low pass filter (Bensmaïa and Hollins,
70 2005). Thus, the observed metamodal coupling with the right rather than the left STG provides intriguing
71 support for the asymmetric spectrotemporal modulation theory of hemispheric processing (Flinker et al., 2019).

32 Given that subjects were able to learn token-based and vocoded VT stimuli as words with roughly equal
33 proficiency, how do token-based stimuli engage spoken word representations? Due to the slower temporal
34 dynamics of token-based stimuli, we initially hypothesized that these stimuli may map onto higher order speech
35 representations in areas such as the superior temporal sulcus (STS) or anterior STG that integrate temporal
36 information on longer timescales (Hullett et al., 2016; Overath et al., 2015). However, we did not find evidence
37 for this in the current study. An anatomical tracer study by De La Mothe (Mothe et al., 2006b) showed strong
38 evidence of connectivity between somatosensory cortex and mid and posterior but not anterior superior
39 temporal areas. Thus, a homologous lack of connectivity between somatosensory and anterior superior
40 temporal areas in humans may explain why we observed no engagement of those areas after training.
41 However, we did find evidence that token-based stimuli engage neural representations in the left hippocampus.
42 This result fits with previous proposals that learned associations can be retrieved using paired-associate recall
43 circuits in the medial temporal lobe (Miyashita, 2019). A more thorough understanding of this process through
44 future studies will shed additional insight into which pathways and mechanisms are leveraged to learn different
45 types of associations.

46 Previous studies have suggested that metamodal engagement is a result of top-down processes such as
47 mental imagery rather than bottom-up processes (Lacey et al., 2009). However, given that in our study,
48 subjects in both algorithm groups were equally proficient at recognizing VT stimuli as words, mental-imagery
49 accounts (Borst and Gelder, 2016; Li et al., 2020; Oh et al., 2013; Tian et al., 2018) in this case would predict
50 that both groups should engage auditory perceptual representations in the mid-STG. Yet, we found no
51 evidence that the token-based VT stimuli engaged this area after training in the same way as auditory speech
52 (see also (Siuda-Krzywicka et al., 2016; Striem-Amit et al., 2012)). Thus, it is unlikely that metamodal
53 engagement of the mid-STG by vocoded stimuli is driven by top-down mechanisms.

54 Most prior studies (Amedi et al., 2002, 2007; Reich et al., 2011; Siuda-Krzywicka et al., 2016; Striem-Amit et
55 al., 2012, 2015; Vetter et al., 2020) have demonstrated metamodal engagement in visual cortex. Our study
56 extends these findings to show that metamodal engagement is possible in auditory cortex as well. To our
57 knowledge, metamodal engagement of auditory cortex has been limited to posterior auditory association cortex
58 (pSTS) and has only been found in congenitally deaf but not hearing individuals (Benetti et al., 2017, 2020;
59 Bola et al., 2017; Twomey et al., 2017). Furthermore, these studies did not find evidence of metamodal
60 engagement in neurotypical individuals. In contrast, our study provides novel evidence for metamodal
61 engagement of intermediate auditory areas. This is particularly noteworthy given the sparse evidence for
62 metamodal engagement of intermediate sensory areas (Heimler and Amedi, 2020). The dearth of evidence is
63 likely due to a lack of knowledge of the structure of stimulus representations in these regions, which our work
64 suggests is critical for successful metamodal engagement.

65 In summary, our results provide further evidence for the metamodal theory and advance it by demonstrating
66 the importance of matching representational formats between functional hierarchies for achieving metamodal
67 engagement. In particular, our results suggest that matching the temporal dynamics of representations is an
68 important consideration when considering the feasibility of learning the appropriate mapping. This extends

39 theories (Heimler et al., 2015; Pascual-Leone and Hamilton, 2001) that emphasize a cross-modal congruences
70 at the Computational/Task level by additionally highlighting the need for an algorithmic congruence. Taking this
71 need for algorithmic congruence into account may provide insight into how the brain learns to map between
72 various levels of different functional hierarchies like sub-lexical and lexical orthography and phonology (Share,
73 1999). Furthermore, it suggests that therapeutic sensory substitution devices might benefit from different
74 designs for patients with acquired rather than congenital sensory deprivation. For the former group, careful
75 consideration should be given to the type of sensory substitution device that best interfaces with spared
76 sensory representations. The ability to “piggyback” onto an existing processing hierarchy (e.g., auditory speech
77 recognition) may facilitate the rapid learning of novel stimuli presented through a spared sensory modality
78 (e.g., VT). Here we demonstrate that an algorithm (vocoding) that improves this interfacing is able to more
79 efficiently convey the same information than an algorithm (token) that does not. Future work should explore
30 whether this observed integration into existing processing streams leads to improved generalization and
31 transfer of learning.
32

33 **Acknowledgments**

34 A portion of the funding for this research was provided by Facebook. We would also like to acknowledge Ali
35 Israr, Frances Lau, Keith Klumb, Robert Turcott, and Freddy Abnoui for their involvement in the early stages
36 of the project, including the design and evaluation of the token-based encoding scheme. Finally, we would like
37 to acknowledge Dr. Ella Striem-Amit for helpful feedback on earlier versions of this manuscript.

References

- Albouy, P., Benjamin, L., Morillon, B., and Zatorre, R.J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* 367, 1043–1047.
- Alcorn, S. (1945). Development of the Tadoma Method for the Deaf-Blind. *Except Children* 11, 117–119.
- Amedi, A., Jacobson, G., Hendler, T., Malach, R., and Zohary, E. (2002). Convergence of Visual and Tactile Shape Processing in the Human Lateral Occipital Complex. *Cereb Cortex* 12, 1202–1212.
- Amedi, A., Stern, W.M., Camprodon, J.A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P., and Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nat Neurosci* 10, 687–689.
- Arno, P., Vanlierde, A., Streel, E., Wanet-Defalque, M. -C., Sanabria-Bohorquez, S., and Veraart, C. (2001). Auditory substitution of vision: pattern recognition by the blind. *Appl Cognitive Psych* 15, 509–519.
- Auer, E.T., and Bernstein, L.E. (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., and Kercel, S.W. (2003). Sensory substitution and the human–machine interface. *Trends Cogn Sci* 7, 541–546.
- Barsalou, L.W. (2017). What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia* 105, 18–38.
- Benetti, S., Ackeren, M.J. van, Rabini, G., Zonca, J., Foa, V., Baruffaldi, F., Rezk, M., Pavani, F., Rossion, B., and Collignon, O. (2017). Functional selectivity for face processing in the temporal voice area of early deaf individuals. *Proc National Acad Sci* 114, E6437–E6446.
- Benetti, S., Zonca, J., Ferrari, A., Rezk, M., Rabini, G., and Collignon, O. (2020). Visual motion processing recruits regions selective for auditory motion in early deaf individuals. *Biorxiv* 2020.11.27.401489.
- Bensmaïa, S., and Hollins, M. (2005). Pacinian representations of fine surface texture. *Percept Psychophys* 67, 842–854.
- Bernstein, L.E., Demorest, M.E., Coulter, D.C., and O’Connell, M.P. (1991). Lipreading sentences with vibrotactile vocoders: performance of normal-hearing and hearing-impaired subjects. *J Acoust Soc Am* 90, 2971–2984.
- Bi, Y., Wang, X., and Caramazza, A. (2016). Object Domain and Modality in the Ventral Visual Pathway. *Trends Cogn Sci* 20, 282–290.
- Boemio, A., Fromm, S., Braun, A., and Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat Neurosci* 8, 389–395.

- 17 Bola, Ł., Zimmermann, M., Mostowski, P., Jednoróg, K., Marchewka, A., Rutkowski, P., and Szwed, M. (2017). Task-
18 specific reorganization of the auditory cortex in deaf humans. *Proc National Acad Sci* 114, E600–E609.
- 19 Bola, Ł., Yang, H., Caramazza, A., and Bi, Y. (2020). Preference for animate domain sounds in the fusiform gyrus of blind
20 individuals is modulated by shape-action mapping. *Biorxiv* 2020.06.20.162917.
- 21 Borst, A.W. de, and Gelder, B. de (2016). fMRI-based Multivariate Pattern Analyses Reveal Imagery Modality and
22 Imagery Content Specific Representations in Primary Somatosensory, Motor and Auditory Cortices. *Cereb Cortex New*
23 *York N Y* 1991.
- 24 Brooks, P.L., and Frost, B.J. (1983). Evaluation of a tactile vocoder for work recognition. *J Acoust Soc Am* 74, 34–39.
- 25 Chevillet, M.A., Jiang, X., Rauschecker, J.P., and Riesenhuber, M. (2013). Automatic Phoneme Category Selectivity in the
26 Dorsal Auditory Stream. *The Journal of Neuroscience* 33, 5208–5215.
- 27 Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English* (Harper and Row).
- 28 Cieśla, K., Wolak, T., Lorens, A., Heimler, B., Skarżyński, H., and Amedi, A. (2019). Immediate improvement of speech-in-
29 noise perception through multisensory stimulation via an auditory to tactile sensory substitution. *Restor Neurol Neuros* 37,
30 155–166.
- 31 Davison, A.P., and Frégnac, Y. (2006). Learning Cross-Modal Spatial Transformations through Spike Timing-Dependent
32 Plasticity. *J Neurosci* 26, 5604–5615.
- 33 DeWitt, I., and Rauschecker, J.P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of*
34 *the National Academy of Sciences* 109, E505–E514.
- 35 Fairhall, S.L., Porter, K.B., Bellucci, C., Mazzetti, M., Cipolli, C., and Gobbini, M.I. (2017). Plastic reorganization of neural
36 systems for perception of others in the congenitally blind. *Neuroimage* 158, 126–135.
- 37 Fischl, B., Sereno, M.I., Tootell, R.B.H., and Dale, A.M. (1999). High-resolution intersubject averaging and a coordinate
38 system for the cortical surface. *Hum Brain Mapp* 8, 272–284.
- 39 Flinker, A., Doyle, W.K., Mehta, A.D., Devinsky, O., and Poeppel, D. (2019). Spectrotemporal modulation provides a
40 unifying framework for auditory cortical asymmetries. *Nat Hum Behav* 3, 393–405.
- 41 Foxe, J.J., Wylie, G.R., Martinez, A., Schroeder, C.E., Javitt, D.C., Guilfoyle, D., Ritter, W., and Murray, M.M. (2002).
42 Auditory-Somatosensory Multisensory Processing in Auditory Association Cortex: An fMRI Study. *J Neurophysiol* 88,
43 540–543.
- 44 Gault, R.H. (1924). Progress in experiments on tactual interpretation of oral speech. *J Abnorm Psychology Soc*
45 *Psychology* 19, 155–159.

- 46 Gault, R.H. (1926). Touch as a Substitute for Hearing in the Interpretation and Control of Speech. Archives
47 Otolaryngology - Head Neck Surg 3, 121–135.
- 48 Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and
49 operations. Nat Neurosci 15, 511–517.
- 50 Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J.,
51 Beckmann, C.F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178.
- 52 Hagoort, P. (2020). The meaning-making mechanism(s) behind the eyes and between the ears. Philosophical
53 Transactions Royal Soc B 375, 20190301.
- 54 Hamilton, L.S., Edwards, E., and Chang, E.F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the
55 Human Superior Temporal Gyrus. Curr Biol 28, 1860-1871.e4.
- 56 Hamilton, L.S., Oganian, Y., and Chang, E.F. (2020). Topography of speech-related acoustic and phonological feature
57 encoding throughout the human core and parabelt auditory cortex. Biorxiv 2020.06.08.121624.
- 58 Hannagan, T., Amedi, A., Cohen, L., Dehaene-Lambertz, G., and Dehaene, S. (2015). Origins of the specialization for
59 letters and numbers in ventral occipitotemporal cortex. Trends Cogn Sci 19, 374–382.
- 60 Heimler, B., and Amedi, A. (2020). Are critical periods reversible in the adult brain? Insights on cortical specializations
61 based on sensory deprivation studies. Neurosci Biobehav Rev 116, 494–507.
- 62 Heimler, B., Striem-Amit, E., and Amedi, A. (2015). Origins of task-specific sensory-independent organization in the visual
63 and auditory brain: neuroscience evidence, open questions and clinical implications. Curr Opin Neurobiol 35, 169–177.
- 64 Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience 8,
65 nrn2113.
- 66 Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., and Chang, E.F. (2016). Human Superior Temporal Gyrus
67 Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. The Journal of Neuroscience 36, 2014–
68 2026.
- 69 Iverson, P., Bernstein, L.E., and Jr, E.T.A. (1998). Modeling the interaction of phonemic intelligibility and lexical structure
70 in audiovisual word recognition. Speech Commun 26, 45–63.
- 71 Kayser, C., Petkov, C.I., and Logothetis, N.K. (2009). Multisensory interactions in primate auditory cortex: fMRI and
72 electrophysiology. Hearing Res 258, 80–88.
- 73 Kell, A., Yamins, D., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A Task-Optimized Neural
74 Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy.
75 Neuron 98, 630-644.e16.

- 76 Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain.
77 *Trends in Cognitive Sciences* 17, 401–412.
- 78 Kriegeskorte, N., Mur, M., and Bandettini, P.A. (2008). Representational similarity analysis - connecting the branches of
79 systems neuroscience. *Frontiers in Systems Neuroscience* 2, 4.
- 30 Lacey, S., Tal, N., Amedi, A., and Sathian, K. (2009). A Putative Model of Multisensory Object Representation. *Brain*
31 *Topogr* 21, 269–274.
- 32 Lewicki, M.S. (2002). Efficient coding of natural sounds. *Nature Neuroscience* 5, 356.
- 33 Lewis, C.M., Baldassarre, A., Comitteri, G., Romani, G.L., and Corbetta, M. (2009). Learning sculpts the spontaneous
34 activity of the resting human brain. *P Natl Acad Sci Usa* 106, 17558–17563.
- 35 Li, Y., Luo, H., and Tian, X. (2020). Mental operations in rhythm: Motor-to-sensory transformation mediates imagined
36 singing. *Plos Biol* 18, e3000504.
- 37 Lomber, S.G., Meredith, M.A., and Kral, A. (2010). Cross-modal plasticity in specific auditory cortices underlies visual
38 compensations in the deaf. *Nat Neurosci* 13, 1421–1427.
- 39 Mahon, B.Z., and Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends Cogn Sci*
30 15, 97–103.
- 31 Malone, P.S., Eberhardt, S.P., Wimmer, K., Sprouse, C., Klein, R., Glomb, K., Scholl, C.A., Bokeria, L., Cho, P., Deco, G.,
32 et al. (2019). Neural mechanisms of vibrotactile categorization. *Hum Brain Mapp* 40, 3078–3090.
- 33 Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual*
34 *Information* (Henry Holt and Co., Inc.).
- 35 Mattioni, S., Rezk, M., Battal, C., Bottini, R., Mendoza, K.E.C., Oosterhof, N.N., and Collignon, O. (2020). Categorical
36 representation from sound and sight in the ventral occipito-temporal cortex of sighted and blind. *Elife* 9, e50732.
- 37 McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995). Why there are complementary learning systems in the
38 hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory.
39 *Psychol Rev* 102, 419–457.
- 30 Meijer, P.B.L. (1992). An experimental system for auditory image representations. *Ieee T Bio-Med Eng* 39, 112–121.
- 31 Meredith, M.A., Kryklywy, J., McMillan, A.J., Malhotra, S., Lum-Tai, R., and Lomber, S.G. (2011). Crossmodal
32 reorganization in the early deaf switches sensory, but not behavioral roles of auditory cortex. *Proc National Acad Sci* 108,
33 8856–8861.
- 34 Miyashita, Y. (2019). Perirhinal circuits for memory processing. *Nat Rev Neurosci* 20, 577–592.

- 35 Moore, J.M., and Woolley, S.M.N. (2019). Emergent tuning for learned vocalizations in auditory cortex. *Nat Neurosci* 22,
36 1469–1476.
- 37 Mothe, L.A. de la, Blumell, S., Kajikawa, Y., and Hackett, T.A. (2006a). Cortical connections of the auditory cortex in
38 marmoset monkeys: Core and medial belt regions. *J Comp Neurol* 496, 27–71.
- 39 Mothe, L.A. de la, Blumell, S., Kajikawa, Y., and Hackett, T.A. (2006b). Cortical connections of the auditory cortex in
40 marmoset monkeys: Core and medial belt regions. *J Comp Neurol* 496, 27–71.
- 41 Obleser, J., Eisner, F., and Kotz, S.A. (2008). Bilateral Speech Comprehension Reflects Differential Sensitivity to Spectral
42 and Temporal Features. *J Neurosci* 28, 8116–8123.
- 43 Oh, J., Kwon, J.H., Yang, P.S., and Jeong, J. (2013). Auditory Imagery Modulates Frequency-specific Areas in the Human
44 Auditory Cortex. *J Cognitive Neurosci* 25, 175–187.
- 45 Oosterhof, N.N., Wiestler, T., Downing, P.E., and Diedrichsen, J. (2011). A comparison of volume-based and surface-
46 based multi-voxel pattern analysis. *Neuroimage* 56, 593–600.
- 47 O'Reilly, R.C., and Rudy, J.W. (2001). Conjunctive representations in learning and memory: Principles of cortical and
48 hippocampal function. *Psychol Rev* 108, 311–345.
- 49 Overath, T., McDermott, J.H., Zarate, J.M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal
50 structure revealed by responses to sound quilts. *Nat Neurosci* 18, 903–911.
- 51 Pascual-Leone, A., and Hamilton, R. (2001). The metamodal organization of the brain. *Prog Brain Res* 134, 427–445.
- 52 Perrachione, T.K., and Ghosh, S.S. (2013). Optimized Design and Analysis of Sparse-Sampling fMRI Experiments. *Front*
53 *Neurosci-Switz* 7, 55.
- 54 Pietrini, P., Furey, M.L., Ricciardi, E., Gobbi, M.I., Wu, W.-H.C., Cohen, L., Guazzelli, M., and Haxby, J.V. (2004).
55 Beyond sensory images: Object-based representation in the human ventral pathway. *P Natl Acad Sci Usa* 101, 5658–
56 5663.
- 57 Pouget, A., and Sejnowski, T.J. (1997). Spatial Transformations in the Parietal Cortex Using Basis Functions. *J Cognitive*
58 *Neurosci* 9, 222–237.
- 59 Pouget, A., and Snyder, L.H. (2000). Computational approaches to sensorimotor transformations. *Nat Neurosci* 3, 1192–
60 1198.
- 61 Ptito, M., Moesgaard, S.M., Gjedde, A., and Kupers, R. (2005). Cross-modal plasticity revealed by electrotactile
62 stimulation of the tongue in the congenitally blind. *Brain* 128, 606–614.
- 63 Pugh, K.R., Mencl, W.E., Jenner, A.R., Katz, L., Frost, S.J., Lee, J.R., Shaywitz, S.E., and Shaywitz, B.A. (2001).
64 Neurobiological studies of reading and reading disability. *Journal of Communication Disorders* 34, 479–492.

- 35 Rauschecker, J.P. (1995). Compensatory plasticity and sensory substitution in the cerebral cortex. *Trends Neurosci* 18,
36 36–43.
- 37 Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate
38 human speech processing. *Nature Neuroscience* 12, 718–724.
- 39 Rauschecker, J.P., Tian, B., Korte, M., and Egert, U. (1992). Crossmodal changes in the somatosensory vibrissa/barrel
40 system of visually deprived animals. *Proc National Acad Sci* 89, 5063–5067.
- 41 Reed, C.M., Tan, H.Z., Perez, Z.D., Wilson, E.C., Severgnini, F.M., Jung, J., Martinez, J.S., Jiao, Y., Israr, A., Lau, F., et
42 al. (2018). A Phonemic-Based Tactile Display for Speech Communication. *Ieee T Haptics* 12, 2–17.
- 43 Reich, L., Szwed, M., Cohen, L., and Amedi, A. (2011). A Ventral Visual Stream Reading Center Independent of Visual
44 Experience. *Curr Biol* 21, 363–368.
- 45 Renier, L., Volder, A.G.D., and Rauschecker, J.P. (2014). Cortical plasticity and preserved function in early blindness.
46 *Neurosci Biobehav Rev* 41, 53–63.
- 47 Rezk, M., Cattoir, S., Battal, C., Occelli, V., Mattioni, S., and Collignon, O. (2020). Shared Representation of Visual and
48 Auditory Motion Directions in the Human Middle-Temporal Cortex. *Curr Biol* 30, 2289-2299.e8.
- 49 Ro, T., Ellmore, T.M., and Beauchamp, M.S. (2013). A Neural Link Between Feeling and Hearing. *Cereb Cortex* 23,
50 1724–1730.
- 51 Saad, Z.S., and Reynolds, R.C. (2011). SUMA. *Neuroimage* 62, 768–773.
- 52 Sadato, N., Pascual-Leone, A., Grafman, J., Ibañez, V., Deiber, M.-P., Dold, G., and Hallett, M. (1996). Activation of the
53 primary visual cortex by Braille reading in blind subjects. *Nature* 380, 526–528.
- 54 Saygin, Z.M., Osher, D.E., Koldewyn, K., Reynolds, G., Gabrieli, J.D.E., and Saxe, R.R. (2012). Anatomical connectivity
55 patterns predict face selectivity in the fusiform gyrus. *Nat Neurosci* 15, 321–327.
- 56 Saygin, Z.M., Osher, D.E., Norton, E.S., Youssoufian, D.A., Beach, S.D., Feather, J., Gaab, N., Gabrieli, J.D.E., and
57 Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nat Neurosci* 19,
58 1250–1255.
- 59 Schroeder, C.E., Lindsley, R.W., Specht, C., Marcovici, A., Smiley, J.F., and Javitt, D.C. (2001). Somatosensory Input to
60 Auditory Association Cortex in the Macaque Monkey. *J Neurophysiol* 85, 1322–1327.
- 61 Schroeder, C.E., Smiley, J., Fu, K.G., McGinnis, T., O’Connell, M.N., and Hackett, T.A. (2003). Anatomical mechanisms
62 and functional implications of multisensory convergence in early cortical processing. *Int J Psychophysiol* 50, 5–17.
- 63 Share, D.L. (1999). Phonological Recoding and Orthographic Learning: A Direct Test of the Self-Teaching Hypothesis.
64 *Journal of Experimental Child Psychology* 72, 95–129.

- 35 Simoncelli, E.P., and Olshausen, B.A. (2001). Natural Image Statistics and Neural Representation. *Annu Rev Neurosci*
36 24, 1193–1216.
- 37 Siuda-Krzywicka, K., Bola, Ł., Paplińska, M., Sumera, E., Jednoróg, K., Marchewka, A., Śliwińska, M.W., Amedi, A., and
38 Szwed, M. (2016). Massive cortical reorganization in sighted Braille readers. *Elife* 5, e10762.
- 39 Smiley, J.F., Hackett, T.A., Ulbert, I., Karmas, G., Lakatos, P., Javitt, D.C., and Schroeder, C.E. (2007). Multisensory
40 convergence in auditory cortex, I. Cortical connections of the caudal superior temporal plane in macaque monkeys. *J*
41 *Comp Neurol* 502, 894–923.
- 42 Striem-Amit, E., Cohen, L., Dehaene, S., and Amedi, A. (2012). Reading with Sounds: Sensory Substitution Selectively
43 Activates the Visual Word Form Area in the Blind. *Neuron* 76, 640–652.
- 44 Striem-Amit, E., Ovadia-Caro, S., Caramazza, A., Margulies, D.S., Villringer, A., and Amedi, A. (2015). Functional
45 connectivity of visual cortex in the blind follows retinotopic organization principles. *Brain* 138, 1679–1695.
- 46 Théoret, H., Merabet, L., and Pascual-Leone, A. (2004). Behavioral and neuroplastic changes in the blind: evidence for
47 functionally relevant cross-modal interactions. *J Physiology-Paris* 98, 221–233.
- 48 Tian, X., Ding, N., Teng, X., Bai, F., and Poeppel, D. (2018). Imagined speech influences perceived loudness of sound.
49 *Nat Hum Behav* 2, 225–234.
- 50 Twomey, T., Waters, D., Price, C.J., Evans, S., and MacSweeney, M. (2017). How Auditory Experience Differentially
51 Influences the Function of Left and Right Superior Temporal Cortices. *J Neurosci* 37, 9564–9573.
- 52 Urner, M., Schwarzkopf, D.S., Friston, K., and Rees, G. (2013). Early visual learning induces long-lasting connectivity
53 changes during rest in the human brain. *Neuroimage* 77, 148–156.
- 54 Vetter, P., Bola, Ł., Reich, L., Bennett, M., Muckli, L., and Amedi, A. (2020). Decoding Natural Sounds in Early “Visual”
55 Cortex of Congenitally Blind Individuals. *Curr Biol* 30, 3039-3044.e2.
- 56 Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures
57 for multi-voxel pattern analysis. *NeuroImage* 137, 188–200.
- 58 Zatorre, R.J., and Belin, P. (2001). Spectral and Temporal Processing in Human Auditory Cortex. *Cereb Cortex* 11, 946–
59 953.