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

Email addresses: saulf@bu.edu (S.A. Frankford); alfnie@gmail.com (A. Nieto-Castañón); jtour@bu.edu (J.A. Tourville); guenther@bu.edu (F.H. Guenther)

Declarations of Interest: None

Send correspondence to:

Saul A. Frankford
Boston University
Department of Speech, Language and Hearing Sciences
677 Beacon St.
Boston, MA 02215
saulf@bu.edu
(215) 510-7179

Abstract

Speech neuroimaging research targeting individual speakers could help elucidate differences that may be crucial to understanding speech disorders. However, this research necessitates reliable brain activation across multiple speech production sessions. In the present study, we evaluated the reliability of speech-related brain activity measured by functional magnetic resonance imaging data from twenty neuro-typical subjects who participated in two experiments involving reading aloud simple speech stimuli. Using traditional methods like the Dice and intraclass correlation coefficients, we found that most individuals displayed moderate to high reliability. We also 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

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 4,500 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 neurotypical participants, in many cases compared to similar-sized cohorts of patients with neurological conditions that impact speech or language function. Collectively, these studies have revealed a network of brain areas that are commonly active during speech production (Guenther, 2016; Price, 2012). When brain responses are compared between groups, however, the results are often less consistent (e.g., Connelly et al., 2018 vs. Chang et al., 2009). This could result from the relatively small sample sizes of typical fMRI study designs lacking sufficient power, a shortcoming that is being addressed in more recent studies with larger samples sizes and data pooling (Brown et al., 2005; Costafreda, 2009; Turkeltaub et al., 2002), i.e., measuring across larger groups.

Larger groups, however, cannot address another factor that is becoming more apparent to those mapping the functional components of the speech production network: high between-subjects variability in the location and level of speech-related BOLD responses. Attempts to localize the locus of "crucial" neural damage in acquired apraxia of speech (AOS), for instance, have reported a variety of locations (Dronkers, 1996; Hillis et

[1] Derived from a search of articles on pubmed.com on February 25, 2020 containing the terms "fMRI" or "functional magnetic resonance imaging" and "speech" or "language" in their title or abstract.

al., 2004; Moser et al., 2016). Moreover, there is tremendous variability in the location and extent of stroke-related damage to neural tissue across *individuals*. This individual variability found in AOS and other speech network disturbances (e.g., stuttering, Wymbs et al., 2013) can mask group differences in fMRI analyses, and make it difficult to map the neural locus (or loci) of a given disorder.

An alternative approach for studying speech disorders is to use subject-specific study designs that are unaffected by between-subjects variability. A number of studies have demonstrated the utility of single-subject fMRI study designs or encouraged its future use for a range of purposes. These include mapping 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) improving diagnosis of disorders (Raschle et al., 2012; Sundermann et al., 2014), and determining whether neural plasticity following stroke can predict outcomes (Chen & Small, 2007; Kiran et al., 2013; Meltzer et al., 2009). In the speech domain, single-subject approaches have been used to evaluate responses to treatment in AOS (e.g., Farias et al., 2014), but these could be expanded to tracking natural neural organization changes over time in developmental speech disorders like stuttering. Due to the individuality of the presentation of these disorders, subject-specific approaches could provide more meaningful measures of change not captured in group average analyses.

However, the suitability of subject-specific studies of speech and language processes, depends heavily on the reliability of speech-related activity in individual brains. 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.

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 have used a covert speech task (Brannen et al., 2001; Harrington et al., 2006; Maldjian et al., 2002; Mayer et al., 2006; Otzenberger et al., 2005; Rutten et al., 2002) or have focused on a limited set of 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). Four recent studies (Gorgolewski et al., 2013; Nettekoven et al., 2018; Paek et al., 2019; Wilson et al., 2017) have assessed reliability in neurologically normal participants across the cortex during overt word production. These studies report moderate to high levels of reliability and each provides unique insight into the factors that impact test-retest reliability, especially pertaining to older adults and clinical populations.

Our aim in the present study was to determine whether such reliability is robust in the speech production network across different speaking tasks and interscan intervals. To do this, we performed a retrospective analysis of participants who had taken part in more than one fMRI study of speech production in our lab. This had the advantage of assessing the reliability of general speech network activation patterns in an individual rather than the reliability of a specific task to allow for greater generalization of the results herein. Compared to previous work, we included studies with stimuli that limited higher-level linguistic processing. Doing so allowed us to assess the reliability of neural activity specific

to speech motor control processes. Finally, since these datasets were collected for basic research purposes in healthy individuals, they were composed of much longer sessions which may improve the reliability of an individual's speech network activity.

We used the Dice coefficient to measure the spatial overlap of active brain regions within individuals across multiple speech production studies. This easily interpretable measure can be compared to numerous previous studies of fMRI reliability (Bennett & Miller, 2010). For a more thorough reliability measure that accounts for both the location and relative scale of activity 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 to a between-subjects standard. We therefore 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). This measure estimated the reliability and discriminability of activation across the entire brain at a vertex level. Finally, we directly tested whether an individual speaker's neural activation patterns during speech in one study could predict activation in a second study using a machine learning classifier. Reliability measures were compared to two benchmarks: a chance-level baseline derived from random data maps, and a residual signal map derived from anatomy-related information in the BOLD signal that we would expect to have high reliability.

2. Materials and Methods

2.1. Participants

Our dataset comprises seventy-five individuals who previously participated in fMRI studies of speech production in the SpeechLab 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 or three 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 2.5.4. Subject Classifier). All participants were right-handed native speakers of American English and reported normal or corrected-to-normal vision as well as no history 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 formed by two or more 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. Details of the four studies from which repeated measures were taken (CCRS, FRS, APE, and PBB) are described here. More detailed

information on the other studies (OP, SylSeq, and CAT) is provided in the publications listed in Table 1.

The CCRS and FRS experiments were block-design fMRI studies in which subjects produced sequences of pseudowords during continuous scanning. Both studies included multiple speech conditions and a baseline condition. During speech trials, subjects simultaneously viewed an orthographic representation and heard a recording of the pseudoword to be produced. A white cross replacing the orthographic representation cued the subject to produce the pseudoword. On baseline trials, subjects saw a series of asterisks on the screen rather than orthographic stimulus and rested quietly. Functional runs were organized into blocks of 6 trials of the same condition with a 3 s pause between blocks. Pseudowords and conditions were randomized within runs.

Sequences in the CCRS study comprised pairs of two-syllable pseudowords that varied in the number of unique phonemes, consonant clusters and syllables in the sequence. The conditions were: exact repetition (e.g., 'GROI SLEE, GROI SLEE'); same phonemes and consonant clusters, different syllables (e.g. 'GROI SLEE, GREE SLOI'); and different phonemes, consonant clusters, and syllables (e.g. 'KWAI BLA, SMOO KROI'). Each trial lasted 2.5 s. Runs consisted of fifteen blocks, and lasted approximately 5 min. Each subject completed 7 runs that optimally allowed for approximately 21 blocks per condition per subject. In total, 120 fMRI volumes were acquired continuously during each run.

Sequences in the FRS study were pairs of monosyllabic pseudowords that varied in the number of unique phonemes, syllables, and syllabic frames (see MacNeilage, 1998). The conditions were: exact repetition (e.g. 'TWAI, TWAI'); same frames, different phonemes and syllables (e.g. 'FAS REEN'); same phonemes, different frames and syllables (e.g. 'RAUD

DRAU'); and different frames, phonemes, and syllables (e.g. 'DEEF GLAI'). Each trial lasted 2 seconds. Runs consisted of eighteen blocks and lasted approximately 4.5 min. Each pseudoword or pseudoword pair was maximally used once per block and in 2-3 blocks throughout the experiment to maintain novelty. Each subject completed 6 runs that optimally allowed for approximately 27 blocks per condition per subject. In total, 108 fMRI volumes were acquired continuously during a run.

The APE (Tourville et al., 2008) and PBB studies (Golfinopoulos et al., 2011), used a sparse fMRI acquisition design that allowed subjects to produce speech during silent intervals between fMRI volume acquisitions. In both experiments, subjects were instructed to read aloud the speech stimulus presented orthographically at the onset of each trial or to remain silent if a control stimulus (the letter string 'yyy') was presented. Stimuli in the APE study consisted of 8 /CɛC/ words (e.g., beck, bet, debt). Stimuli remained onscreen for 2 s. An experimental run consisted of 64 speech trials (8 presentations of each word) and 16 control trials (Tourville et al., 2008). On 25% of speech trials, the first formant (F1) of the subject's speech was altered before being fed back to the subject. Trial order was randomly permuted within each run such that consecutive presentation of the same stimulus and consecutive F1 shifts in the same direction were prohibited. Subjects performed 3 or 4 functional runs. *Only speech trials with normal feedback* and baseline trials were included in the present study.

Speech stimuli in the PBB study (Golfinopoulos et al., 2011) consisted of eight pseudowords that required a jaw closure after producing an initial vowel (e.g., /au/, /ani/, /ati/). Stimuli remained onscreen for 3 s. Each experimental run consisted of 56 speech trials (seven presentations of each pseudoword) and 16 baseline trials. On one seventh of

all speech trials and half of all baseline trials, jaw closure was restricted by the rapid inflation of a small balloon positioned between the subjects' upper and lower molars. Trial order was randomly permuted within each run such that consecutive perturbation trials were prohibited. Subjects included in the present analysis completed between three and five runs. *No perturbation trials* were included in the present analysis.

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

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

Subject	Study 1			Study 2			Days Between Studies
	ID	Speech Trials	Baseline Trials	ID	Speech Trials	Baseline Trials	
1	CCRS	378	126	FRS	258	66	6
2	CCRS	378	126	FRS	258	66	14
3	CCRS	324	108	FRS	258	66	52
4	CCRS	378	126	FRS	258	66	7
5	CCRS	324	108	FRS	258	66	6
6	CCRS	378	126	FRS	258	66	20
7	CCRS	324	108	FRS	258	66	7
8	CCRS	378	126	FRS	258	66	13

9	CCRS	378	126	FRS	258	66	19
10	CCRS	378	126	FRS	258	66	12
11	CCRS	324	108	FRS	258	66	7
12	CCRS	324	108	FRS	258	66	7
13	CCRS	378	126	FRS	258	66	7
14	CCRS	324	108	FRS	258	66	7
15	APE	191	64	PBB	192	32	75
16	APE	191	64	PBB	144	24	163
17	APE	191	64	PBB	192	32	196
18	APE	187	63	PBB	240	40	21
19	APE	192	64	PBB	192	32	70
20	APE	143	48	PBB	240	40	28

Table 2. Studies in which each test subject participated, total number of trials, and time between studies. Study identification codes refer to abbreviations in the 'Study' column of Table 1.

2.3. 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 CCRS and FRS, data were acquired using a 3 Tesla Siemens Trio Tim scanner with a 32-channel head coil. For each subject, a whole-brain high-resolution T1-weighted MPRAGE volume was acquired (voxel size: 1 mm^3, 256 sagittal images, TR: 2530 ms, TE: 3.44 ms). T2*-weighted volumes consisting of 41 gradient echo – echo planar axial images (in plane resolution: 3.1 mm, slice thickness: 3 mm, gap: 25%, TR: 2.5 s, TA: 2.5 s, TE: 20 ms) were collected continuously during functional runs.

For APE and PBB, a high-resolution T1-weighted anatomical volume (128 slices in the sagittal plane, slice thickness: 1.33 mm, in-plane resolution: 1 mm^2, TR: 2530 ms, TE: 3.3 ms) was obtained for each subject prior to functional imaging. Functional volumes consisted of 32 gradient echo - echo planar axial images (in plane resolution: 3.125 mm^2,

slice thickness: 5 mm, TR: 2000 ms, TE: 30 ms). A sparse sampling (Hall et al., 1999) clustered volume acquisition method, consisting of silent intervals between consecutive volume acquisitions, was used. Two consecutive volumes (each volume acquisition taking 2 s) were acquired 5 s after the onset of each trial.

See Peeva et al. (2010), Ghosh, Tourville, & Guenther (2008), and Niziolek & Guenther (2013) for acquisition parameters for the SylSeq, OP, and CAT studies, respectively (refer to Table 1 for study codes).

2.4. 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 first functional image, and coregistered to their structural image using SPM12's inter-modality coregistration procedure with a normalized mutual information cost function (Collignon et al., 1995; Studholme et al., 1998). 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 APE and PBB, the BOLD response for each event was modeled using a single-bin finite impulse response (FIR) basis function spanning the time of acquisition of the two consecutive volumes. 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 session effect (the intercept for that run). See Peeva et al. (2010), Ghosh, Tourville, & Guenther (2008), and Niziolek & Guenther (2013)

for first-level design details in the other studies. Functional data were also censored (Power et al., 2014) by including additional regressors for all studies to remove the effects of volumes with excessive motion and global signal change, as identified using ART (https://www.nitrc.org/projects /artifact_detect/) with a scan-to-scan motion threshold of 0.9 mm and a scan-to-scan signal intensity threshold of 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). 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 using iterative diffusion smoothing with 40 diffusion steps (equivalent to a 8 mm full-width half maximum smoothing kernel, Hagler et al., 2006). This level of smoothing has previously been shown to optimize reliability of task-related BOLD response data in individuals (Caceres et al., 2009).

In addition to the above *speech* maps, we computed two other sets of maps for comparison purposes. The first was *random* maps, representing randomly generated data

with similar spatial properties, and processed in exactly the same way as the *speech* maps. We expected these maps to show minimal reliability (chance-level). Reliability measures derived from *random* maps served as a baseline reference, and to eliminate the possibility that our preprocessing and estimation procedure would artifactually introduce unexpected biases in reliability metrics. The second was *null* maps, representing anatomical information about each subject like tissue morphology and neurovasculature present in the average BOLD signal, and, again, were processed in exactly the same way as the *speech* maps. We expected these maps to show high reliability, as anatomical information is expected to vary minimally over the time spans considered in this study. Reliability measures derived from *null* maps served as references for comparison purposes, and to explore the possibility that reliability of speech-related functional activation may be influenced by, or related to, reliability of anatomical features.

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 (mean of 0 and a standard deviation of 1) and smoothing the data to the same degree as the *speech* maps. To obtain maps of average MRI signal (*null* maps) that is not affected by task effects, estimates of the constant regression term of each run were averaged for each subject in each study. These maps represent the average T2* signal after the effects of speech, baseline, motion, and outliers have been removed. Similar to the *speech* maps, they were then projected to each individual's structural surface. Because there is individual variability in the T2* signal across the cortex, these maps represent individual features of a subject's cortical anatomy

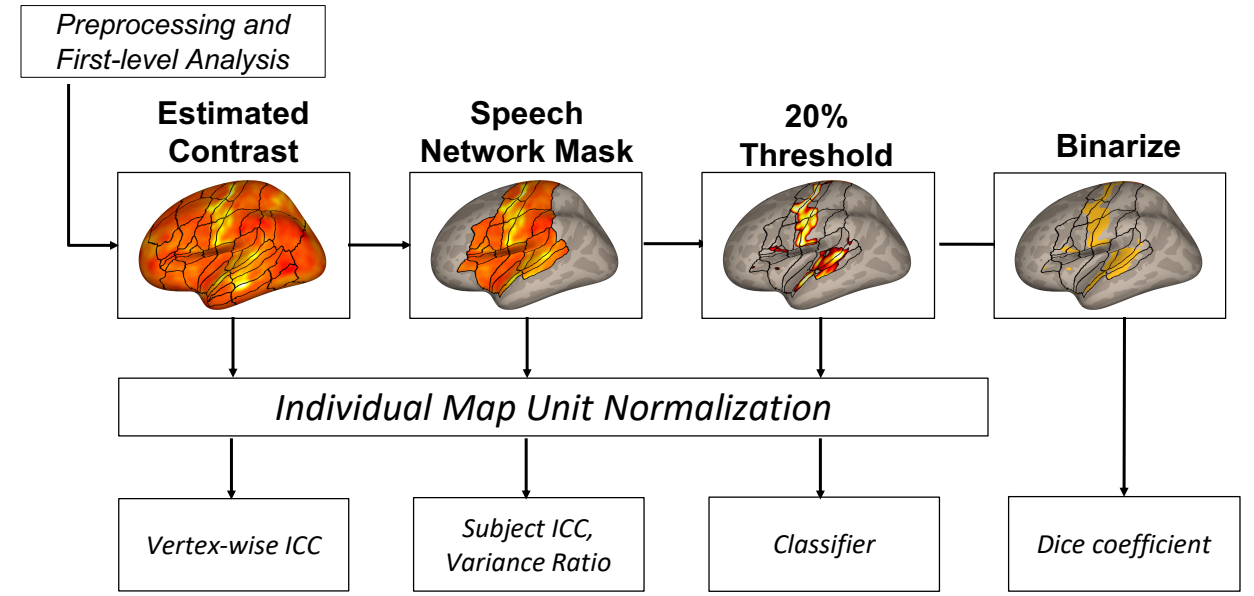

Figure 1. Thresholding pipeline map for each of the reliability analyses. After preprocessing and estimation of first-level condition effects, the *speech*, *null*, and *random* maps were calculated, and submitted to the vertex-wise ICC analysis. A speech network mask was applied, so that only vertices inside this mask were used for the single-subject ICC and variance ratio measures. Next, the 20% of vertices with the highest activation levels were kept for the classifier analysis. Finally, these thresholded maps were binarized for the Dice coefficient analysis. Prior to calculating reliability measures (except the Dice coefficient), 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.

2.5. Reliability Measures

We used two measures to quantify individual-subject activation reliability across different sessions in individuals (while sessions come from two separate studies, for clarity the term *session* will be used going forward to refer to a data collection time point): the Dice coefficient and a single-subject intraclass correlation coefficient. Two further measures were used to examine sampled-normed reliability: a vertex-wise intraclass

correlation coefficient, and a machine-learning classifier. Each of these measures was applied to the *speech, random,* and *null* maps.

2.5.1. Single-subject Spatial Overlap

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 of individual maps and their average size and yields values between 0 (no overlap) and 1 (complete overlap). A strength of this measure is that it is 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 these maps are thresholded (Duncan et al., 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 approximately 35% of cortex (see Figure 1; Tourville & Guenther, 2012). Activation maps

were then thresholded to retain only the highest 20% of surface vertices within the masked area (approximately 7% of total cortex; see Figure 2 for examples of these thresholded maps). Finally, this map was binarized (active voxels = 1, all other voxels = 0).

2.5.2. Single-subject ICC

To obtain a measure of reliability that was not threshold-dependent and took into account the level of activation at each vertex, 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 there is no correlation across all vertices, while a value of 1 signifies perfect correlation across all vertices. 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 sessions, 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:

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

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). Following the convention of Koo and Li (2016), ICC values below 0.5 indicate poor reliability, between 0.5 and 0.75, moderate reliability, between 0.75 and 0.9, good reliability, and above 0.9, excellent reliability.

In addition, to determine whether reliability in individual subjects across sessions was higher than that across the sample, we also computed a between-subjects ICC analysis. This was accomplished by averaging each individual's *speech* maps across sessions, and estimating the same ICC defined in *Eq.* 2 and *Eq*. 3. Thus, the *s* term estimated the between-subjects effect rather than the between-session effect.

For this analysis, activation maps were masked with the same speech production network mask described for the overlap analysis but no activation threshold was applied, 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 unit normalized within each map prior to each analysis by dividing the value at each vertex by the Euclidian norm of all the vertices in the map.

2.5.3. 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 2.5.2, 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 & Li's (2016) convention to threshold the resulting ICC map, keeping only those vertices with good or excellent reliability (values greater than or equal to 0.75). Because this measure is calculated with respect to the sample variance, it also provides a measure of discriminability – greater differences between subjects leads to higher values. We applied this analysis to all cortical vertices (without a speech network mask) in order to compare the discriminability of vertices within speech-related areas to those not usually associated with speech. As with the previously described analyses, activation values in each map were unit normalized.

2.5.4. 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 trained a nearest-neighbor subject classifier to identify individual subjects from their functional maps, in order to assess both the reliability and discriminability of *speech* and *null* maps (separately) for individual subjects. First, a session map from the 20 subjects who were scanned twice was

set aside as the testing map. A randomly selected single-session activation map from all 75 subjects was then used as the training set (excluding the testing map). The training set data were converted to a set of activation vectors, demeaned, and whitened using the observed between-subjects covariance within the training set (Strang, 1998). The nearest-neighbor classifier then selected the subject within the training set that had the smallest Euclidean distance to the test map. This was repeated for all 40 activation test maps in the dataset (2 maps from each of the 20 subjects with repeated measures) and a percent accuracy score was obtained. This whole procedure was repeated 100 times, each time selecting different sets of random single-session activation maps for training, and the mean accuracy value across these repetitions was taken as the classifier predictive accuracy. Bias-corrected and accelerated (BCa) bootstrapping 95% confidence intervals (Efron, 1987) for accuracy were estimated with 1000 resamples.

For this analysis, we used maps that were masked, thresholded, and unit normalized (see Figure 2B for examples). This meant that subjects were classified by the patterns of relative activation within the most active vertices. We also ran this same classifier on *random* maps (described in section 2.4) to provide an estimate of the accuracy expected based on chance, given the thresholding steps and type of classifier used.

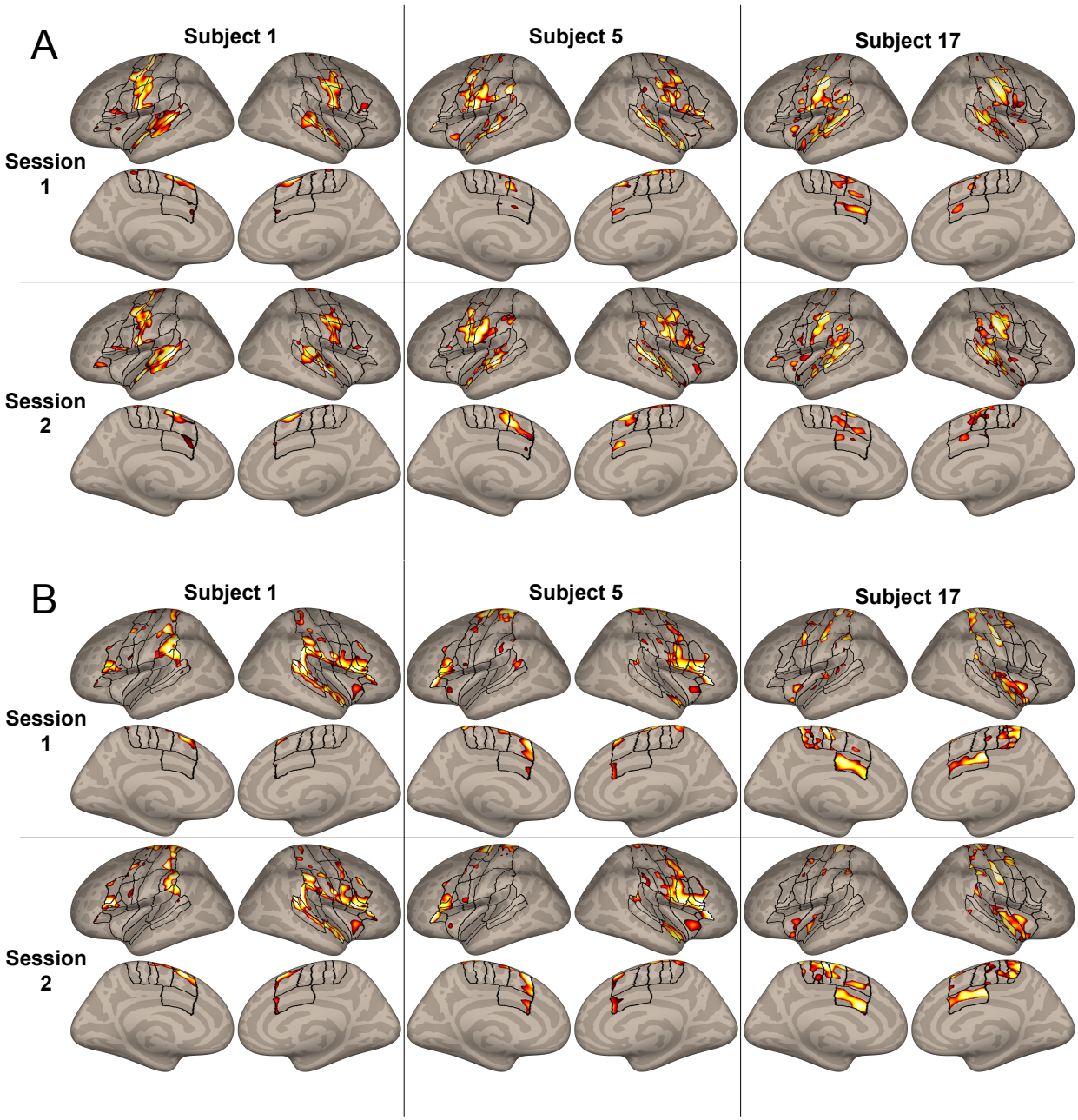

Figure 2. A. Masked and thresholded *speech* maps for three example subjects in both sessions. Outlines of regions of interest covering the masked speech network previously described in Tourville & Guenther (2012) are included for reference. B. Masked and thresholded *null* maps for the same subjects. 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.

2.6. Group-level Statistical Analyses

Dice coefficient and single-subject ICC reliability measures from the *speech, null, and random* maps, which were not assumed to follow a normal distribution, were compared using Wilcoxon Signed-Ranks tests. For the single-subject ICC analysis, we also compared individual ICC values with the between-subjects ICC group measure. In addition, we 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 corresponded with high reliability in the *null* condition).

2.7. 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

The Dice coefficient for each subject's thresholded *speech* maps compared between scanning sessions can be found in Figure 3A. 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 sessions 1 and 2 are also shown in Figure 3A. On average, individuals had 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 0.205 (SD: 0.016; this is 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). However, both conditions were significantly different from the *random* maps (z = 3.92, p < 0.001 for both). Further, there was no correlation between Dice coefficients for *speech* and *null* maps (Spearman's r = 0.098, p = 0.681).

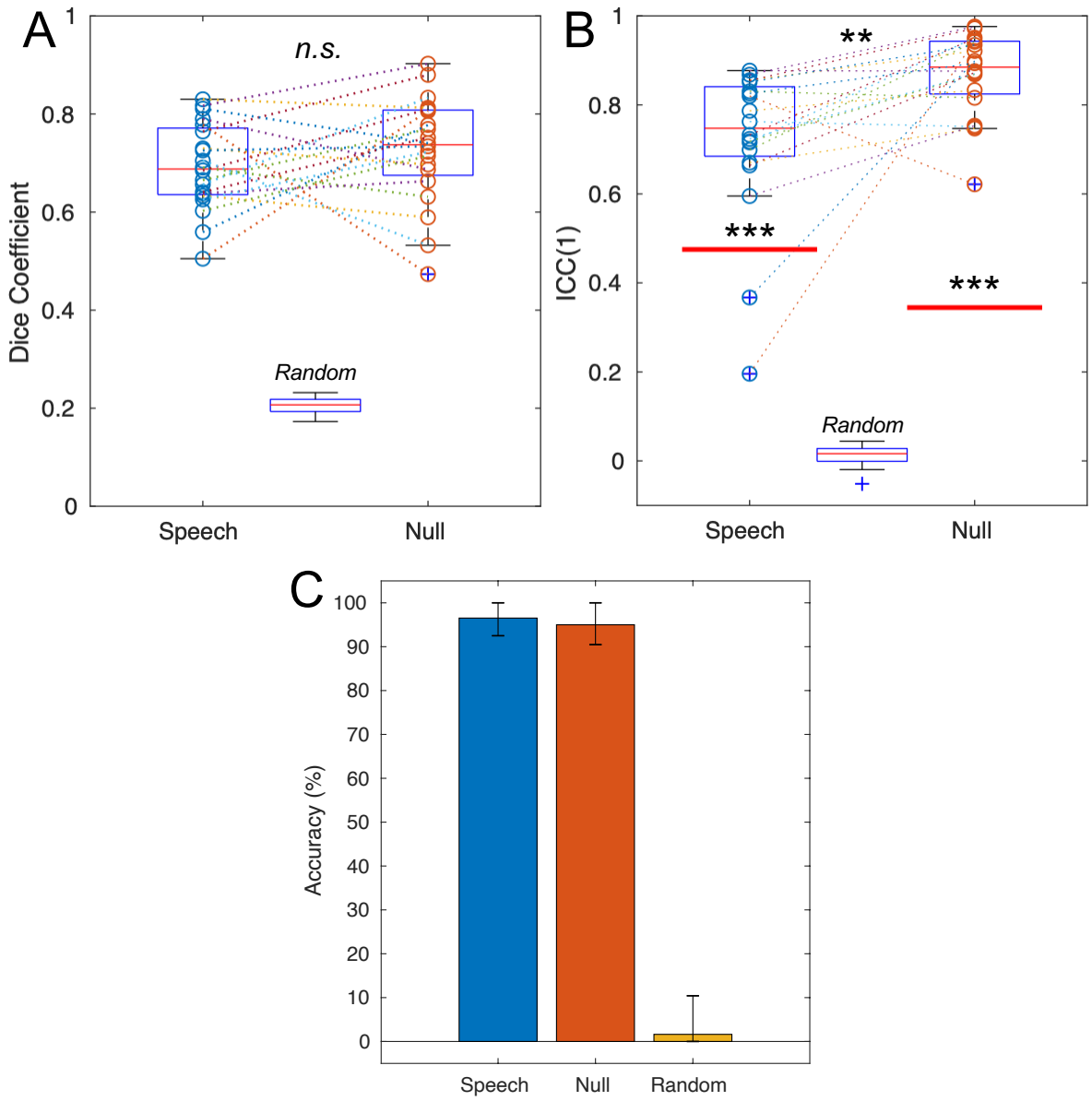

Figure 3. Comparison of reliability measures across conditions. A. Dice coefficient values. Values for individual subjects are shown as circles in each condition, and dashed lines connect results from individual subjects across conditions. For each condition: thin red line = median; blue box = interquartile 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; blue crosses signify outliers – values that fall outside the black lines. B. Single-subject intraclass correlation coefficients. Circles and box plots represent the same information as in A. The thick red lines show the between-subjects intraclass correlation values. Asterisks in line with each condition show comparisons between the distribution of individual points and the Between-Subjects ICC. C. Classifier accuracy. Error bars denote the bias-corrected and accelerated bootstrapping 95% confidence intervals (see section 2.5.4 for details). *n.s.:* non-significant at alpha = 0.05; **: $p < 0.01$; ***: $p < 0.001$.

3.2. Single-subject ICC

The distribution of single-subject *speech* ICC values across sessions can be found in Figure 3B. Subjects exhibited poor (0.196) to good (0.868) reliability according to the convention of 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-Rank test shows that the median of the 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 *random* maps yielded a mean ICC of 0.013 (SD: 0.025). Within-subject ICCs for the *null* maps were significantly greater than the ICCs for the *speech* maps ($z=3.17$, $p=0.002$), and both were significantly greater than random maps ($z = 3.92$, $p < 0.001$ for both). Similar to the Dice coefficient, there was no significant correlation between ICC values in the *speech* and *null* conditions (Spearman's $r = 0.173$, $p = 0.464$).

3.3. Vertex-wise Reliability

The vertex-wise ICC map for the *speech* data thresholded at 0.75 can be found in Figure 4. 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, reflecting good or excellent reliability; Koo & Li, 2016) appeared in areas commonly activated during speech production including, on the lateral surface: bilateral motor and somatosensory cortex, bilateral secondary 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 discriminability 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.

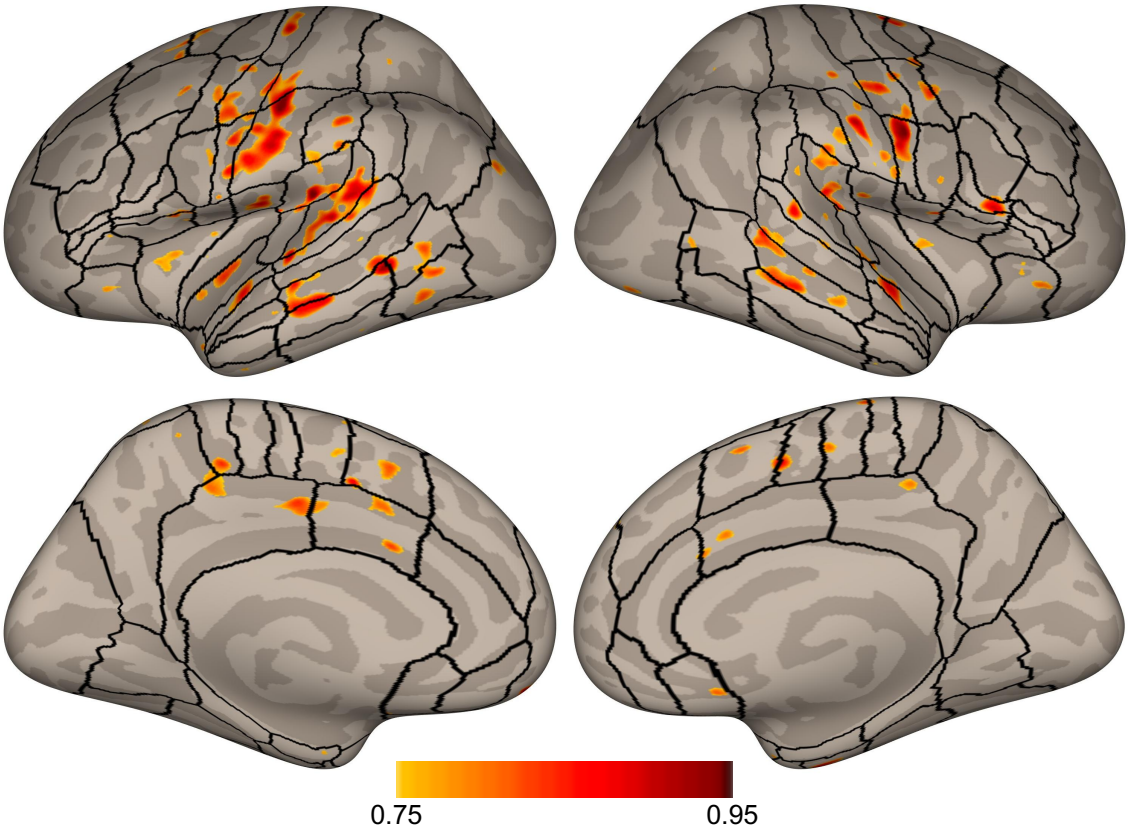

Figure 4. Vertex-wise ICC values for the *speech* activation maps thresholded at 0.75. Regions of interest previously described in Tourville & Guenther (2012) are included for reference.

3.4. Subject Classifier

Accuracy of the subject classifier for the *speech* and *null* maps is displayed in Figure 3C. For the *speech* maps, classifier accuracy for untrained test data was 96.52% (BCa bootstrapping 95% confidence interval: 92.5% – 100%). Similarly, the accuracy of this classification method reached 95% for the *null* activation maps (BCa bootstrapping 95% confidence interval: 90.48% – 100%). 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 each subject has and the thresholding pipeline). These results show that for random data, the classifier accuracy was 1.63% (BCa bootstrapping 95% confidence interval: 0% – 10.42%).

4. Discussion

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

4.1. Subject-specific 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, but not all, individuals showed moderate to high amounts of reliability across tasks and timepoints. The Dice values found in this study were generally larger than those found in previous overt expressive language studies (Gorgolewski et al., 2013; Nettekoven et al., 2018; Paek et al., 2019; Wilson et al., 2017). There are several possibilities as to why this was the case. First, the high number of trials for each subject included herein likely increased power which could have improved the robustness of the activation patterns. Indeed, in Gorgolewski et al. (2013), participants had 36 speech trials and 36 baseline compared to an average of 271.9 speech trials in the present analysis (range: 143 – 378) and 78.7 baseline trials (range 24 – 126), likely leading to differences in power as shown previously (Friedman et al., 2008). Paek et al. (2019), on the other hand, included 60 speech trials and 60 baseline trials which may have contributed to its relatively higher Dice coefficients. It is important to note that sessions in these studies were intentionally shortened to accommodate clinical populations, future studies will need to determine how to balance the duel needs of maximal power with minimal scan time.

In addition, as previously discussed, the Dice coefficient is inherently tied to the thresholding scheme used. Gorgolewski et al. (2013), Nettekoven et al., (2018), and Paek et al. (2019) used statistically thresholded maps rather than effect size maps with a percent threshold. Statistically thresholded maps such as these can be strongly affected by multiple factors including noise from head motion and total scan time (Bennett & Miller, 2010; Gross & Binder, 2014). Furthermore, 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. Finally, the older cohorts used in Gorgolewski et al. (2013; age range: 50-58 years), Wilson et al. (2017; age range: 70-76 years), and Paek et al., (2019; age range: 64-83 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 et al., 2003). Future work will have to confirm the relationship between age and speech activation reliability.

The single-subject ICC applied 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 reliability. One noteworthy difference between this measure and the Dice coefficient was

significantly higher ICC for the *null* maps compared to that of the *speech* maps with some subjects attaining near perfect between-session *null* map correspondence. This demonstrates that once all task and motion parameters are accounted for, the underlying signal patterns that reflect individual anatomy maintain 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.

There were, however, two participants (Subject 6 and Subject 7) whose within-subjects ICC scores for the *speech* maps were less than the between-subjects ICC estimate. In both cases, the median beta value across vertices for one of the two scanning sessions (the CCRS study session) was more negative than that of 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 had the two highest scan-to-scan global signal change values for the other study (FRS). Changes in global signal are often artifacts associated with subject motion, (although other physiological sources contribute to this measure; see Liu, Nalci, & Falahpour, 2017) which was found to be detrimental to reliability measures in previous work (Gorgolewski et al., 2013). 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 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. Indeed, the distribution of activation values was shifted between

the two sessions to a greater extent for these subjects than for others. We attempted to correct for this by unit-normalizing vertex values for each subject in each study, but this is not a perfect method. Thus, both data quality and methodological choices likely drove down their reliability scores. Minimizing motion will therefore be especially important for future subject-specific analyses.

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. Population-normed Reliability

The other two measures we calculated assessed population-normed reliability by comparing response variability within subjects (across sessions) to variability between subjects. These measures assess individual reliability relative to the sample, but additionally characterize 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 discriminability. In a review of fMRI studies of speech and language processing (Price, 2012), both IFG pars orbitalis and MTG were associated with semantic processing, 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 measure of population-normed reliability was the classifier analysis. This type of analysis, which has not previously been used to determine the reliability of an individual's neural activation patterns, has the added advantage of 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 akin to a neural "fingerprint." In fact, the only subject that was ever mis-classified was Subject 7, who also had the lowest within-subject ICC value and Dice coefficient, thus demonstrating consistency across measures. The same classification method trained on the *null* maps also demonstrated high accuracy, roughly equivalent to that achieved by the *speech* map classifier. It is 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. We thus interpret the current result as a lower bound of discriminability of speech activation maps among individuals which might be improved with more sophisticated machine learning algorithms.

4.3. *Speech* vs. *Null* Reliability

As expected, the portion of the mean BOLD signal associated with brain morphology and neurovasculature demonstrated high reliability within subjects and high discriminability. However, the lack of a correlation between reliability measures in the *speech* and *null* maps suggests that unique activation patterns during the speech task are not dependent on underlying individual anatomy.

4.4. Reliability for Speech Production across Tasks

The speech tasks used to assess within-subject reliability herein 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, it means that the reported reliability (and discriminability) measures reflect consistency of the speech production network response rather than the response to 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 bi-syllabic words and pseudowords). This is important for assessing the validity of future subject-specific analyses that use speaking tasks that depart from those in the present study.

5. Conclusion

Based on the results of four measures of reliability, we conclude that speech activation maps for most neurologically-healthy speakers are generally highly reliable, providing justification for subject-specific neuroimaging research of 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 et al., 2008] and mild cognitive impairment [Zanto et al., 2014] are promising),

and ultimately whether subject-specific neuroimaging techniques can be used 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 et al., 2016; Reinkensmeyer et al., 2016), and has the potential to improve diagnosis and treatment for people with speech disorders.

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.

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., Wheless, J. W., & 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., Aquino, D., Marras, C., Caldiroli, D., & 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.

Brown, S., Ingham, R. J., Ingham, J. C., Laird, A. R., & Fox, P. T. (2005). Stuttered and fluent speech production: An ALE meta-analysis of functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 105–117. https://doi.org/10.1002/hbm.20140

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

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., & Marchal, G. (1995). Automated multimodality image registration using information theory. In *Information processing in medical imaging* (pp. 263–274). Kluwer Academic Publishers.

Costafreda, S. (2009). Pooling fMRI data: Meta-analysis, mega-analysis and multi-center studies. *Frontiers in Neuroinformatics*, *3*. https://doi.org/10.3389/neuro.11.033.2009

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

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

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

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Farias, D., Davis, C. H., & Wilson, S. M. (2014). Treating apraxia of speech with an implicit protocol that activates speech motor areas via inner speech. *Aphasiology*, *28*(5), 515–532. https://doi.org/10.1080/02687038.2014.886323

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 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., Gollub, R. L., Lauriello, J., Lim, K. O., Cannon, T., Greve, D. N., Bockholt, H. J., Belger, A., Mueller, B., Doty, M. J., He, J., Wells, W., Smyth, P., Pieper, S., … 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., 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. (2016). *Neural control of speech*. MIT Press.

Hagler, D. J., Saygin, A. P., & Sereno, M. I. (2006). Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage*, *33*(4), 1093–1103. https://doi.org/10.1016/j.neuroimage.2006.07.036

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., Gurney, E. M., & Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping*, *7*(3), 213–223. https://doi.org/10.1002/(sici)1097-0193(1999)7:3<213::aid-hbm5>3.0.co;2-n

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

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., Meinzer, M., & 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

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

Nettekoven, C., Reck, N., Goldbrunner, R., Grefkes, C., & Weiß Lucas, C. (2018). Short- and long-term reliability of language fMRI. *NeuroImage*, *176*, 215–225. https://doi.org/10.1016/j.neuroimage.2018.04.050

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

Paek, E. J., Murray, L. L., Newman, S. D., & Kim, D.-J. (2019). Test-retest reliability in an fMRI study of naming in dementia. *Brain and Language*, *191*, 31–45. https://doi.org/10.1016/j.bandl.2019.02.002

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

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

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 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., Swinnen, S. P., Ward, N. S., & 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

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., Matthews, P. 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

Strang, G. (1998). *Introduction to linear algebra* (2. ed). Wellesley-Cambridge Press.

Studholme, C., Hawkes, D. J., & Hill, D. L. (1998). *Normalized entropy measure for multimodality image alignment* (K. M. Hanson, Ed.; pp. 132–143). https://doi.org/10.1117/12.310835

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

Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. (2002). Meta-Analysis of the Functional Neuroanatomy of Single-Word Reading: Method and Validation. *NeuroImage*, *16*(3), 765–780. https://doi.org/10.1006/nimg.2002.1131

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

Wymbs, N. F., Ingham, R. J., Ingham, J. C., Paolini, K. E., & Grafton, S. T. (2013). Individual differences in neural regions functionally related to real and imagined stuttering. *Brain and Language*, *124*(2), 153–164. https://doi.org/10.1016/j.bandl.2012.11.013

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