bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Reliability of single-subject neural activation patterns in speech production tasks

Saul A. Frankford[a], Alfonso Nieto-Castañón[a], Jason A. Tourville[a], and Frank H. Guenther[a,b,c]

[a]Department of Speech, Language, & Hearing Sciences, Boston University, Boston, MA 02215, USA

[b]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

[c]Department of Radiology, Massachusetts General Hospital, Boston, MA 02114, USA

Declarations of Interest: None

Send correspondence to:

Saul Frankford
Boston University
Department of Speech, Language and Hearing Sciences
635 Commonwealth Avenue
Boston, MA 02215
saulf@bu.edu
(215) 510-7179

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Traditional group fMRI (functional magnetic resonance imaging) analyses are not designed to detect individual differences that may be crucial to better understanding speech disorders. Single-subject research could therefore provide a richer characterization of the neural substrates of speech production in development and disease. Before this line of research can be tackled, however, it is necessary to evaluate whether healthy individuals exhibit reproducible brain activation across multiple sessions during speech production tasks. In the present study, we evaluated the reliability and discriminability of cortical functional magnetic resonance imaging data from twenty neuro-typical subjects who participated in two experiments involving reading aloud mono- or bisyllabic speech stimuli. Using traditional methods like the Dice and intraclass correlation coefficients, we found that most individuals displayed moderate to high reliability, with exceptions likely due to increased head motion in the scanner. Further, this level of reliability for speech production was not directly correlated with reliable patterns in the underlying average blood oxygenation level dependent signal across the brain. Finally, we found that a novel machine-learning subject classifier could identify these individuals by their speech activation patterns with 97% accuracy from among a dataset of seventy-five subjects. These results suggest that single-subject speech research would yield valid results and that investigations into the reliability of speech activation in people with speech disorders are warranted.

Keywords: speech production; fMRI; reliability; classifier

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Introduction

Our understanding of the neural mechanisms responsible for speech and language has dramatically improved in recent decades due to the development of non-invasive techniques for measuring whole-brain activity. Perhaps the most widely used technique of this type is functional magnetic resonance imaging (fMRI); at least 49,000 papers have been published on this topic in pubmed since 2000[1]. To date, the vast majority of fMRI studies of speech and language have involved analyzing group average results from cohorts of 10 or more neurologically normal participants, in many cases compared to similar-sized cohorts of patients with neurological conditions that impact speech or language function. However, many speech disorders result from heterogeneous neural disturbances not easily identified in traditional group analyses (Moser, Basilakos, Fillmore, & Fridriksson, 2016). For example, in the case of acquired apraxia of speech (AOS), a neurogenic disorder that affects speech motor planning and prosody (Duffy, 2013), there is considerable variability in the literature regarding the crucial location of neural damage (Dronkers, 1996; Hillis et al., 2004; Moser et al., 2016). More broadly, there is tremendous variability in the location and extent of stroke-related damage to neural tissue across individuals, which severely limits our ability to characterize brain function in stroke-based disorders using standard group-based fMRI analyses.

An alternative approach to studying stroke-based disorders is to investigate brain activity in individual disordered subjects. A number of studies covering a range of potential purposes have demonstrated or encouraged the use of single-subject fMRI. These include:

[1] Derived from a search of articles on pubmed.com on April 27, 2019 containing the terms “fMRI” or “functional magnetic resonance imaging” in their title or abstract.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

studies of healthy variability and changes over the lifespan (Dosenbach et al., 2010), mapping of language areas prior to resective surgery for patients with epilepsy or gliomas (Babajani-Feremi et al., 2016; Bizzi et al., 2008; Chen & Small, 2007; Gross & Binder, 2014) improved diagnosis of disorders (Raschle, Zuk, & Gaab, 2012; Sundermann, Herr, Schwindt, & Pfleiderer, 2014), and whether neural plasticity following stroke can predict outcomes (Chen & Small, 2007; Kiran et al., 2013; Meltzer, Postman-Caucheteux, McArdle, & Braun, 2009).

These individual-subject approaches depend heavily on the assumption that fMRI data from a single scanning session is reliable. The main purpose of the current study is to test this assumption by assessing the reliability of single-subject fMRI measured during speech production tasks across scanning sessions. Although several prior studies have examined within-subject reliability of BOLD responses during language production tasks (e.g. Mayer, Xu, Paré-Blagoev, & Posse, 2006; Otzenberger, Gounot, Marrer, Namer, & Metz-Lutz, 2005; Wilson, Bautista, Yen, Lauderdale, & Eriksson, 2017), many of these used a covert speech task (Brannen et al., 2001; Harrington, Buonocore, & Farias, 2006; Maldjian, Laurienti, Driskill, & Burdette, 2002; Mayer et al., 2006; Otzenberger et al., 2005; Rutten, Ramsey, van Rijen, & van Veelen, 2002) or have only looked at reliability within language regions of interest (ROIs) like Broca's area and temporo-parietal cortex (e.g., Brannen et al., 2001; Harrington et al., 2006; Mayer et al., 2006; Otzenberger et al., 2005; Rau et al., 2007). However, speech requires overt motor actions and the integration of sensory feedback supported by large and often distant areas of the brain (Guenther, 2016; Sato, Vilain, Lamalle, & Grabski, 2015). Thus far, only two studies (Gorgolewski, Storkey, Bastin, Whittle, & Pernet, 2013; Wilson et al., 2017) have assessed reliability in larger areas of

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cortex during overt word production, but these suffered from low sample sizes, narrow age ranges, and limited reliability measures.

The present study assessed the reliability of speech activation in healthy speakers across multiple sessions and speech tasks. We used the Dice coefficient to measure the spatial overlap of active brain regions within individuals across multiple speech production studies Though crude, it is an easily interpretable measure that can be compared to numerous previous studies of fMRI reliability (Bennett & Miller, 2010). For a more thorough measure that also takes into account the relative scale of activation levels across the brain, we calculated a single-subject intraclass correlation coefficient (ICC; as in Raemaekers et al., 2007). While each of these provides an estimate of similarity that can be used in a single-subject context, further information can be gleaned from measures that assess reliability in relation a between-subjects standard. To evaluate the relative network-wide reliability of activations levels from among the study sample, we calculated a simple ratio of within-subject (across-session) variance compared to between-subject variance. We also wanted to determine which areas of the brain are not only reliable, but highly discriminable across individuals, so we computed an ICC for each vertex on the cortical surface to yield a map of reliability (as in Aron, Gluck, & Poldrack, 2006; Caceres, Hall, Zelaya, Williams, & Mehta, 2009; Freyer et al., 2009; Meltzer et al., 2009). Finally, we directly tested whether individual speakers' neural activation patterns during speech in one study could predict activation in the second study using a machine learning classifier. As our aim was to assess reliability of neural activity specific to speech motor control in healthy individuals measured by fMRI, we included studies with stimuli that removed most higher linguistic processing.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We also investigated whether observed response reliability was associated with the speech task or with intrinsic anatomical and/or resting state functional properties of individual brains. For example, an individual's brain morphometry is largely stable across time, which would make it simple to distinguish individual brains using functional responses if they are highly correlated with brain anatomy. Similarly, it is conceivable that reliability seen during a speech task using fMRI is largely established by unique patterns of the BOLD signal during rest (Jann et al., 2015; Shehzad et al., 2009) or consistent neurovascular organization. To characterize whether within-subject reliability was specific to the speaking tasks or a general property of the BOLD signal in humans, we also assessed the reliability of brain activation not associated with a particular task.

2. Materials and Methods

2.1 Participants

We previously collected data from seventy-five individuals who participated in fMRI studies of speech production in the Speech Lab at Boston University. Of these, data from twenty individuals (mean age: 28.95 years, range: 19-44, 10 female/10 male) who participated in at least two fMRI studies (see Tables 1 and 2) were used to evaluate reliability (median number of days between studies: 13.5, range: 6 - 196). Data from the remaining fifty-five speakers (age range: 18-51) from these and five other speech production studies (see Table 2) were added in the classifier analysis to train the subject classifier and to generalize its features to the broader population of healthy speakers (see section on the classifier analysis). All participants were right-handed native speakers of American English and reported normal or corrected-to-normal vision as well as no history

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of speech, language, hearing, or neurological disorders. Informed consent was obtained from all participants, and each study was approved by the Boston University Institutional Review Board.

2.2. Speech Tasks

All speech tasks included in the present study were overt productions of either real words or pseudowords with at least two consecutive phonemes. These characteristics ensure a distribution of tasks used in neuroimaging studies of speech, while limiting activation patterns to those associated with overt speech production that includes phonemic transitions. A list of speaking tasks and their visual baseline control conditions from each study is included in Table 1.

Study	**Subjects Included**	**Speech Task**	**Visual Baseline**	**Associated Publications**
Consonant Cluster Representation (**CCRS**)	16 Ages: 20-43	Repeating bisyllabic pseudowords that varied in terms of their phonemic, cluster, or syllabic content	"****"	
Syllable Frame Representation (**FRS**)	17 Ages: 20-43	Repeating monosyllabic pseudowords that varied in terms of their phonemic, frame, or syllabic content	"****"	
Auditory Perturbation (**APE**)	6 Ages: 23-36	Monosyllable CVC words (non-perturbed only)	"yyy"	Tourville, Reilly, & Guenther (2008)
Somatosensory Perturbation (**PBB**)	12 Ages: 23-51	VV or VCV pseudowords (non-perturbed only)	"yyy"	Golfinopoulos et al. (2011)
Overt Production (**OP**)	10 Ages: 19-47	CV and CVCV pseudowords	"xxxx"	Ghosh, Tourville, & Guenther (2008)
Syllable Sequence Representation (**SylSeq**)	15 Ages: 18-30	Bisyllabic pseudowords that varied in terms of their phonemic or suprasyllabic content	"XXXXX"	Peeva et al. (2011)
Auditory Category Perturbation (**CAT**)	15 Ages: 19-33	Monosyllable CVC words (non-perturbed only)	"***"	Niziolek and Guenther (2013)

Table 1. Information about the studies from which activation maps were included in the present analyses. C = consonant, V = vowel.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Subject	Studies
1	CCRS, FRS
2	CCRS, FRS
3	CCRS, FRS
4	CCRS, FRS
5	CCRS, FRS
6	CCRS, FRS
7	CCRS, FRS
8	CCRS, FRS
9	CCRS, FRS
10	CCRS, FRS
11	CCRS, FRS
12	CCRS, FRS
13	CCRS, FRS
14	CCRS, FRS
15	APE, PBB
16	APE, PBB
17	APE, PBB
18	APE, PBB
19	APE, PBB
20	APE, PBB

Table 2. Studies in which each test subject participated. Study identification codes refer to abbreviations in the 'Study' column of Table 1.

2.2. Image Acquisition

MRI data were acquired at the Athinoula A. Martinos Center for Biomedical Imaging at Massachusetts General Hospital (APE, PBB, OP, CCRS, FRS), the Athinoula A. Martinos imaging Center at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology (CAT), and the fMRI Centre of Marseille (SylSeq). For each study, a high-resolution structural scan was acquired in addition to the task fMRI scans. For CCRS and FRS, data were acquired using a 3 Tesla Siemens Trio Tim scanner with a 32-channel head coil. For each subject, a high-resolution T1-weighted volume was acquired (MPRAGE, voxel size: 1 mm^3, 256 sagittal images, TR: 2530 ms, TE: 3.44 ms, flip angle: 7°). Functional gradient echo – echo planar imaging (EPI) scans (41 horizontal slices, in plane resolution: 3.1 mm, slice thickness: 3 mm, gap: 25%, TR: 2.5 s, TA: 2.5 s, TE: 20 ms) were automatically

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

registered to the AC-PC line and were collected continuously. See Peeva et al. (2010), Tourville, Reilly, & Guenther (2008), Golfinopoulos et al., (2011), Ghosh, Tourville, & Guenther (2008), and Niziolek & Guenther (2013) for acquisition parameters for the Sylseq, APE, PBB, OP, and CAT studies, respectively (refer to Table 1 for study codes).

2.3. Preprocessing and first-level analysis

Preprocessing was carried out using SPM12 (http://www.fil.ion.ucl.ac.uk/spm) and the CONN toolbox (Whitfield-Gabrieli & Nieto-Castanon, 2012) preprocessing modules. Each participant's functional data were motion corrected to their mean functional image, and coregistered to their structural image. For CCRS and FRS, BOLD responses were high-pass filtered with a 128-second cutoff period and estimated at each voxel using a general linear model (GLM). The hemodynamic response function (HRF) for each stimulus block was modeled using a canonical HRF convolved with the trial duration from each study. For each run, a linear regressor was added to the model to remove linear effects of time, as were six motion covariates and a 'constant' effect (the intercept for that run). See Peeva et al. (2010), Tourville, Reilly, & Guenther (2008), Golfinopoulos et al., (2011), Ghosh, Tourville, & Guenther (2008), and Niziolek & Guenther (2013) for first-level design details in the other studies. Regressors were added for all studies to remove the effects of volumes with excessive motion and global signal change using ART (https://www.nitrc.org/projects/artifact_detect/) with a scan-to-scan motion threshold of 0.9mm and a scan-to-scan signal intensity threshold of at least 5 standard deviations above the mean.

In all studies and subjects, first-level model estimates for each speech condition and baseline were contrasted at each voxel and averaged across all study-specific speech conditions to obtain speech activation maps (*speech* maps). To obtain maps of average

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BOLD signal activity not explained by task effects (*null* maps), estimates of the constant effect of each run were averaged for each subject in each study. These maps represent the average BOLD signal after the effects of speech, baseline, motion, and outliers have been removed. Effect size maps were used for subsequent analyses rather than significance (*p*-value) maps because a) significance maps are not as consistent for individual subjects as they are for group analyses (Gross & Binder, 2014; Voyvodic, 2012) and b) previous research has demonstrated greater overlap in effect size maps (Wilson et al., 2017).

T1 volume segmentation and surface reconstruction were carried out using the FreeSurfer image analysis suite (freesurfer.net; Fischl, Sereno, & Dale, 1999). Activation maps were then projected to each individual's inflated structural surface. To align subject data, individual surfaces were inflated to a sphere and coregistered with the FreeSurfer mean surface template (fsaverage; see Figure 1). Surface maps were then smoothed with 40 diffusion steps (equivalent to a 10.8mm full-width half maximum smoothing kernel). This level of smoothing has previously been shown to optimize reliability of task-related BOLD response data in individuals (Caceres et al., 2009).

Maps of random activation (*random* maps) were created by independently replacing effect sizes at each vertex with a randomly chosen value from a normal distribution with a mean of 0 and a standard deviation of 1. These maps were included to estimate the results for each analysis under the assumption that no systematic relationship exists between maps from each subject and session.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

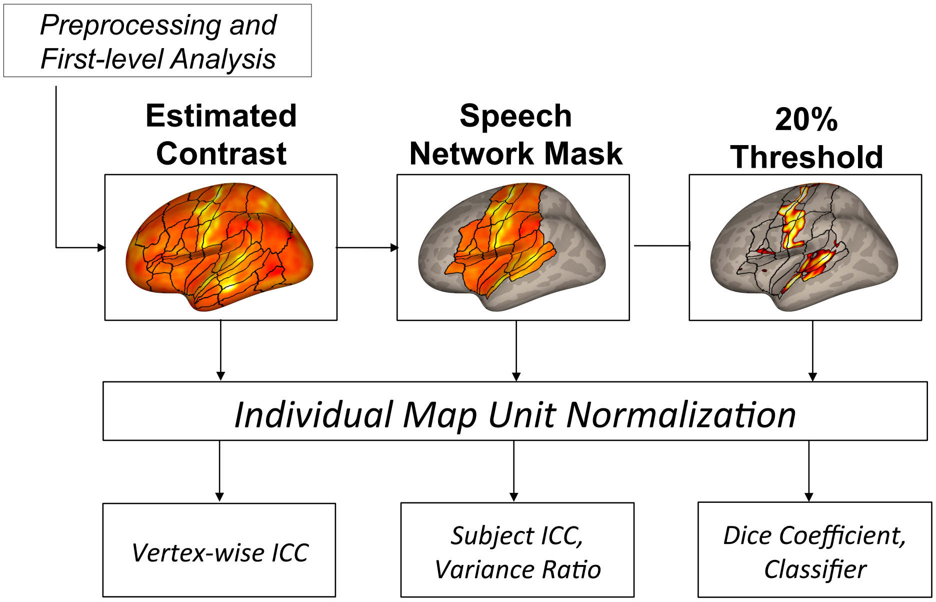
Preprocessing and First-level Analysis
Estimated Contrast
Speech Network Mask
20% Threshold
Individual Map Unit Normalization
Vertex-wise ICC
Subject ICC, Variance Ratio
Dice Coefficient, Classifier

Figure 1. Thresholding pipeline map for each of the reliability analyses. After preprocessing and estimation of first-level condition effects, the *speech* and *null* maps were calculated and fed into the vertex-wise ICC analysis. A speech network mask is applied, so that only vertices inside this mask are used for the single-subject ICC and variance ratio measures. Finally, the 20% of vertices with the highest activation levels are kept for the Dice coefficient and classifier analysis. Prior to calculating reliability measures, all maps were normalized to account for differences in effect size scaling between subjects and studies. Outlines for regions of interest previously described in Tourville & Guenther (2012) are included for reference, and appear only in areas of cortex on which a given analysis was carried out.

2.4. Reliability Measures

We used five different measures to quantify individual-subject activation reliability across different sessions (the term *session* will be used going forward to refer to a data collection time points): the Dice coefficient, a single-subject intraclass correlation coefficient, a simple variance ratio, a vertex-wise intraclass correlation coefficient, and a machine-learning classifier. Each of these measures was applied to both the *speech* and *null* maps.

2.4.1. Single-subject Spatial Overlap

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To measure the spatial overlap of supra-threshold vertices, we used the Dice coefficient, a metric widely used in fMRI reliability studies (see Bennett & Miller, 2010 for a review). It is the ratio between the extent of overlap and the average size of the individual maps and yields values between 0 (no overlap) and 1 (complete overlap). A strength of this measure is that it is generally straightforward to interpret and provides a simple way to characterize the reproducibility of thresholded activation maps (Bennett & Miller, 2013). On the other hand, the Dice coefficient is sensitive to how this map is thresholded (Duncan, Pattamadilok, Knierim, & Devlin, 2009; Smith et al., 2005), and the area over which the calculation is made (Gorgolewski et al., 2013), where lower thresholds and whole-brain analyses will tend to increase overlap. Despite this, the Dice coefficient provides a rough estimate of neural response reliability.

The Dice coefficient is formally given by:

$$R_{overlap} = \frac{2 * A_{overlap}}{A_1 + A_2} \qquad (Eq. 1),$$

where A_1 and A_2 are defined as the number of supra-threshold vertices for individual sessions and $A_{overlap}$ is the total number of vertices that exceeds the threshold in both sessions(Bennett & Miller, 2010). Because we were only interested in assessing reliability in brain areas commonly activated during speech production, we masked each map to only analyze activation within a predefined speech production network area covering about 35% of cortex (see Figure 1; J.A. Tourville & Guenther, 2012). Since Dice coefficients operate on binary maps, activation maps were then thresholded, keeping the highest 20% of all surface vertices within the masked area (7% of total cortex; see Figure 2 for examples of these thresholded maps).

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.4.2. Single-subject ICC

To assess the reliability of relative activation levels in the speech network (as opposed to merely measuring the overlap of suprathreshold vertices), we calculated a single-subject ICC (see Raemaekers et al., 2007) for each subject that compares variance between sessions to within-session (across-vertex) variance. Like the Dice coefficient, the ICC is relatively straightforward to interpret: a value of 0 means that relative activation levels are entirely unreliable, while a value of 1 signifies complete reliability. Of the many types of ICCs described in the literature, we used the ICC(1) as defined in McGraw and Wong (1996). This type of ICC is based on an analysis of variance (ANOVA) of the following one-way random effects model:

$$y_{ij} = \mu + b_i + s_{ij} \qquad (Eq. 2),$$

where y_{ij} is the value for the i^{th} vertex and the j^{th} session, μ is the mean value across all vertices and session, b_i is the between-vertices effect at vertex i, and s_{ij} is the residual, representing the between-sessions effect. ICC(1) estimates the degree of absolute agreement across multiple repetitions of a set of measurements. Formally, it is an estimate of

$$ICC(1) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_s^2} \qquad (Eq. 3),$$

where σ_b^2 is the between-vertex variance and σ_s^2 is the between-sessions variance. Based on McGraw and Wong (1996), the sample estimate, $\widehat{ICC(1)}$, can be calculated using the following formula:

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\widehat{ICC(1)} = \frac{MS_b - MS_s}{MS_b + (k-1)MS_s} \qquad (Eq. 4),$$

where MS_b is the mean squares across vertices, MS_s is the mean squares of the residuals, and k is the number of within-subjects measurements (in this case, 2 sessions).

For this analysis, we used the same speech production mask as in the overlap analysis. No threshold was applied, however, since, unlike Dice coefficients, ICC does not require the maps to be binarized. This allows us to characterize the reliability not only in spatial location but also in the relative scale of the activation responses. To account for any gross scaling differences in effect sizes across contrasts and sessions that could affect the this ICC (McGraw & Wong, 1996), effect sizes were normalized within each map immediately prior to each analysis by dividing the value at each vertex by the Euclidian norm of all the vertices in the map.

2.4.3. Variance Ratio

To quantify how reliable the average vertex activation is for individual subjects across sessions compared to across subjects, we calculated a simple ratio of within-subject variance and between-subject variance, averaged across all masked vertices. The within-subject variance measure for each subject was calculated as:

$$Var_w = \frac{1}{2n}\sum_{i=1}^{n}(y1_i - y2_i)^2 \qquad (Eq. 5),$$

where $y1_i$ and $y2_i$ are the activation levels of the i^{th} vertex in each session, and n is the total number of vertices. Thus, Var_w was the mean squared difference between session maps

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from each subject (similar to the t$_{\text{diff}}$ measure in Gorgolewski et al. (2013) but with effect sizes instead of t-statistics). The between-subject variance was calculated as:

$$Var_b = \frac{k}{n(N-1)} \sum_{i=1}^{n} \sum_{j=1}^{N} \left(\overline{y_{\imath\jmath}} - \overline{y}_{\imath}\right)^2 \qquad (Eq. 6),$$

where $\overline{y_{s\imath}}$ is the mean activation level in the i^{th} vertex from the j^{th} subject across sessions, $\overline{y}_{\imath}$ is the mean activation level of the i^{th} vertex across subjects and sessions, *k* is the number of sessions, *n* is the total number of vertices, and *N* is the total number of subjects. In other words, Var_b was the squared difference between each subject's mean map and the grand mean map, averaged across vertices. Then, the variance ratio for each subject was defined by:

$$Var_{rat} = \frac{Var_w}{Var_b} \qquad (Eq. 7).$$

Values of Var_{rat} below 1 would indicate that activation was relatively more consistent across sessions for a given subject than across subjects. We calculated this ratio for the *speech, null, and* r*andom* maps. To capture only the reliability within the speech network, these measures were calculated on the masked activation maps that were unit normalized.

2.4.4. Vertex-wise Reliability

As in previous fMRI reliability studies (Aron et al., 2006; Caceres et al., 2009; Freyer et al., 2009; Meltzer et al., 2009), we used the ICC to determine the vertex-wise reliability of individuals across sessions. This analysis used the ICC(1) as in 2.4.2, but we defined MS_b in Eq. 4 as the mean squares between subjects, while MS_s and *k* remained the same. Then, to focus our results on vertices that exhibited 'good' or 'excellent' reliability, we used Koo &

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Li's (2016) convention to threshold the resulting ICC map, keeping only those vertices with a value greater than or equal to 0.75. We applied this analysis to all cortical vertices (without a speech network mask) in order to compare the reliability of vertices within speech-related areas to those not usually associated with speech. Doing so would reveal whether high reliability was specific to the speech network or whether other areas not commonly active during speech production also demonstrate high reliability during speech tasks. As with the previously described analyses, activation values in each map were unit normalized.

2.4.5. Subject Classifier

Machine-learning tools have recently been applied to MRI data to detect whether subject groups (e.g., patient and control) are discriminable by their neural structure and function (see Sundermann et al., 2014 for a review). Here, we implemented a nearest-neighbor subject classifier to assess both the reliability and discriminability of *speech*, *null,* and *random* maps (separately) for individual subjects. A leave-one-out cross-validation procedure was employed to ensure maximum training data for the classifier. On each iteration, one session activation map from among the 20 subjects who were scanned twice was used as the test dataset. A session activation map from all 75 subjects was then used for training; the classifier was always trained on one activation map from each subject to avoid biasing effects due to uneven training samples. For the subjects that had two maps, one training map was chosen at random (excepting the test dataset). The entire procedure was repeated for both session activation maps from each subject (a total of 40 times). For this analysis, we used maps that were masked, thresholded, and unit normalized (see

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2 for examples). This meant that subjects were classified by the spatial extent and relative activation values of the most active vertices.

Due to the high dimensionality of fMRI activation data, we used singular-value decomposition (SVD) to extract a small number of features that account for the most variance across all subjects and sessions in the training set. Before computing these features, data maps were de-meaned by subtracting the mean vertex value in each subject map. The training and testing activation maps were then projected onto this low-dimensional space, and the resulting scores were divided by the singular values (characterizing the standard deviation of the original data across each dimension). This ensured that all components were weighted equally, independent of the variance explained by each component. The nearest-neighbor classifier then selected the subject within the training set that had the smallest Euclidean distance to the test map. This procedure was repeated for all activation test maps in the dataset and a percent accuracy score was obtained across the whole dataset.

The number of features used to train this classifier was varied between 1 and the maximum number extracted from the SVD (equal to the total number of subjects used) to determine the power of the classifier in both *speech* and *null* conditions. To ensure that high accuracy was not due to some bias of the thresholding steps or type of classifier we used, *random* maps were also run through the classifier. Finally, confidence intervals for each number of features were estimated by repeating each analysis 20 times.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

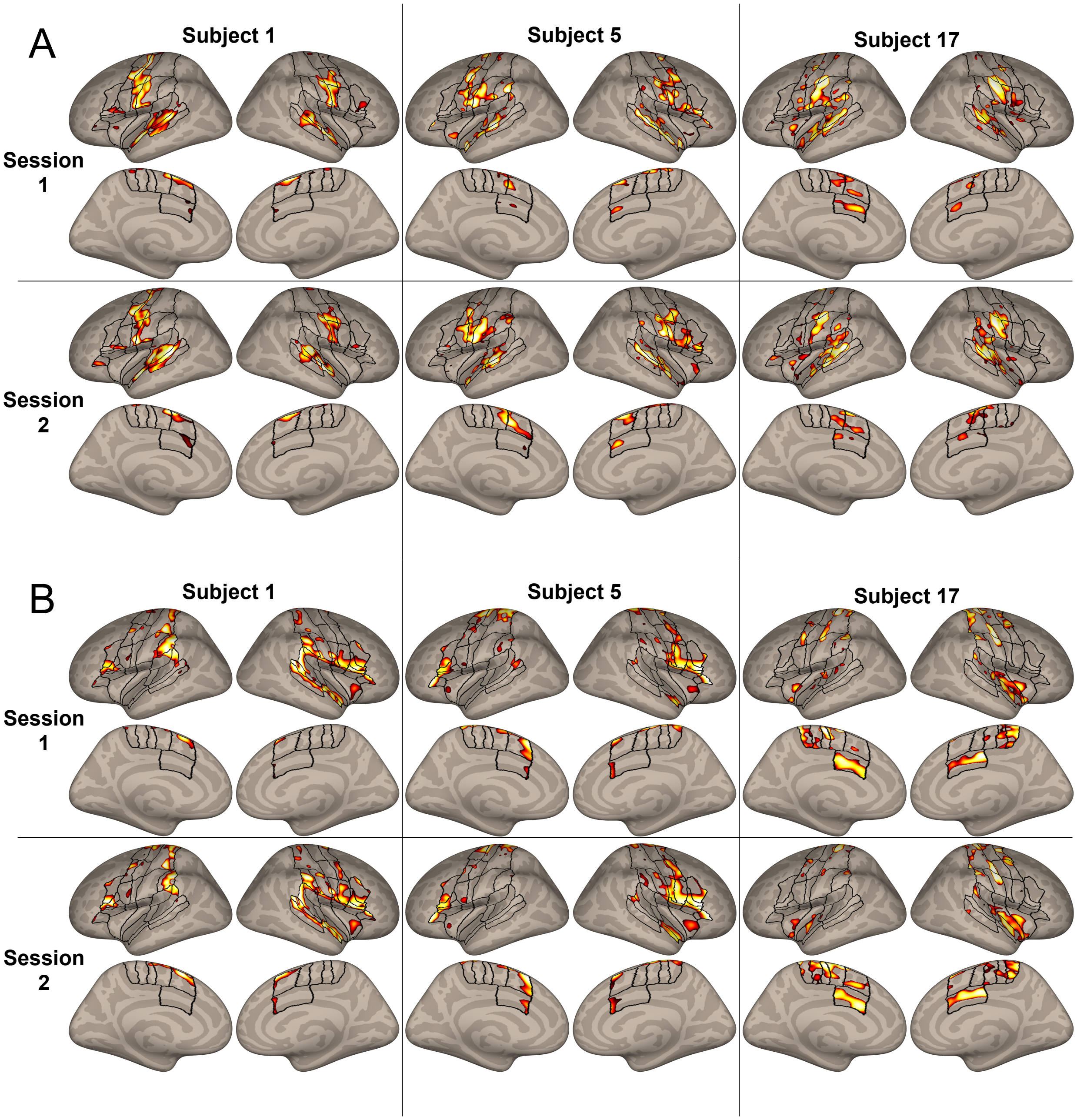

Figure 2. A. Masked and thresholded *speech* maps for three subjects in both sessions. Outlines of regions of interest covering the masked speech network previously described in Tourville & Guenther (2012) are included for reference. B. Masked and thresholded *null* maps for the same subjects. In both cases, the activation peaks display broad visual similarity between sessions. Note: the color scale indicates the rank of vertex activation within each map, where lighter colors indicate higher activation.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.5. Group-level Statistical Analyses

To determine whether reliability during the speech task was greater than that of the brain activity not related to speech, we directly compared the Dice coefficient and single-subject ICC values from the *speech* and *null* conditions across subjects. Because these values were not assumed to follow a normal distribution, repeated-measures Wilcoxon Signed-Ranks tests were run to test for these differences. For the single-subject ICC analysis, we also compared individual ICC values with a Between-Subjects ICC group measure. This measure was calculated in the same way as the individual ICC values, substituting in subject activation maps averaged across sessions for individual session maps. We also calculated the Spearman correlations between the *speech* and *null* maps in these measures to determine whether reliability in these two conditions was related (i.e. whether high reliability in the *speech* condition also meant high reliability in the *null* condition). For the variance ratio analysis, we used a Kruskal-Wallace one-way ANOVA to compare *speech*, *null*, and randomized data, with post-hoc comparisons between each pair of conditions using Tukey's honestly significant difference procedure.

2.6. Data and Code Sharing Statement

All anonymized data and analysis code are available upon reasonable request in accordance with the requirements of the institute, the funding body, and the institutional ethics board.

3. Results

3.1. Single-subject Spatial Overlap

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Dice coefficient for each subject's thresholded *speech* maps compared between scanning sessions can be found in Figure 3. On average, their Dice coefficient was 0.693 (SD: 0.089), demonstrating approximately 69% spatial overlap of individual activation maps. For individual *null* maps, the Dice coefficient between the activation peaks in Experiment 1 and Experiment 2 are also shown in Figure 3. On average, individuals have a Dice coefficient of 0.726 (SD: 0.110), indicating about 73% spatial overlap across sessions. To understand how these values would compare to subjects with completely uncorrelated activation maps, *random* maps yielded a Dice coefficient of 0.2 (as expected, since only voxels with the highest 20% of effect sizes in each map were included). For the group comparison, although *speech* scores were lower than *null* scores, this comparison was not significant (z=-1.31, p=0.191). Further, there was no correlation between Dice coefficients for *speech* and *null* maps (Spearman's r = 0.098, p = 0.681).

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

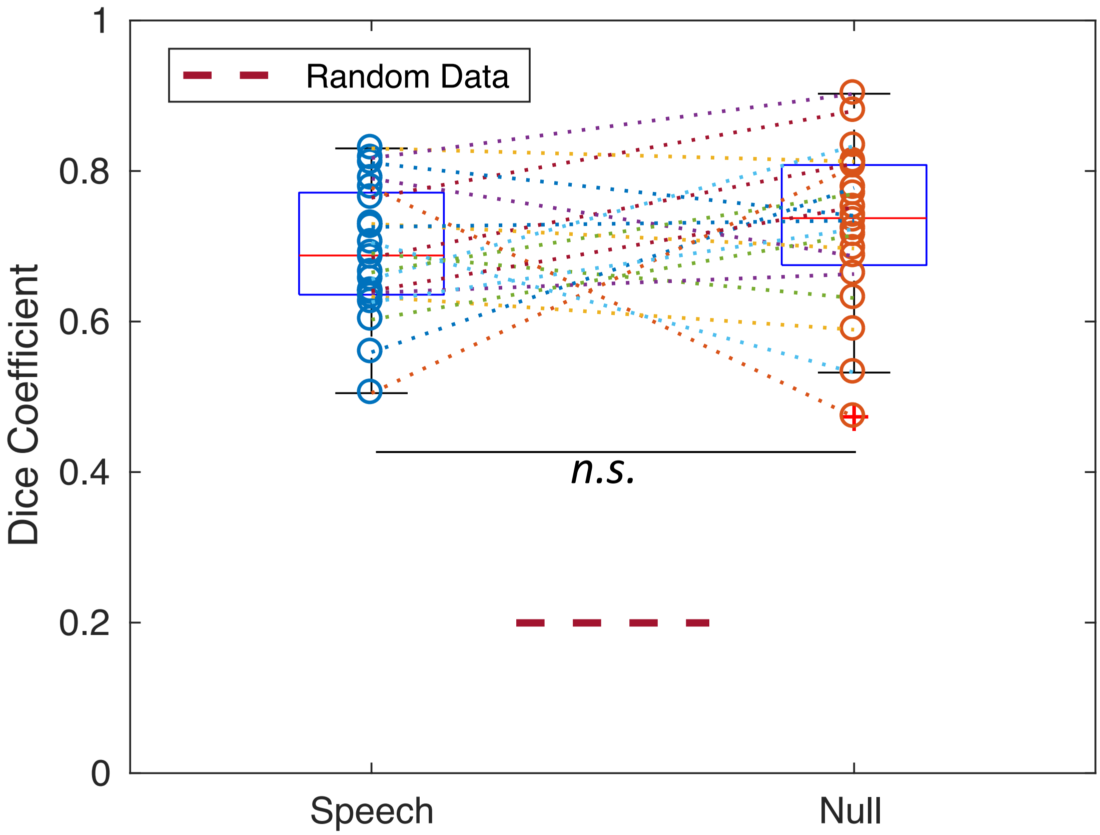

Figure 3. Comparison of single-subject Dice coefficient values in the *speech* and *null* thresholded activation maps. Values for individual subjects are shown as circles in each condition, and dashed lines connect results from individual subjects across conditions. For each condition: red line = median; blue box = interquartile range (25th-75th percentile); black lines = boundary of values for data points that fall within 1.5 times the IQR away from the edges of the box; red crosses signify outliers – values that fall outside the black lines. *n.s.:* non-significant at alpha = 0.05.

3.2. Single-subject ICC

The distribution of within-subject *speech* ICC values across sessions can be found in Figure 4. Individual subjects exhibited poor (0.196) to good (0.868) reliability (Koo & Li, 2016), with a mean ICC(1) of 0.721 (SD: 0.172). As a comparison, the between-subjects correlation, calculated on the averaged individual activation maps across both sessions, was poor with a value of 0.475. A Wilcoxon Signed-Ranks test shows that the median of the

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

within-subject ICCs was significantly higher than the between-subject ICC (z=3.51, p<0.001). For the *null* condition, individuals showed moderate (0.622) to excellent (0.976) within-subject reliability, with a mean ICC(1) of 0.870 (SD: 0.092). The between-subjects correlation for this condition was poor at 0.345, and the median of the within-subject coefficients was significantly greater than this value (z=3.92, p<0.001). The within-subject ICCs for the *null* maps were significantly greater than the ICCs for the *speech* maps (z=3.17, p=0.002), and *random* maps yielded an ICC of 0 as expected. Similar to the Dice coefficient, there was no significant correlation between ICC values in the *speech* and *null* conditions (Spearman's r = 0.173, p = 0.464).

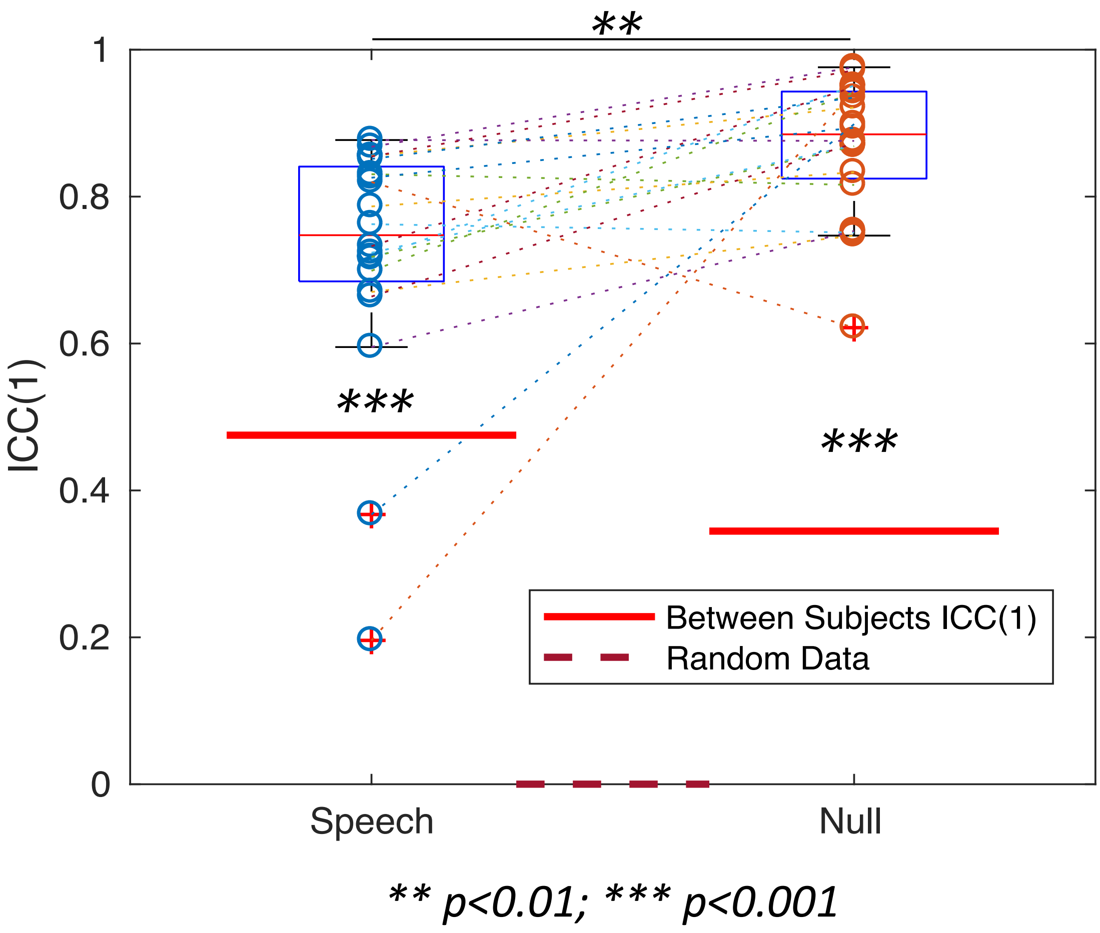

Figure 4. Comparison of single-subject ICC values in the *speech* and *null* activation maps.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Values for individual subjects are shown as circles in each condition, and dashed lines connect results from individual subjects across conditions. Asterisks in line with each condition show comparison between the distribution of individual points and the Between-Subjects ICC. Red crosses signify outliers - data points that fall at least 1.5 times the IQR away from the edges of the box.

3.3. Variance Ratio

The ratios between within-subject variance and between-subject variance in all conditions are found in Figure 5. A ratio of 0.296 (SD: 0.212) for the *speech* activation maps, 0.096 (SD: 0.083) for the *null* maps, and 1.000 (SD: 0.004) for the randomized maps (as expected). There was an overall significant effect of condition (χ^2 = 47.95, p < 0.001), and post-hoc tests showed significant differences between each pair of conditions (*speech* vs. *null*: p = 0.009; *speech* vs. random: p < 0.001; *null* vs. random: p < 0.001).

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

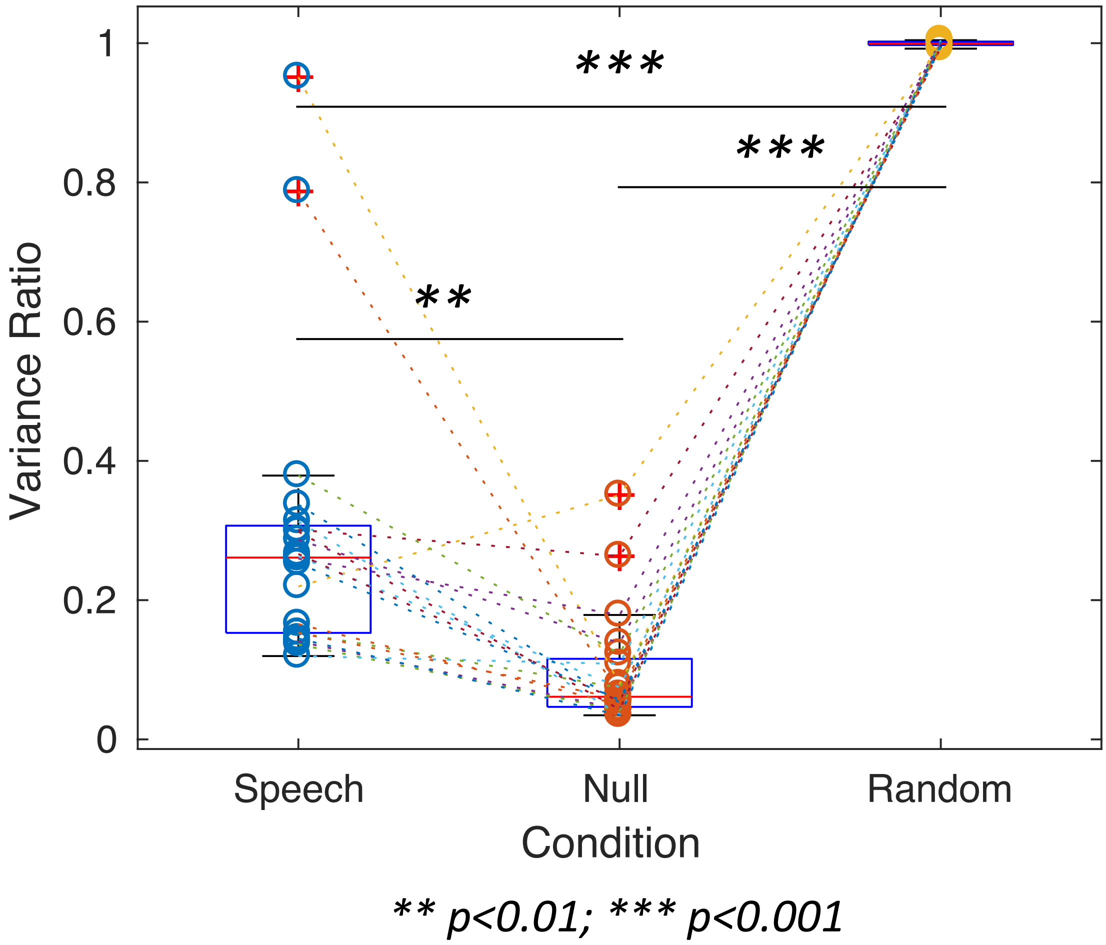

Figure 5. Variance ratios compared across *speech*, *null*, and *random* maps. Values for individual subjects are shown as circles in each condition, and dashed lines connect results from individual subjects across conditions. Red crosses signify outliers - data points that fall at least 1.5 times the IQR away from the edges of the box.

3.4. Vertex-wise Reliability

The vertex-wise ICC map for the *speech* data thresholded at 0.75 can be found in Figure 6. While much of cortex was found to have ICC values greater than 0.5 (see Supplementary Figures 1 and 2 for an unthresholded ICC map of *speech* and *null* data), the highest within-subject reliability (>0.75) appeared in areas commonly activated during speech production including on the lateral surface: bilateral motor and somatosensory

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cortex, bilateral auditory cortex, bilateral inferior frontal gyrus (IFG) *pars opercularis*, left anterior insula, and bilateral anterior supramarginal gyrus; and on the medial surface, bilateral supplementary and pre-supplementary motor areas, and bilateral cingulate motor area. Some additional regions showed high reliability as well: bilateral IFG pars orbitalis, right anterior insula, bilateral middle temporal gyrus, and bilateral posterior cingulate cortex. Thus, the speech production network accounts for most of the regions with high within-subject reliability.

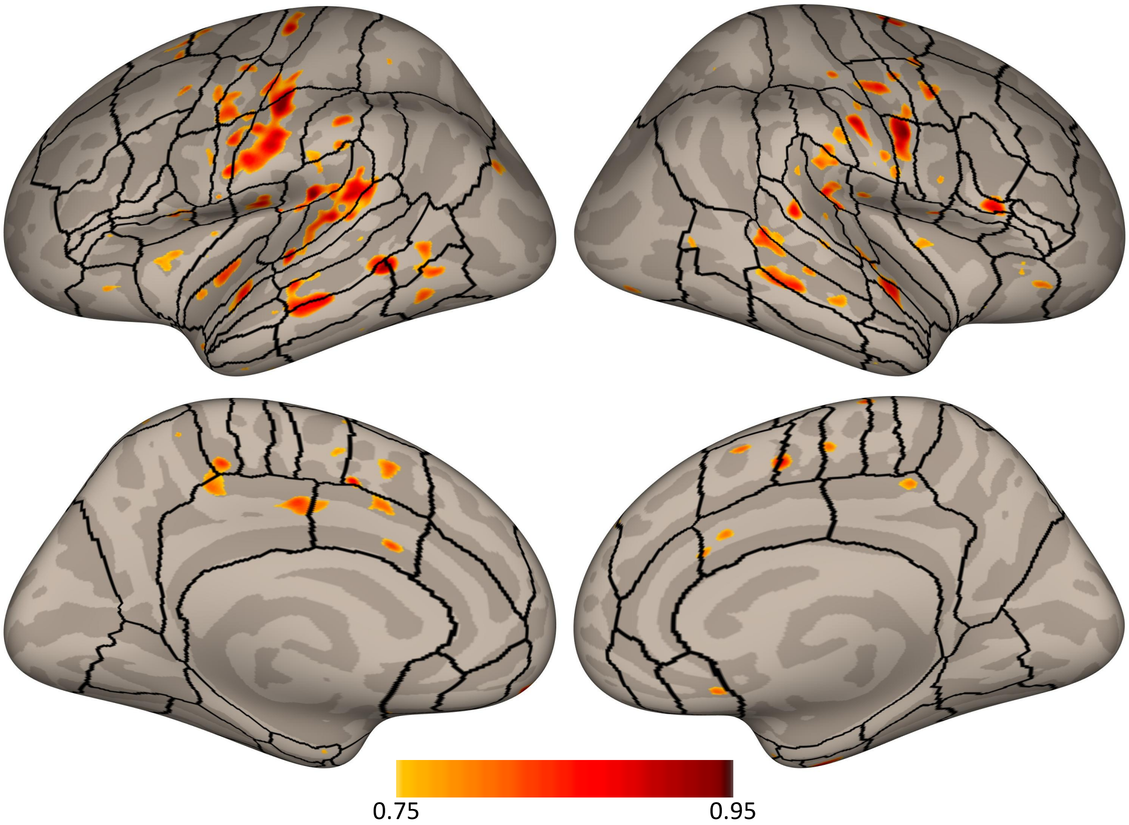

Figure 6. Vertex-wise ICC values for the *speech* activation maps thresholded at 0.75. Regions of interest previously described in Tourville & Guenther (2012) are included for reference.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3.5. Subject Classifier

Accuracy of the subject classifier for the *speech* and *null* maps is displayed in Figure 7. For the *speech* maps, classifier accuracy for untrained test data approached 97% when 75 features were used. Similarly, the accuracy of this classification method reached 95% for the *null* activation maps when all principal components were included. In addition, both analyses surpassed 75% accuracy with as few as 23 features. Thus, subjects were highly distinct in that they could be discriminated using a relatively small number of features.

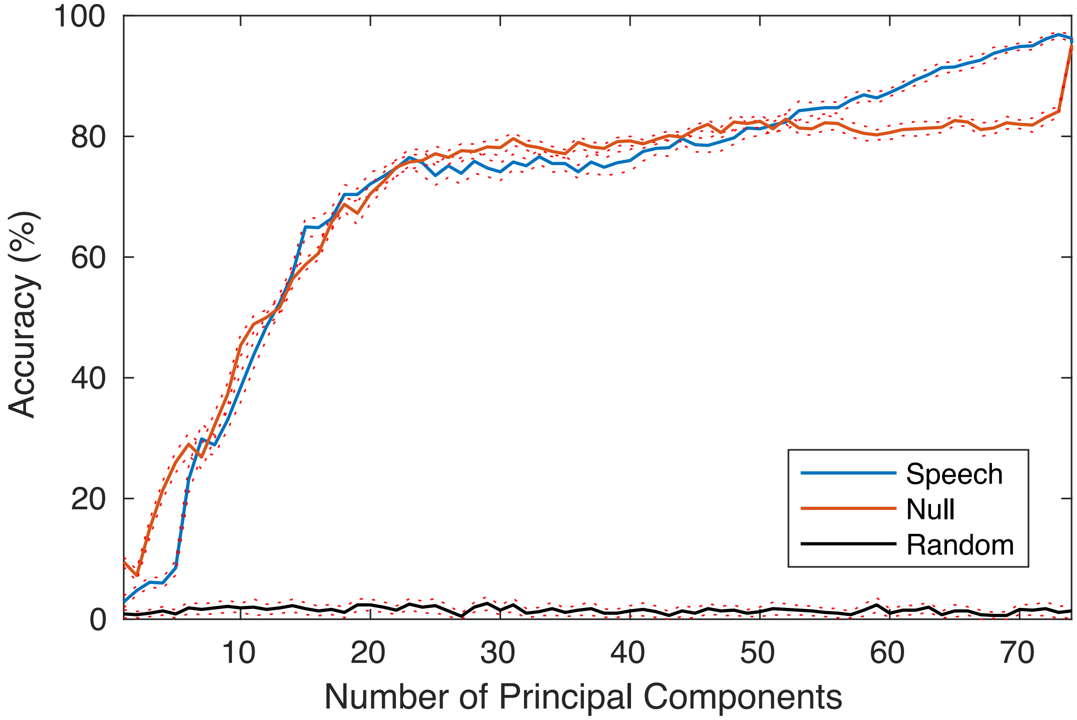

Figure 7. Subject classifier accuracy for *speech*, *null*, and random data across 75 dimensions. Solid and dashed lines show the mean accuracy and 95% confidence intervals, respectively, across 20 repetitions of the analysis.

To assess whether these results were better than chance, we substituted *random* maps for each subject's *speech* surface maps (while maintaining the number of maps that

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

each subject has). These results show that random data performed near 1%, as expected given 75 potential classes.

4. Discussion

Characterizing individual reliability in speech activation is an important step toward validating single-subject speech research in persons with and without speech disorders. In this study, we used five methods to assess reliability and discriminability in a group of 20 healthy speakers.

4.1. Activation Reliability

The Dice coefficient and single-subject ICC results in this study demonstrated that both the extent and degree of activation patterns during speech production in most individuals showed moderate to high amounts of reliability. The Dice values found in this study were generally larger than those found in previous overt expressive language studies (Gorgolewski et al., 2013; Wilson et al., 2017). There are several possibilities as to why this was the case. First, as previously discussed, the Dice coefficient is inherently tied to the thresholding scheme used. Gorgolewski et al. (2013) used statistically thresholded maps (although with an advanced thresholding procedure; Gorgolewski, Storkey, Bastin, & Pernet, 2012) as opposed to effect size maps with a percent threshold; statistically thresholded maps can be strongly affected by multiple factors including noise from head motion and total scan time (Bennett & Miller, 2010; Gross & Binder, 2014). Indeed, scan time in Gorgolewski et al. (2013) was less than 8 minutes (36 speech trials and 36 baseline) compared to an average of 38 (range: 27 – 65) minutes in the present analysis (mean of

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

114 speech trials, range: 72 – 240; mean of 28 baseline trials, range 18 – 64), likely leading to differences in power as shown previously (Friedman et al., 2008). Second, even at similar levels of thresholding (Wilson et al., 2017), reducing the region of interest to pre-defined cortical speech areas in the present study eliminates extraneous regions that show session-specific activations not related to speech *per se*. In Wilson et al. (2017), Dice values in predefined language regions were notably lower than when they looked at all supratentorial voxels, suggesting that higher-level language processing may lead to more variable activation, have lower signal change, and/or contain more noise. Gorgolewski et al. (2013) reported the opposite effect, although Dice values for this task were only specified for auditory cortices. Third, Wilson et al. (2017) did not use a control condition because their goal was to test language-mapping paradigms for individuals with aphasia who may have task-switching difficulties. This may have led to activation variability in brain areas not directly related to the task. Finally, the older age cohorts used in Gorgolewski et al. (2013; age range: 50-58 years) and Wilson et al. (2017; age range: 70-76 years) may have had reduced reliability due to various factors that decrease signal-to-noise ratio in the BOLD signal in older adults (D'Esposito, Deouell, & Gazzaley, 2003) .

The single-subject ICC measure we employed in this study measured the degree of reliability between two cortical activation maps. While it relied only on within-subject sources of variance, it was highly correlated with the Dice coefficient (*speech*: Spearman's r = 0.902, p < 0.001; *null*: r = 0.949, p < 0.001) thus demonstrating its validity as a measure of test-retest reliability. One major difference that appeared between this measure and the Dice coefficient was that the *null* condition yielded higher ICC values than the *speech* maps with some subjects attaining near perfect correspondence. This seems to demonstrate that

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

once all task and motion parameters are accounted for, the underlying BOLD signal maintains high reliability for individuals across scanning sessions. Nonetheless, both *speech* and *null* maps generally demonstrated greater within-subject reliability than a matched between-subjects measure.

The variance ratio we calculated was a simple comparison of within-subject variance to between-subject variance where values below 1 demonstrate greater reliability within-subjects. Because the within-subject variance component is the same as in the single-subject ICC measure above and affects the outcome in inversely proportional ways, the variance ratio largely mirrored those results. There were, however, two participants (Subject 6 and Subject 7) whose ICC scores for the *speech* maps were less than the between-subjects score and whose variance ratio scores approached 1. In each of these cases, the median beta value across vertices for one of the two scanning sessions (the CCRS study session) was more negative for these two subjects than for any other subjects. This might imply that these subjects had less power for the *speech* contrasts in CCRS. Although they had similar numbers of speech trials as the other subjects, they were among the subjects with the highest scan-to-scan motion and global signal change for this study. They also happened to have the two highest scan-to-scan global signal change values for the other study session (FRS), though motion was more average for this study. Changes in global signal are often artifacts associated with subject motion, although other physiological sources contribute to this measure (see Liu, Nalci, & Falahpour, 2017). However, their motion was not excessive for typical neuroimaging sessions and other subjects with similar amounts of scan-to-scan motion and signal change maintained among the highest ICC values. Another potential reason that these two subjects had much lower

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ICC scores is methodological: since the ICC(1) measures absolute agreement rather than consistency (McGraw & Wong, 1996), it does not account for global differences in effect sizes across studies. We attempted to correct for this by unit-normalizing vertex values for each subject in each study, but this is not a perfect method. Indeed, the absolute difference in the median vertex values from each study was much greater for these subjects. Thus, both data quality and methodological choices likely drove down their reliability scores.

In sum, we found high within-subject reliability of activation in the speech network, except in two cases where motion may have negatively impacted the signal-to-noise ratio.

4.2. Activation Discriminability

The other two measures we used assess reliability by comparing response variability within subjects (across sessions) to variability between subjects. These measures characterize individual reliability relative to the sample, but additionally assess how discriminable individuals are from one another. The vertex-wise *speech* ICC map paralleled previous studies that calculated this metric – many of the areas where ICC values were high corresponded to areas commonly activated during the task (Aron et al., 2006; Caceres et al., 2009; Freyer et al., 2009; Meltzer et al., 2009). Thus, for speech production, speech-related areas in somato-motor cortex, medial and lateral pre-motor cortex and extended areas of auditory cortex were consistent for individual subjects across scanning sessions. In addition, even areas of cortex inconsistently active during speech production like IFG *pars orbitalis*, middle temporal gyrus (MTG), and posterior cingulate gyrus (PCG) showed high reliability. In a review of fMRI studies of speech and language processing (Price, 2012), both IFG pars orbitalis and MTG were associated with semantic processing,

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

while MTG was also associated with translating orthography into sound. This second explanation would be relevant because all tasks involve reading aloud, but it is less clear why semantic processing centers would be highly reliable for pseudoword speaking tasks. The PCG is part of the default mode network and appears to help modulate attentional control (Leech & Sharp, 2014). Thus individuals may consistently activate or deactivate this region depending on their level of attention during speaking tasks. Previous studies of higher-level cognitive tasks have found reliable activation outside of areas commonly associated with the task, but this usually occurred in sensory and motor regions needed to complete the task (Aron et al., 2006; Freyer et al., 2009). Caceres et al. (2009) suggested that areas with high reliability but low significance values have time-series that are reliable but do not fit the task/HRF model, and demonstrated this pattern for half of their participants in one ROI. This may also be the case in the present study.

It may be worth pointing out that bilateral primary auditory cortex appears less reliable by this vertex-wise ICC measure. While it is counter-intuitive that a low-level sensory region of cortex would be least reliable, this may be an example of one of the drawbacks of this type of measure – since between-subject variance is an important component of this calculation, areas that are more reliable *across* speakers would tend to have *lower* ICC values, given constant within-subject reliability. Thus it may be more accurate to say that vertices with a high ICC value in this map are the most discriminable areas among a group of subjects.

The final and most direct measure of discriminability was the classifier analysis. This type of analysis has not previously been applied for the purpose of determining the reliability of an individual's neural activation patterns, but it has the added advantage of

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

characterizing the distinctiveness of an individual's brain activation maps. From the near perfect accuracy in identifying a subject correctly from among 75 potential classes given 1 training sample, it is clear that individuals are not only quite reliable but also have distinct activation patterns during speech production – a neural "fingerprint". In fact, the only subject that was mis-classified using the full complement of extracted features is Subject 7, who also had the lowest within-subject ICC value and Dice coefficient, thus demonstrating consistency across measures. Even with only 23 features, accuracy reaches 75%. The *null* maps also demonstrated clear distinctness based on the accuracy of the same classifier trained *null* data and show that the same amount of accuracy as the *speech* maps can be obtained using the reduced feature set. It is also important to mention that the classification method used in the current study is among the simplest of modern machine learning options, and that using only one training map per subject severely reduces the power of the method. Nonetheless, classification accuracy was very high, and we interpret the current result as a lower bound of discriminability of speech activation maps among individuals that might be improved with more sophisticated machine learning algorithms.

4.3. *Speech* vs. *Null* Reliability

Our main goal for including maps of BOLD signal not associated with a particular task was to assess whether reliability found for speech activation could be explained by differences in resting BOLD patterns or underlying neurovasculature. Indeed, we found that the *null* maps showed similar reliability and discriminability than *speech* activation maps. This may indicate that the *null* measure corresponds to underlying anatomical or

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

physiological features for individuals that are reliable (Jann et al., 2015; Shehzad et al., 2009). However the lack of a correlation between *speech* and *null* maps suggests that unique activation patterns during the speech task are not dependent on underlying individual BOLD patterns.

4.4. Reliability for Speech Production across Tasks

One benefit of the results herein is that the speech tasks used to assess reliability, despite their similarity, differed across sessions. This has two important consequences for interpretation of the results. First, the present results do not account for activation variance attributable to inter-task reliability. There may be differences in activation between the studies simply because the speech stimuli were different. Thus they are potentially conservative compared to the results for a consistent speaking task as well as other published fMRI reliability literature. Second, this setup means that the reliability and discriminability discussed applies to the speech production network rather than a particular task. Therefore, the results are more generalizable to other speech production tasks (at least of the same characteristics – reading orthographic representations of mono and bisyllabic words and pseudowords). This is important for assessing the validity of future work mapping individual speech networks derived from speaking tasks that depart from those in the present study.

5. Conclusion

Based on the results of five measures of reliability and discriminability, we conclude that speech activation maps for most neurologically-healthy speakers are generally highly

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reliable, providing justification for single-subject neuroimaging research for speech production. Exceptions were found for subjects who exhibited higher levels of scan-to-scan motion and signal change, reinforcing the widely-held understanding that minimizing motion is crucial for trusting neuroimaging data. Future work analyzing activation patterns from patients with neurogenic speech disorders will be needed to determine whether these individuals are similarly reliable (though extant work examining reliability in patients with stroke (Kimberley, Khandekar, & Borich, 2008) and mild cognitive impairment (Zanto, Pa, & Gazzaley, 2014) are promising), and ultimately whether it is feasible to map the speech production network in individuals and track changes in these patterns across time. This future research would be an important contribution to the growing body of literature characterizing disease progression and neurorehabilitation (Herbet, Maheu, Costi, Lafargue, & Duffau, 2016; Reinkensmeyer et al., 2016), especially for people with speech disorders.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements:

This research was supported by the National Institutes of Health [R01 DC002852, R01 DC007683, and T32 DC013017]. Imaging data were acquired at the Athinoula A. Martinos Center for Biomedical Imaging using resources provided by the Center for Functional Neuroimaging Technologies, P41RR14075 a P41 Regional Resource supported by the National Institute of Biomedical Imaging and Bioengineering, NIH.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Aron, A. R., Gluck, M. A., & Poldrack, R. A. (2006). Long-term test–retest reliability of functional MRI in a classification learning task. *NeuroImage*, *29*(3), 1000–1006. https://doi.org/10.1016/j.neuroimage.2005.08.010

Babajani-Feremi, A., Narayana, S., Rezaie, R., Choudhri, A. F., Fulton, S. P., Boop, F. A., … Papanicolaou, A. C. (2016). Language mapping using high gamma electrocorticography, fMRI, and TMS versus electrocortical stimulation. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *127*(3), 1822–1836. https://doi.org/10.1016/j.clinph.2015.11.017

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*(1), 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x

Bennett, C. M., & Miller, M. B. (2013). fMRI reliability: Influences of task and experimental design. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 690–702. https://doi.org/10.3758/s13415-013-0195-1

Bizzi, A., Blasi, V., Falini, A., Ferroli, P., Cadioli, M., Danesi, U., … Broggi, G. (2008). Presurgical Functional MR Imaging of Language and Motor Functions: Validation with Intraoperative Electrocortical Mapping. *Radiology*, *248*(2), 579–589. https://doi.org/10.1148/radiol.2482071214

Brannen, J. H., Badie, B., Moritz, C. H., Quigley, M., Meyerand, M. E., & Haughton, V. M. (2001). Reliability of functional MR imaging with word-generation tasks for mapping Broca's area. *AJNR. American Journal of Neuroradiology*, *22*(9), 1711–1718.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., & Mehta, M. A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage*, *45*(3), 758–768. https://doi.org/10.1016/j.neuroimage.2008.12.035

Chen, E., & Small, S. (2007). Test–retest reliability in fMRI of language: Group and task effects. *Brain and Language*, *102*(2), 176–185. https://doi.org/10.1016/j.bandl.2006.04.015

D'Esposito, M., Deouell, L. Y., & Gazzaley, A. (2003). Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. *Nature Reviews Neuroscience*, *4*(11), 863–872. https://doi.org/10.1038/nrn1246

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., … Schlaggar, B. L. (2010). Prediction of Individual Brain Maturity Using fMRI. *Science*, *329*(5997), 1358–1361. https://doi.org/10.1126/science.1194144

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature*, *384*(6605), 159–161. https://doi.org/10.1038/384159a0

Duffy, J. R. (2013). *Motor speech disorders: substrates, differential diagnosis, and management* (Third edition). St. Louis, Missouri: Elsevier.

Duncan, K. J., Pattamadilok, C., Knierim, I., & Devlin, J. T. (2009). Consistency and variability in functional localisers. *NeuroImage*, *46*(4), 1018–1026. https://doi.org/10.1016/j.neuroimage.2009.03.014

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical Surface-Based Analysis. *NeuroImage*, *9*(2), 195–207. https://doi.org/10.1006/nimg.1998.0396

Freyer, T., Valerius, G., Kuelz, A.-K., Speck, O., Glauche, V., Hull, M., & Voderholzer, U. (2009). Test–retest reliability of event-related functional MRI in a probabilistic reversal

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

learning task. *Psychiatry Research: Neuroimaging*, *174*(1), 40–46. https://doi.org/10.1016/j.pscychresns.2009.03.003

Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., … Potkin, S. G. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Human Brain Mapping*, *29*(8), 958–972. https://doi.org/10.1002/hbm.20440

Ghosh, S. S., Tourville, J. A., & Guenther, F. H. (2008). A neuroimaging study of premotor lateralization and cerebellar involvement in the production of phonemes and syllables. *Journal of Speech, Language, and Hearing Research*, *51*(5), 1183–1202.

Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., Nieto-Castanon, A., & Guenther, F. H. (2011). fMRI investigation of unexpected somatosensory feedback perturbation during speech. *NeuroImage*, *55*(3), 1324–1338. https://doi.org/10.1016/j.neuroimage.2010.12.065

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., & Pernet, C. R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00245

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I., & Pernet, C. (2013). Single subject fMRI test–retest reliability metrics and confounding factors. *NeuroImage*, *69*, 231–243. https://doi.org/10.1016/j.neuroimage.2012.10.085

Gross, W. L., & Binder, J. R. (2014). Alternative thresholding methods for fMRI data optimized for surgical planning. *NeuroImage*, *84*, 554–561. https://doi.org/10.1016/j.neuroimage.2013.08.066

Guenther, F. H. (2015). *Neural control of speech*. Cambridge, MA: The MIT Press.

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Harrington, G. S., Buonocore, M. H., & Farias, S. T. (2006). Intrasubject reproducibility of functional MR imaging activation in language tasks. *AJNR. American Journal of Neuroradiology*, *27*(4), 938–944.

Herbet, G., Maheu, M., Costi, E., Lafargue, G., & Duffau, H. (2016). Mapping neuroplastic potential in brain-damaged patients. *Brain*, *139*(3), 829–844. https://doi.org/10.1093/brain/awv394

Hillis, A. E., Work, M., Barker, P. B., Jacobs, M. A., Breese, E. L., & Maurer, K. (2004). Re-examining the brain regions crucial for orchestrating speech articulation. *Brain*, *127*(7), 1479–1487. https://doi.org/10.1093/brain/awh172

Jann, K., Gee, D. G., Kilroy, E., Schwab, S., Smith, R. X., Cannon, T. D., & Wang, D. J. J. (2015). Functional connectivity in BOLD and CBF data: Similarity and reliability of resting brain networks. *NeuroImage*, *106*, 111–122. https://doi.org/10.1016/j.neuroimage.2014.11.028

Kimberley, T. J., Khandekar, G., & Borich, M. (2008). fMRI reliability in subjects with stroke. *Experimental Brain Research*, *186*(1), 183–190. https://doi.org/10.1007/s00221-007-1221-8

Kiran, S., Ansaldo, A., Bastiaanse, R., Cherney, L. R., Howard, D., Faroqi-Shah, Y., … Thompson, C. K. (2013). Neuroimaging in aphasia treatment research: standards for establishing the effects of treatment. *NeuroImage*, *76*, 428–435. https://doi.org/10.1016/j.neuroimage.2012.10.011

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Leech, R., & Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain*, *137*(1), 12–32. https://doi.org/10.1093/brain/awt162

Liu, T. T., Nalci, A., & Falahpour, M. (2017). The global signal in fMRI: Nuisance or Information? *NeuroImage*, *150*, 213–229. https://doi.org/10.1016/j.neuroimage.2017.02.036

Maldjian, J. A., Laurienti, P. J., Driskill, L., & Burdette, J. H. (2002). Multiple reproducibility indices for evaluation of cognitive functional MR imaging paradigms. *AJNR. American Journal of Neuroradiology*, *23*(6), 1030–1037.

Mayer, A. R., Xu, J., Paré-Blagoev, J., & Posse, S. (2006). Reproducibility of activation in Broca's area during covert generation of single words at high field: A single trial FMRI study at 4 T. *NeuroImage*, *32*(1), 129–137. https://doi.org/10.1016/j.neuroimage.2006.03.021

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Meltzer, J. A., Postman-Caucheteux, W. A., McArdle, J. J., & Braun, A. R. (2009). Strategies for longitudinal neuroimaging studies of overt language production. *NeuroImage*, *47*(2), 745–755. https://doi.org/10.1016/j.neuroimage.2009.04.089

Moser, D., Basilakos, A., Fillmore, P., & Fridriksson, J. (2016). Brain damage associated with apraxia of speech: evidence from case studies. *Neurocase*, *22*(4), 346–356. https://doi.org/10.1080/13554794.2016.1172645

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Niziolek, C. A., & Guenther, F. H. (2013). Vowel Category Boundaries Enhance Cortical and Behavioral Responses to Speech Feedback Alterations. *Journal of Neuroscience*, *33*(29), 12090–12098. https://doi.org/10.1523/JNEUROSCI.1008-13.2013

Otzenberger, H., Gounot, D., Marrer, C., Namer, I. J., & Metz-Lutz, M.-N. (2005). Reliability of individual functional MRI brain mapping of language. *Neuropsychology*, *19*(4), 484–493. https://doi.org/10.1037/0894-4105.19.4.484

Peeva, M. G., Guenther, F. H., Tourville, J. A., Nieto-Castanon, A., Anton, J.-L., Nazarian, B., & Alario, F.-X. (2010). Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. *NeuroImage*, *50*(2), 626–638. https://doi.org/10.1016/j.neuroimage.2009.12.065

Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R. J. A., Kahn, R. S., & Ramsey, N. F. (2007). Test–retest reliability of fMRI activation during prosaccades and antisaccades. *NeuroImage*, *36*(3), 532–542. https://doi.org/10.1016/j.neuroimage.2007.03.061

Raschle, N. M., Zuk, J., & Gaab, N. (2012). Functional characteristics of developmental dyslexia in left-hemispheric posterior brain regions predate reading onset. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(6), 2156–2161. https://doi.org/10.1073/pnas.1107721109

Rau, S., Fesl, G., Bruhns, P., Havel, P., Braun, B., Tonn, J.-C., & Ilmberger, J. (2007). Reproducibility of Activations in Broca Area with Two Language Tasks: A Functional

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MR Imaging Study. *American Journal of Neuroradiology*, *28*(7), 1346–1353. https://doi.org/10.3174/ajnr.A0581

Reinkensmeyer, D. J., Burdet, E., Casadio, M., Krakauer, J. W., Kwakkel, G., Lang, C. E., … Schweighofer, N. (2016). Computational neurorehabilitation: modeling plasticity and learning to predict recovery. *Journal of NeuroEngineering and Rehabilitation*, *13*(1). https://doi.org/10.1186/s12984-016-0148-3

Rutten, G. J. M., Ramsey, N. F., van Rijen, P. C., & van Veelen, C. W. M. (2002). Reproducibility of fMRI-Determined Language Lateralization in Individual Subjects. *Brain and Language*, *80*(3), 421–437. https://doi.org/10.1006/brln.2001.2600

Sato, M., Vilain, C., Lamalle, L., & Grabski, K. (2015). Adaptive Coding of Orofacial and Speech Actions in Motor and Somatosensory Spaces with and without Overt Motor Behavior. *Journal of Cognitive Neuroscience*, *27*(2), 334–351. https://doi.org/10.1162/jocn_a_00711

Shehzad, Z., Kelly, A. M. C., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., … Milham, M. P. (2009). The Resting Brain: Unconstrained yet Reliable. *Cerebral Cortex*, *19*(10), 2209–2229. https://doi.org/10.1093/cercor/bhn256

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., … McGonigle, D. J. (2005). Variability in fMRI: A re-examination of inter-session differences. *Human Brain Mapping*, *24*(3), 248–257. https://doi.org/10.1002/hbm.20080

Sundermann, B., Herr, D., Schwindt, W., & Pfleiderer, B. (2014). Multivariate classification of blood oxygen level-dependent FMRI data with diagnostic intention: a clinical

bioRxiv preprint doi: https://doi.org/10.1101/807925; this version posted October 17, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

perspective. *AJNR. American Journal of Neuroradiology*, *35*(5), 848–855. https://doi.org/10.3174/ajnr.A3713

Tourville, J.A., & Guenther, F. H. (2012). Automatic cortical labeling system for neuroimaging of normal and disordered speech. *42nd Annual Meeting for the Society for Neuroscience*.

Tourville, Jason A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, *39*(3), 1429–1443. https://doi.org/10.1016/j.neuroimage.2007.09.054

Voyvodic, J. T. (2012). Reproducibility of single-subject fMRI language mapping with AMPLE normalization. *Journal of Magnetic Resonance Imaging*, *36*(3), 569–580. https://doi.org/10.1002/jmri.23686

Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). *Conn* : A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. *Brain Connectivity*, *2*(3), 125–141. https://doi.org/10.1089/brain.2012.0073

Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., & Eriksson, D. K. (2017). Validity and reliability of four language mapping paradigms. *NeuroImage: Clinical*, *16*, 399–408. https://doi.org/10.1016/j.nicl.2016.03.015

Zanto, T. P., Pa, J., & Gazzaley, A. (2014). Reliability measures of functional magnetic resonance imaging in a longitudinal evaluation of mild cognitive impairment. *NeuroImage*, *84*, 443–452. https://doi.org/10.1016/j.neuroimage.2013.08.063