fMRI Insights on Word Similarity Within the Brain:

Identifying Distinguishable Word Features for Speech BCI

Caya Dreessen

Faculty of Medicine, Utrecht University

Major Research Internship Report

Supervisor: Dr. Mathijs Raemaekers

Word Count: 4819

Due Date: November 2023

Abstract

With its ability to investigate the similarity between brain responses to different stimuli, representational similarity analysis (RSA) may be a powerful tool to identify distinguishable linguistic features for speech BCIs. But despite RSAs potential, its adoption in speech BCI research has been somewhat limited and only moderately successful. The current study aims to bridge this gap by applying RSA to word production fMRI data – allowing the investigation of neural similarity between different words. 11 healthy subjects pronounced 28 words during fMRI image acquisition at 7T. Representational Similarity Matrices (RSMs) were computed from activity within the sensorimotor cortex, the cerebellum and the superior temporal area. The magnitude and distribution of similarity levels within these RSMs suggested inconsistency of neural responses to words. In an attempt to improve the quality of our data, various correction methods were applied. These included the reduction of general noise by regressing out white matter principal components, as well as accounting for pronunciation-related movements by excluding trial-pair comparisons with substantial head position divergence and regressing out trial means. None of these correction methods succeeded in revealing consistent neural responses to words, rendering further interpretation of word similarities inappropriate. The successful cross-validation of our RSA configurations with gesture data indicates that our implementation is fundamentally sound, and simultaneously hints towards reasons as to why our word dataset may be unsuitable for RSA. The findings of the present study are discussed with regards to fMRIs temporal resolution, pronunciation-related motion artifacts, voxel selection, and amounts of trial repetitions.

Keywords: Word Production, Speech, Representational Similarity Analyses, Functional MRI, Brain Computer Interface

Introduction

The ability to communicate is fundamental for our existence as social beings, and its loss can have detrimental effects on an individual's quality of life. This is the reality for many people who lost the control over their motor functions. For these cases, the past decades have generated numerous neurotechnological solutions, with Brain-Computer Interfaces (BCIs) emerging as a particularly promising one. By letting computer systems execute mental commands, BCIs allow their users to interact with the environment without relying on own movements (RSA; Hramov et al., 2021). In the context of restoring communication, these mental commands can be retrieved from a range of neural responses. However, the most intuitive source may be activity associated with the final stages of language production.

Various imaging techniques are being employed to investigate language representation within the brain, with the ultimate goal of optimizing speech BCIs. Among them, functional Magnetic Resonance Imaging (fMRI) stands out for its non-invasiveness and exceptional spatial resolution. These attributes make it particularly suitable for exploring the intricate pronunciation-related response patterns across a large subject pool. Nonetheless, utilizing fMRI for the study of speech production presents its unique set of challenges. Speech-related motion (Gracco et al., 2005; Hirsch et al., 2018) and respiration (Gracco et al., 2005) during volume acquisition can for instance cause signal changes that mimic or mask task-related BOLD responses. Moreover, fMRIs low temporal resolution renders it inadequate to capture the dynamics of neural processes during speech (Grootswagers et al., 2013). But despite these drawbacks, several fMRI studies have successfully decoded diverse utterances from various brain areas: Bleichner et al. (2015) reached for instance 90% accuracy in labelling four mouth movements, with a pattern-correlation classifier analysing activity within the left sensorimotor cortex. Furthermore, Correia et al. (2020) achieved significant distinction of syllables, using a support-vector machine (SVM) that interpreted M1, cerebellum and basal ganglia activity. Likewise, Markiewicz and Bohland (2016) effectively employed an SVM, here for the identification of vowels based on activity within the bilateral superior temporal sulcus.

Speech BCI research commonly utilizes classification-based methods to analyse fMRI data (e.g., Bleichner et al., 2015; Correia et al., 2020; Markiewicz & Bohland, 2016; Otaka et al., 2008; Vitória et al., 2023). Nevertheless, for the purpose of revealing the information encoded by a region of interest (ROI), Representational Similarity Analysis (RSA; Kriegeskorte et al., 2006) may be a more suitable choice. By quantifying the similarity of brain responses to different stimuli, RSA holds promise for being a powerful tool to identify distinguishable linguistic features (Evans & Davis, 2015) for speech BCIs. But despite RSA's potential, its adoption in speech BCI research has been somewhat limited. And while performing RSA on syllable production fMRI data led to contradictory findings (Carey et al., 2017; Zhang et al., 2020), its application on word production fMRI data has, to date, been unsuccessful (Bailey et al., 2021).

In light of these observations, our study strived to revisit the analysis of word production fMRI data using RSA. This was done by having subjects pronounce different words during scanning, followed by similarity computations based on their underlying BOLD responses. Importantly, response patterns used for the analysis were derived from brain regions known to be involved in late stages of language production – namely the bilateral sensorimotor cortex (e.g., Lotze et al., 2000), the superior temporal area, and the cerebellum (e.g., Zhang et al., 2020). With this study, we aim to explore the neural similarity distribution of words, shed light on novel methodologies for exploring language production and in doing so, play a part in the advancement of user-friendly speech BCIs.

Methods

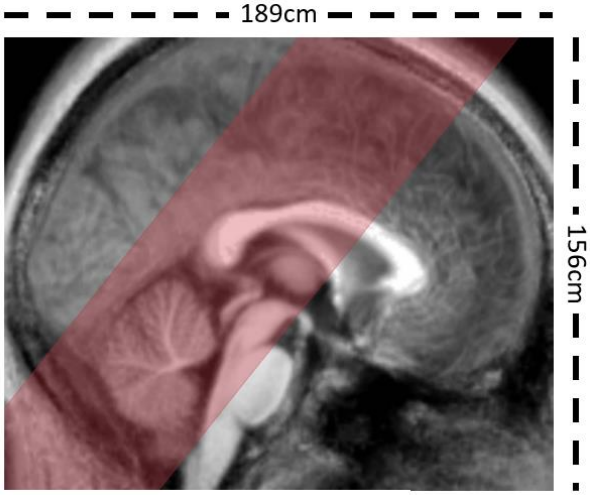
Subjects and Task Preparation

11 Dutch-speaking subjects (mean age: 28, age range: 22-57; 7 males and 4 females) without any neurological disease or contraindications to MRI scanning, participated in this study. Informed consent in agreement with the Declaration of Helsinki (World Medical Association, 2013) was obtained prior to participation. The experimental procedure was approved by the medical-ethical committee of the University Medical Center Utrecht.

Scanner Protocol

For data collection, the subjects were positioned in a whole-body 7 Tesla MR scanner (Achieva, Philips Health Care, Best, Netherlands) equipped with a 32-channel head-coil (Nova Medical, MA, USA). Prior to the main task, a T1-weighted MP2RAGE image of the entire brain was obtained. Functional data was collected using a gradient-echo echo-planar imaging (EPI) sequence (TR = 1500 ms; TE = 25 ms; flip angle = 62°; anterior-posterior phase encoding direction; 33 slices; ascending interleaved slice acquisition order; voxel size = 1.75 x 1.586 x 1.586 mm^3; FOV = 57.750 *AP* x 184.000 *FH* x 226.462 *RL* mm^3). The field of view was configured to cover the bilateral superior temporal area, ventral sensorimotor cortex, and the cerebellum while minimizing inclusion of major brain-penetrating arteries in the phase-encoding direction (see Figure 1). Subjects underwent 253 functional image acquisitions per run.

Figure 1. *Field of View*



Note. This figure displays the field of view (57.750 AP x 184.000 FH x 226.462 *FH* mm^3), laid over the mid-sagittal view of the average mean co-registered normalized anatomical image of our sample. The field of view was configured to cover all areas of interest while minimizing the inclusion of major brain-penetrating arteries in the phase-encoding direction.

Stimuli and Task Design

After task practice, subjects were positioned in the scanner, equipped with hearing protection and prism glasses. The prism glasses allowed them to view a mirror on top of the coil, reflecting a waveguide-projected screen. Subjects were instructed to pronounce words appearing on that screen, while restricting any other head movements. Overall 3 runs were completed (two subjects completed respectively 4 and 5 runs since their comfort levels allowed longer scanning times) with each run

presenting the same set of 28 words (inter-stimulus interval = 13.5 secs). Word orders per run were randomly generated, with the same sequences being used for all subjects. The whole procedure, from entering the MRI facility to leaving it, took on average two hours, out of which the subjects spend about one hour inside the scanner.

The stimuli for this task were extracted from the DiaArt register - a list of 70 Dutch words created by Wieling et al. (2016) to investigate dialectical variance within the Netherlands. Words containing more or less than two-syllables were excluded to prevent word similarity from being primarily influenced by word length. This resulted in a list of 28 two-syllable words, collectively covering various linguistic properties (see Table 1).

Table 1. *Word Stimuli Extracted from the DiaArt Register*

Wordlist			
1. ballen	8. kameel	15. ogen	22. treinen
2. bellen	9. kersen	16. paarden	23. uilen
3. bijlen	10. krukken	17. palen	24. vingers
4. bloemkool	11. lepel	18. peren	25. vlaggen
5. bogen	12. molen	19. stoelen	26. vogels
6. brillen	13. muggen	20. taarten	27. wielen
7. dolfijn	14. negen	21. tollen	28. zagen

Note. Listed are the 28 two-syllable Dutch word stimuli, derived from the DiaArt register, covering various linguistic properties. The numbering relates to the baseline stimulus order, represented in all following figures.

Preprocessing

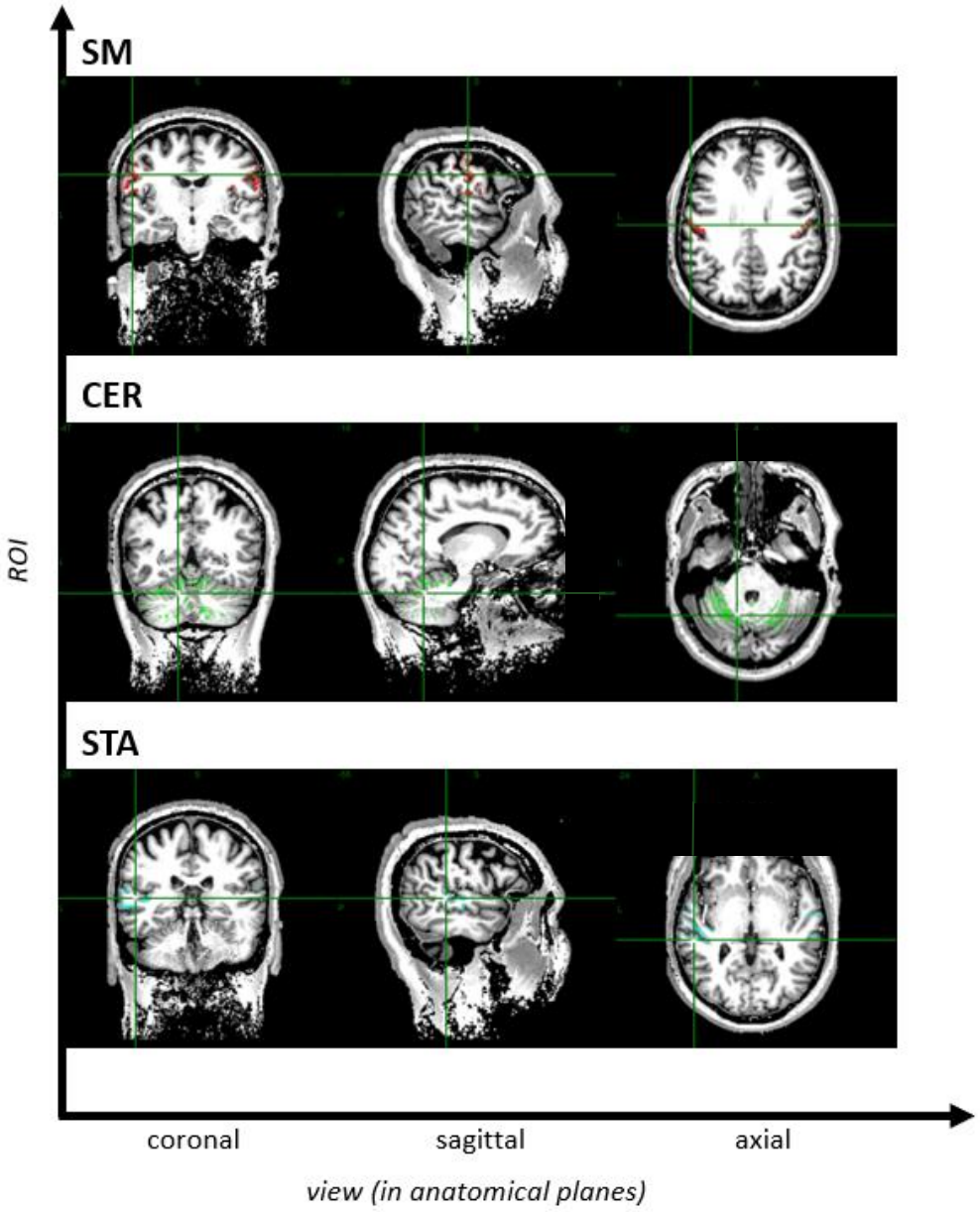
The data was pre-processed incorporating tools from FreeSurfer 7.0 (Fischl, 2012), Fsl 6.0 (Jenkinson et al., 2012) and SPM12 (Friston, 2007), integrated within a MATLAB pipeline (The MathWorks Inc, 2022). All functional images were Nordic-corrected to remove signal components not distinguishable from thermal noise (Moeller et al., 2021), slice-time corrected (SPM), realigned and unwarped to the mean EPI image (SPM) to address head motion, top-up corrected (FSL) to adjust geometrical distortions (Andersson et al., 2003; Smith et al., 2004), co-registered to the subjects' high-resolution T1 scan, and high pass filtered using a kernel with a cut-off at 75 seconds to eliminate low frequency signal drift. The mean activity of each run was subtracted from all time points within that run, followed by the standardization of the data across all runs. We chose to forgo smoothing and

instead fully leverage the spatial resolution of our data in order to preserve fine-scale distinctions in word activity potentially crucial for our investigation (Dimsdale-Zucker & Ranganath, 2018; Kriegeskorte et al., 2006; Zhang et al., 2020).

Representational similarity analysis

RSA was performed separately for the sensorimotor cortex, the cerebellum and the superior temporal area. To identify these ROIs within the individual subject space, cortical surface reconstructions were generated from the anatomical images, and ROIs identified using the Desikan-Killiany atlas as a parcellation scheme (Desikan et al., 2006; FreeSurfer; labels: sensorimotor cortex - left and right precentral cortex, left and right postcentral cortex; cerebellum - left and right cerebellum, superior temporal area - left and right superiortemporal cortex, left and right temporal pole). Voxels within each ROI were further filtered for their task-relevance, with the aim to minimize the influence of noise and irrelevant neural activity. Achieving this involved forming a general linear model (SPM), with a single task regressor and global mean regressor for each run. The task regressor, representing the pronunciation of all 28 words, was convolved with a standard hemodynamic response function. To account for pronunciation-related motion artifacts, we additionally introduced regressors that modelled the first two volumes of each trial. The implicit masking threshold was set at 50%. The final t-values obtained from the analysis were used to identify the voxels with the strongest-task association, with the upper 10% of voxels within an ROI being selected for the representational similarity matrix (RSM) calculation. In absolute numbers, on average 789 voxels within the sensorimotor cortex, 1429 voxels within the cerebellum and 340 voxels within the superior temporal area were included in the analyses. An example for the distribution of selected voxels per ROI can be viewed in Figure 2.

Figure 2. *Subject Example of Voxel Selection within each ROI*

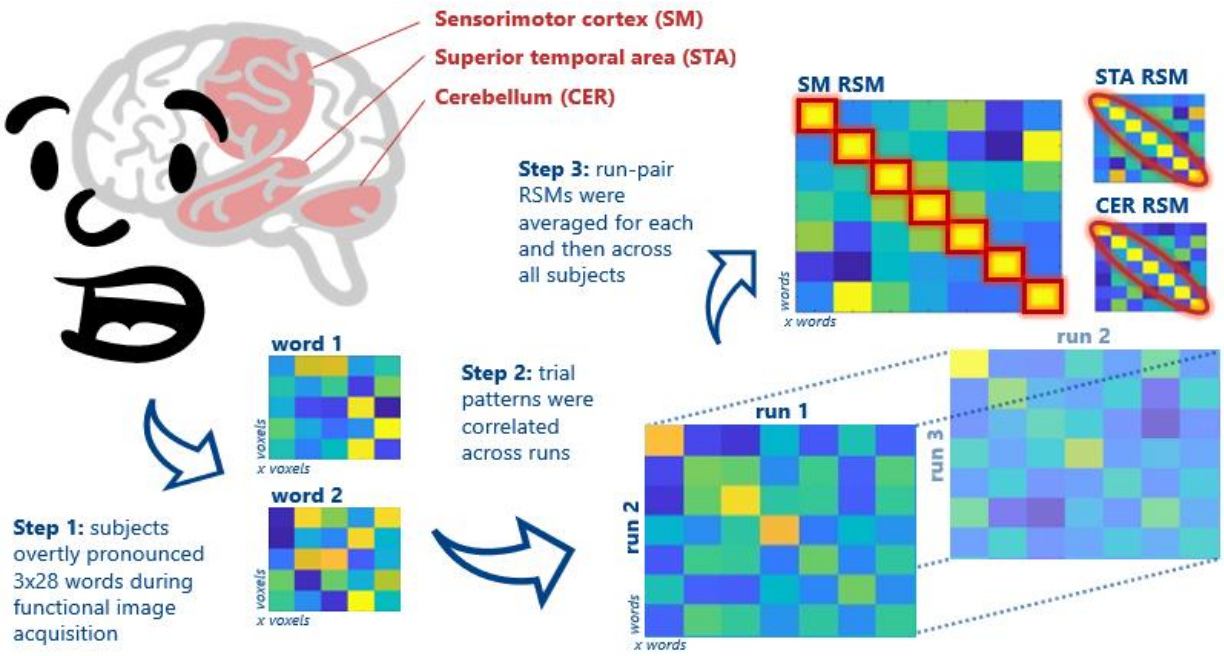


Note. Displayed is a subject example of the task-association based voxel selection within the sensorimotor cortex (first row; selected voxels in red) the cerebellum (second row; selected voxels in green) and the superior temporal area (third row; selected voxels in blue). Voxel distributions are presented from a coronal (first column), sagittal (second column) and axial (third column) perspective, with viewpoint locations per ROI being indicated by green lines.

A subject's RSM was computed by correlating the activity pattern of all trials with each other (using Pearson's correlation), Fisher Transforming the resulting coefficients and averaging across those describing the same word pairs, while leaving out comparisons of trials from the same run. The averaged similarity values were then arranged in a 28 by 28 matrix, where each cell represents the comparison between two words. Lastly, group-level RSMs were obtained by averaging across all individual RSMs per ROI.

Data quality across the sample was evaluated by inspecting the respective RSMs for a visible diagonal (see Figure 3). Since on-diagonal elements represent the average neural similarity between a word and its repetitions, high values along the diagonal imply that neural responses to words are consistent across runs. And since off-diagonal elements represent the average neural similarity between one word and another, increased values within the on- compared to the off-diagonal space imply that the variability in neural similarity is at least to some extend driven by word features. Both – consistent neural responses and similarity dependent on word features are a prerequisite for making subsequent judgments about the dissimilarity between different words. To ultimately determine whether the value increase along the diagonal is significantly above zero, one-tailed one-sample t-tests on the sample-mean of the on/off-diagonal difference were performed, utilizing IBM SPSS 29 (IBM Corp., Armonk, N.Y., USA). Prior to these tests, normality (Shapiro-Wilk test) and the absence of outliers was confirmed. The initial alpha level was set at 0.05. However, to account for the familywise error rate in multiple comparisons, a Bonferroni correction was applied, resulting in an adjusted significance threshold. Follow-up analyses are described within the results section.

Figure 3. RSA Configurations

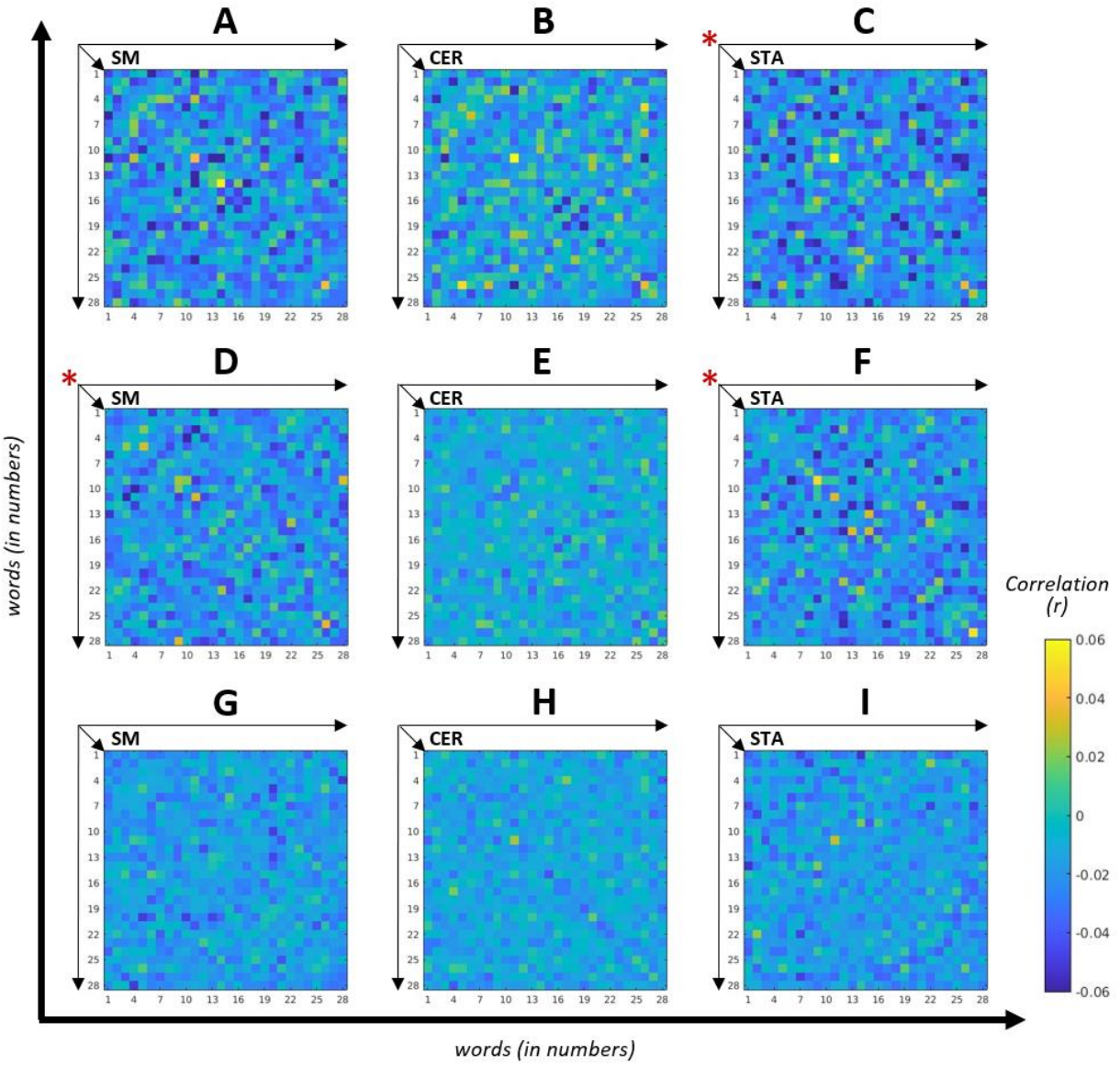


Note. This figure illustrates the steps taken to compute group-level RSMs per ROI, as described in the Methods section. Activity patterns of all trials were derived from word production fMRI data acquired in step 1. Those activity patterns were then correlated with one another (step 2), and the resulting similarity values averaged; first across comparisons of the same word-pair, and eventually across the group (step 3).

Results and Follow-Up Analyses

In order to explore neural similarity between the produced words, group-level RSMs were computed per ROI. But before delving into interpretations, data quality was assessed by inspecting the plotted RSMs and testing the rise of on-diagonal values for significance. As can be seen in Figure 4 (first row), all group-level RSMs exhibit consistently low correlation coefficients, while also lacking a visible diagonal. The latter impression is in line with t-test outcomes on RSMs computed from sensorimotor cortex [Table 2: M = .010, SD = .019, $t(10) = 1.831$, p = .048] and cerebellum activity [Table 2: M = .003, SD = .012, $t(10) = .798$, p = .222], but contradicts outcomes describing the superior temporal area RSM [Table 2: M = .014, SD = .019, $t(10) = 2.50$, p = .016]. These findings suggest the presence of word feature sensitivity within the superior temporal area, and its absence within the other ROIs. Note however that the mean on-diagonal value within the superior temporal area RSM increased on average by only .014. The generally low values along the diagonal in all RSMs suggests that word-related activity patterns are considerably inconsistent across all ROIs.

Figure 4. *Group-level RSMs Computed from Word-Production fMRI Data; Separated by ROIs*



Note. This figure displays all group RSMs, derived from activity within the sensorimotor cortex (SM; first RSM column), the cerebellum (CER; second RSM column) and the superior temporal area (STA; third RSM column). RSMs were computed from word-pronunciation data without additional corrections (first RSM row), corrected for noise (second RSM row), and corrected for motion (third RSM row). The numbers along the RSM axes indicate the words in comparison (corresponding words can be identified within Table 1). Each cell represents a coefficient describing the word-pair similarity. Coefficient values are indicated in colours, with the colour-to-value mapping depicted in the colour ramp in the lower right corner. Red stars next to the upper left corner of each RSM symbolize a significant mean increase in on-diagonal compared to off-diagonal values. This is the case for the RSM computed from uncorrected superior temporal area activity (C), as well as the RSMs computed from noise-corrected sensorimotor (D) and superior temporal area (F) activity.

Noise Exclusion

The low correlation coefficients along the diagonal of our group-level RSMs point towards factors considerably interfering with the word-related BOLD responses. To eliminate these unknown noise sources, functional data underwent correction based on white matter activity – a signal assumed to be devoid of task-related activity so that artifact-related activity can be extracted. The white matter space itself was defined using the FreeSurfer parcellation according to the Desikan-Killiany atlas (labels: white matter space - left & right white matter space), and then eroded with a 3D cube-shaped structuring element (side length = 3 voxels). Run activity recorded from this space was submitted to a principal component analysis. The resulting components were then regressed out of the functional data of that respective run. To prevent the exclusion of task-relevant activity, only so many components were selected for the regression that they would still collectively account for less than 0.35 of the variability within the task-related regressors.

Regressing out white matter components changed the RSM landscape slightly but did not lead to notably higher correlation coefficients or visually more prominent diagonals (Figure y). T-test on the mean on/off-diagonal differences did indeed not reveal a diagonal within the RSM computed from cerebellum activity [Table 2: M = .002, SD = .015, $t(10)$ = .466, p = .325], but confirmed their presence now not only within the superior temporal area RSM [Table 2: M = .020, SD = .023, $t(10)$ = 2.914, p = .008] but also the sensorimotor cortex RSM [Table 2: M = .014, SD = .012, $t(10)$ = 3.868, p = .002].

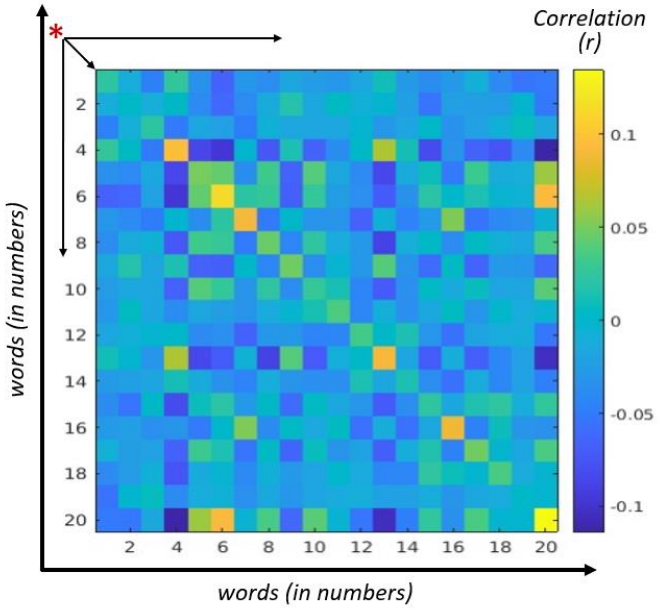
Cross-Validating RSA Configuration with Gesture-Formation fMRI Data

Since our RSM configuration is rather unconventional, the suspicion arose that the adjusted implementation may be the reason for our low on-diagonal values. To cross-validate our approach, we applied it to a dataset, obtained and preprocessed similarly to our own. fMRI recordings were this time acquired from one subject, producing 20 different gestures repeatedly over 10 runs. Other methodological distinctions were the field of view (FOV = 226.462 *AP* x 52.500 *FH* x 184.000 *RL* mm^3), left-sided instead of bilateral motion and the analysis being merely focussed on the right sensorimotor

cortex. The RSM computed from this data was visually and statistically inspected. Though since the experiment only involved a single subject, a permutation test would now serve to determine whether there was a significant rise in on-diagonal values. This involved reshuffling condition labels, reordering RSMs accordingly, and calculating mean differences in these newly arranged RSMs. Iterating the procedure 1000 times generated a distribution of mean differences under the null hypothesis, and allowed to determine the proportion of these mean differences being greater than or equal to the observed difference.

The computed gesture RSM displayed higher correlation coefficients along the diagonal compared to word RSMs, but also off-diagonal elements (see Figure 5). This impression was substantiated with the outcome of the permutation test [M = .0720, p = .000]. Thus, even when working with a limited sample size, our methodology appears to be appropriate – at least for analysing gesture data.

Figure 5. *RSM Computed from Gesture-Formation fMRI Data*

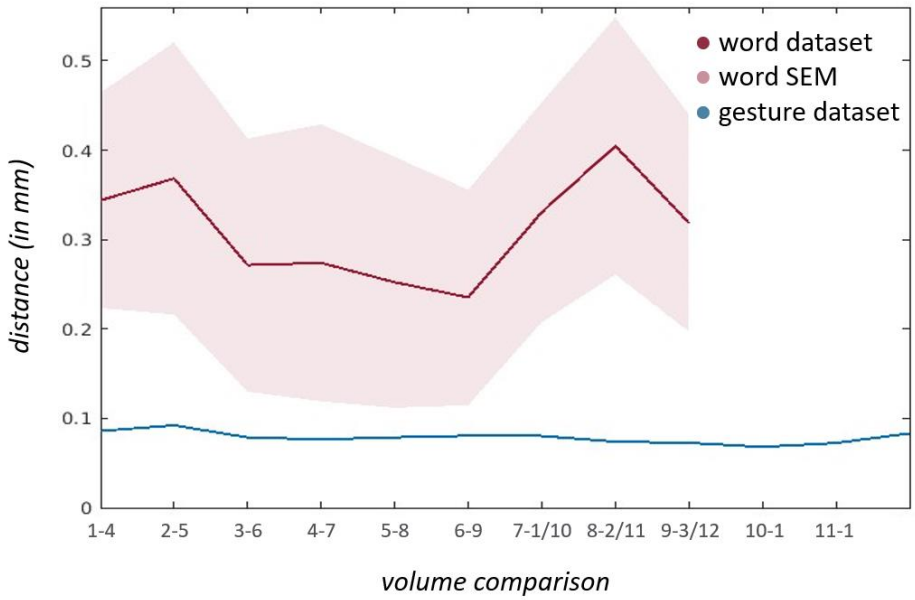


Note. This figure displays an RSM derived from the BOLD activity recorded within the right sensorimotor cortex, in response to gesture-formation. The red star next to the upper left corner of the RSM symbolizes a significant mean increase in on-diagonal compared to off-diagonal values.

Assessing and Controlling for Head Displacement

What may have caused the different RSA outcomes for word and gesture data is the presence of pronunciation-related movements during word data acquisition. These movements may induce head displacements which, once reaching a certain magnitude, will render voxel activity patterns before and after pronunciation non-comparable. To assess the severity of head displacements for both datasets, we first reconstructed the scanner coordinates of each voxel from motion parameters. Coordinates at each time point were then compared with those three timepoints further. The differences were subsequently transformed to distances and lastly averaged across ROIs, trials, runs and subjects, to obtain the distance change between volumes for an average trial. Plotting these averaged distances revealed more pronounced shifts in head position during word compared gesture data acquisition. Furthermore, a plateau at the 9^{th} comparison of the word data plot suggests that subjects tend to move during word production (volume 1 and 2) without returning to their original position afterwards (volume 3).

Figure 6. *Head Position Change; Averaged by Trial, Run, ROIs and Subjects*



Note. This graph displays the head position change during word data (red line) and gesture data acquisition (blue line). Plotted are the distances (y axis) between positions at each volume and those assessed three volumes further (x axis), averaged across trials, runs, ROIs and - for the word dataset - across subjects. The shaded area surrounding the word plot (red line) represents the standard error of the group mean at each comparison. The graph indicates much more pronounced shifts in head position during word compared gesture data acquisition. Furthermore, the word plot trajectory suggests a peculiar distance change throughout and across trials: The distance starts high at the 1st to 4th volume comparison, is slightly increased at the 2nd to 5th comparison, then rapidly decreases to almost 0 for the next 4 comparisons until rising again at the 7th to 1st and 8th to 2nd volume comparison (with the 1st and 2nd volumes being those obtained during the follow-up trial). The two peaks within the word plot can be explained with head motion caused by word pronunciation. The plateau at the 9th comparison, suggests that subjects do not return to their initial head position after pronouncing a word.

In an attempt to avoid the comparison of trials during which head positions diverged too much, only trials pairs with an ROI-averaged divergence of less than a voxel size (1.586mm) were included in the analyses. To obtain the divergence values, we first averaged scanner coordinates per voxel across the volumes 3 to 9 of each trial and compared them across trials. The resulting trial differences were eventually transformed to distances. 15.06% of all word-pairs (across runs and subjects) exhibited distances that surpassed the threshold and were consequently excluded from the analysis.

To additionally exclude motion-related trial mean changes from the analysis input, functional images were replaced by maps representing the trial-by-trial task-following of each voxel: The maps

were generated through a regression analysis conducted on the 3rd to 9th volumes of each trial (the first two volumes were excluded to avoid potential confounding with motion). Resulting regression maps per trial depicted to which extend the activity of each voxel follows the shape of a hemodynamic response function. Beyond their use for computing RSMs, regression maps were also employed to replace the beta t-maps within the voxel selection procedure, since their characteristic to overlook motion-related trial mean changes offers enhanced selection accuracy. The new t-maps were calculated by conducting a t-test on the regression coefficients across trials. As before, the upper 10% of voxels with the highest absolute t-values within each ROI were selected for RSA.

Resulting RSMs appeared to have overall attenuated coefficients (see Figure 4). T-tests performed on the on/off-diagonal difference did not identify an increase of on-diagonal values within any of our ROIs [Table 2: sensorimotor cortex - M = .002, SD = .006, $t(10) = 1.164$, p = .136; cerebellum - M = .002, SD = .006, $t(10) = 1.261$, p = .118; superior temporal area - M = .003, SD = .009, $t(10) = 1.373$, p = .100].

Discussion

In this study, we investigated the similarity distribution of word-production fMRI data using RSA. Group RSMs computed from the sensorimotor cortex, the cerebellum and the superior temporal area, indicated overall low neural similarity between words and their repetitions across runs. Moreover, only within the superior temporal area were neural responses to word repetitions more similar to each other than to those of different words. A follow-up noise correction based on white matter activity had little effect on these RSM landscapes; except for the significant similarity increase for word repetitions now being not only present within the superior temporal area but also in the sensorimotor cortex. Our findings suggest that in both ROIs, neural similarity varies according to word feature similarity. However, since the detected increase is of small magnitude, neural similarities between word repetitions remain low across ROIs. We thus conclude that our words did not evoke consistent neural responses within any of the ROIs.

To address concerns about our methodology, we cross-validated RSA configurations using a gesture dataset. The resulting RSM exhibited the desired similarity distributions, implying increased consistency in neural responses, with the similarity between them being driven by gesture features. The differential suitability of RSA for word and gesture data could be attributed to the presence of pronunciation-related movements during word data acquisition. Assessments of head displacements indeed revealed more severe head position changes during word compared to gesture data acquisition. Subsequent motion corrections, including trial-pair exclusion and regressing out trial means, only attenuated similarity values within the RSMs, to the degree that t-tests could not detect a similarity increase for word repetitions within any of the ROIs. Thus, motion artifacts do either not explain the observed low similarity values for word repetitions or are not fully accounted for. Given that activity patterns across runs appear to remain rather inconsistent, independent of extensive noise and motion correction, any further interpretation of word similarities is deemed invalid.

Since it is not common practice for RSA studies to publish their neural RSMs, our ability to directly compare results is limited. Instead, we will focus on the frequently reported correlations between neural and feature-based similarity distributions, as well as their implications, in order to relate our findings to existing literature. For instance, Bailey et al. (2021) searched within various ROIs – including the sensorimotor cortex and superior temporal area - for the phonological representation of 30 overtly pronounced words. However, none of the ROIs would exhibit word-related activity patterns whose similarity varied according to phonological similarity. As past studies have repeatedly confirmed the representation of phonological information within several of the investigated ROIs (e.g., Schomers & Pulvermüller, 2016), Bailey et al. (2021)'s null results may instead be explained by the same neural inconsistency we had encountered within our dataset. Zhang et al. (2020)'s findings, on the other hand, suggests no such complications. 19 subjects pronounced 16 different syllables during functional data acquisition. As hypothesised, similarity between syllable-related activity varied in motor, somatosensory and auditory regions depending on articulatory similarity, as well as in auditory regions depending on phonetic similarity.

A distinction between the studies, that might explain the varying success of RSA application on language-production fMRI data, is the stimulus length. While Bailey et al. (2021) let subjects pronounce 5 to 10 letter words, and our study employed two-syllable words, Zhang et al. (2020) focussed solely on one-syllable pronunciations. Since fMRI's low temporal resolution leads to neural responses being captured as averages across a certain period, longer utterances may introduce a broader range of speech features during each period, making words harder to distinguish at the neural level than syllables. This reasoning also aligns with our observations regarding the differential suitability of RSA for our word and gesture data: Just like syllables, gestures involve shorter and simpler movement sequences than words, potentially making them more discernible in the brain.

Despite the use of short stimuli, Carey et al. (2017) reported findings similarly puzzling to those of our and Bailey et al. (2021)'s study. They examined the sensorimotor cortex and anterior cerebellum for articulatory representations of four overtly pronounced vowels. In neither ROI did the similarity between vowel-related activities vary according to articulatory similarity, contrasting observations on articulatory representations within the sensorimotor cortex (e.g., Salari et al., 2019; Zhang et al., 2020). The success of RSA application might therefore not only be influenced by the stimulus length, but also the impact of motion artifacts. Accordingly, while the subject in the gesture dataset minimally altered their head position during acquisition, subjects in the word dataset displayed extensive head displacements, likely attributed to pronunciation-related movements. Bailey et al. (2021) confronted similar issues, resulting in the exclusion entire runs. And although not explicitly stated, Carey et al. (2017)'s data could have been affected by motion too. In all three cases, motion correction post-acquisition may not have been sufficient to eliminate motion artifacts from the data. Introducing additional measures at the time of acquisition, like comfortable head immobilization (Gracco et al., 2005), could potentially address this issue. A paradigm repeatedly being used in fMRI studies to mitigate speech-related motion artifacts (e.g., Achim et al., 2021; Correia et al., 2020; Stefaniak et al., 2022) is sparse temporal sampling. Motion during scanning can lead to repeated excitation of certain regions, leaving others unstimulated. The variation in excitation can impact later scans. Sparse temporal

sampling intends to avoid this differential spin history by pausing image acquisition during motion periods (Gracco et al., 2005).

Besides stimulus length and pronunciation-related head displacements, our voxel selection method may have been another contributing factor to an incoherent RSA output. The number of voxels included in our RSM computations range from 340 to 1429, depending on the ROI. Forming activity patterns from such big samples carries the risk of considering noisy voxels. Many studies thus further reduce their sample either through a searchlight, where only 11 to 30 adjacent voxels are examined at a time (Bailey et al., 2021; Carey et al., 2017; Waters et al., 2021), or - in case of Zhang et al. (2020)'s study - by choosing smaller sub areas containing an average of 265 voxels. Considering the relatively dense representation of pronunciation-related activity within the sensorimotor cortex (e.g., Tourville et al., 2019), and clear-cut functional division within the superior temporal area (e.g., Bhaya-Grossman & Chang, 2022), the exploration of a more focused voxel selection strategy may be worthwhile. An alternative approach to limit the influence of noisy voxels on RSM landscapes may be voxel weighting. According to Kaniuth and Hebart (2022), attributing equal importance to all voxels may lead to an underestimation of the informativeness of RSMs. To unveil their potential, one can reweight voxel contributions depending on their importance in trial distinction. This approach essentially adopts the strategy of some multivariate linear decoding strategies, where the focus lays on data channels that carries the signal of interest.

A last factor that may have affected the RSM landscape and certainly limited the exploration of alternative explanations for our findings, is the number of repetitions. While the gesture dataset comprises 10 repetitions, the word dataset features only three. This imbalance was compensated for by forming group RSMs, uniting overall more repetitions (3 x 11 subjects) than the gesture RSM (10 x 1 subject). Nevertheless, group RSMs computed from the word dataset would not exhibit the same desired similarity distribution displayed in the gesture RSM. It is debatable whether this is due to gesture-related activity being more consistent than word-related activity, implying that the word dataset would need many more repetitions to extract similarly robust activity patterns. Notably, at the

syllable level, already three repetitions were sufficient for forming sensible similarity distributions (Zhang et al., 2020). One way or the other, incorporating additional repetitions would improve our understanding and accommodation of subject-specific noise sources. Moreover, it would open up the possibility to apply classifiers to the dataset, offering further insights into aspects such as data quality and appropriate voxel selection.

Conclusion

While classification-based methods remain to be the conventional choice in speech BCI research to analyze fMRI data, RSA presents a promising alternative. And although our study did not yield the expected results of consistent neural responses to words, it highlights potential pitfalls when performing RSA on word production fMRI data, and provides valuable suggestions for future studies. Those include the careful selection of stimulus length and number of repetitions, proactive measures for mitigation of potentially extensive motion artifacts, and a more focal voxel selection.

References

Achim, A. M., Deschamps, I., Thibaudeau, É., Loignon, A., Rousseau, L.-S., Fossard, M., & Tremblay, P. (2021). The neural correlates of referential communication: Taking advantage of sparse-sampling fMRI to study verbal communication with a real interaction partner. *Brain and Cognition*, *154*, 105801.

Andersson, J. L., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*, *20*(2), 870-888.

Association, W. M. (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *Jama*, *310*(20), 2191-2194.

Bailey, L. M., Bodner, G. E., Matheson, H. E., Stewart, B. M., Roddick, K., O'Neil, K., Simmons, M., Lambert, A. M., Krigolson, O. E., & Newman, A. J. (2021). Neural correlates of the production effect: An fMRI study. *Brain and Cognition*, *152*, 105757.

Bhaya-Grossman, I., & Chang, E. F. (2022). Speech computations of the human superior temporal gyrus. *Annual review of psychology*, *73*, 79-102.

Bleichner, M., Jansma, J., Salari, E., Freudenburg, Z., Raemaekers, M., & Ramsey, N. (2015). Classification of mouth movements using 7 T fMRI. *Journal of neural engineering*, *12*(6), 066026.

Carey, D., Miquel, M. E., Evans, B. G., Adank, P., & McGettigan, C. (2017). Vocal tract images reveal neural representations of sensorimotor transformation during speech imitation. *Cerebral cortex*, *27*(5), 3064-3079.

Correia, J. M., Caballero-Gaudes, C., Guediche, S., & Carreiras, M. (2020). Phonatory and articulatory representations of speech production in cortical and subcortical fMRI responses. *Scientific Reports*, *10*(1), 4529.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., & Hyman, B. T. (2006). An automated labeling system for subdividing the

human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968-980.

Dimsdale-Zucker, H. R., & Ranganath, C. (2018). Representational similarity analyses: a practical guide for functional MRI applications. In *Handbook of behavioral neuroscience* (Vol. 28, pp. 509-525). Elsevier.

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cerebral cortex*, *25*(12), 4772-4788.

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774-781.

Friston, K. (2007). A short history of SPM. *Statistical parametrical mapping: The analysis of functional brain images*, 3-9.

Gracco, V. L., Tremblay, P., & Pike, B. (2005). Imaging speech production using fMRI. *Neuroimage*, *26*(1), 294-301.

Grootswagers, T., Dijkstra, K., Ten Bosch, L., Brandmeyer, A., & Sadakata, M. (2013). Word identification using phonetic features: towards a method to support multivariate fMRI speech decoding. Interspeech,

Hirsch, J., Adam Noah, J., Zhang, X., Dravida, S., & Ono, Y. (2018). A cross-brain neural mechanism for human-to-human verbal communication. *Social cognitive and affective neuroscience*, *13*(9), 907-920.

Hramov, A. E., Maksimenko, V. A., & Pisarchik, A. N. (2021). Physical principles of brain–computer interfaces and their applications for rehabilitation, robotics and control of human brain states. *Physics Reports*, *918*, 1-133.

Inc, M. (2022). (Version 9.13.0; R2022b) Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782-790.

Kaniuth, P., & Hebart, M. N. (2022). Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *Neuroimage*, *257*, 119294.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863-3868.

Lotze, M., Erb, M., Flor, H., Huelsmann, E., Godde, B., & Grodd, W. (2000). fMRI evaluation of somatotopic representation in human primary motor cortex. *Neuroimage*, *11*(5), 473-481.

Markiewicz, C. J., & Bohland, J. W. (2016). Mapping the cortical representation of speech sounds in a syllable repetition task. *Neuroimage*, *141*, 174-190.

Moeller, S., Pisharady, P. K., Ramanna, S., Lenglet, C., Wu, X., Dowdle, L., Yacoub, E., Uğurbil, K., & Akçakaya, M. (2021). NOise reduction with DIstribution Corrected (NORDIC) PCA in dMRI with complex-valued parameter-free locally low-rank processing. *Neuroimage*, *226*, 117539.

Otaka, Y., Osu, R., Kawato, M., Liu, M., Murata, S., & Kamitani, Y. (2008). Decoding syllables from human fMRI activity. Neural Information Processing: 14th International Conference, ICONIP 2007, Kitakyushu, Japan, November 13-16, 2007, Revised Selected Papers, Part II 14,

Salari, E., Freudenburg, Z., Branco, M., Aarnoutse, E., Vansteensel, M., & Ramsey, N. (2019). Classification of articulator movements and movement direction from sensorimotor cortex activity. *Scientific Reports*, *9*(1), 14165.

Schomers, M. R., & Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. *Frontiers in human neuroscience*, *10*, 435.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., & Flitney, D. E. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208-S219.

Stefaniak, J. D., Geranmayeh, F., & Lambon Ralph, M. A. (2022). The multidimensional nature of aphasia recovery post-stroke. *Brain*, *145*(4), 1354-1367.

Tourville, J. A., Nieto-Castañón, A., Heyne, M., & Guenther, F. H. (2019). Functional parcellation of the speech production cortex. *Journal of Speech, Language, and Hearing Research*, *62*(8S), 3055-3070.

Vitória, M. A., Fernandes, F. G., van den Boom, M., Ramsey, N., & Raemaekers, M. (2023). Decoding single and paired phonemes using 7T functional MRI.

Waters, S., Kanber, E., Lavan, N., Belyk, M., Carey, D., Cartei, V., Lally, C., Miquel, M., & McGettigan, C. (2021). Singers show enhanced performance and neural representation of vocal imitation. *Philosophical Transactions of the Royal Society B*, *376*(1840), 20200399.

Zhang, W., Liu, Y., Wang, X., & Tian, X. (2020). The dynamic and task-dependent representational transformation between the motor and sensory systems during speech production. *Cognitive Neuroscience*, *11*(4), 194-204.

Table 2. *One-Tailed One-Samples T-Test Results Assessing the Significance of the Mean On/Off-Diagonal Difference; Separated by ROIs*

A Tests on RSMs Computed from Uncorrected Word-Pronunciation Data

	Descriptive Statistics			T-Test Output					Effect Size		
ROI	N	Mean Difference	Std. Deviation	t	df	Significance (One-Sided p)	95% CI		Cohen's d	95% CI	
							Lower	Upper		Lower	Upper
SM	11	.01030	.01866	1.831	10	.048	-.0022	.0228	.552	-.097	1.178
CER	11	.00285	.01184	.798	10	.222	-.0051	.0108	.241	-.365	.835
STA	11	.01425	.01890	2.501	10	**.016**	.0016	.0269	.754	.065	1.415

B Tests on RSMs Computed from Noise-Corrected Word-Pronunciation Data

	Descriptive Statistics			T-Test Output					Effect Size		
ROI	N	Mean Difference	Std. Deviation	t	df	Significance (One-Sided p)	95% CI		Cohen's d	95% CI	
							Lower	Upper		Lower	Upper
SM	11	.01433	.01229	3.868	10	**.002**	.0061	.0226	1.166	.373	1.926
CER	11	.00210	.01491	.466	10	.325	-.0079	.0121	.141	-.457	.731
STA	11	.02017	.02296	2.914	10	**.008**	.0047	.0356	.897	.161	1.566

C Tests on RSMs Computed from Motion-Corrected Word-Pronunciation Data

	Descriptive Statistics			T-Test Output					Effect Size		
ROI	N	Mean Difference	Std. Deviation	t	df	Significance (One-Sided p)	95% CI		Cohen's d	90% CI	
							Lower	Upper		Lower	Upper
SM	11	.00216	.00617	1.164	10	.136	-.0020	.0063	.351	-.267	.953
CER	11	.00221	.00581	1.261	10	.118	-.0017	.0061	.380	-.242	.985
STA	11	.00385	.00931	1.373	10	.100	-.0024	.0101	.414	-.213	1.022

Note. The tables display the descriptive statistics, t-test outputs and respective effect sizes, corresponding to RSMs computed from uncorrected word-pronunciation data (A), noise-corrected word-pronunciation data (B) and motion-corrected word-pronunciation data (C). One-tailed one-samples t-tests compared the mean on/off-diagonal difference to zero. Normality and the absence of outliers is assumed. P-values surpassing the Bonferroni-corrected significance threshold ($p < .017$) appear in bold.