Supplementary Texts

S1. Spatial characteristics of articulator motion components

Articulator motion artifacts have multiple sources of origin and may be driven by two distinct mechanisms. First, noise signals related to jaw motion can propagate into inferior frontal and temporal brain regions (see Fig. 2, Panel 1 in the main text). They are likely caused by perturbations of the static magnetic field (Birn et al., 1998) due to bone movement and muscle contraction. Second, in some other components, signals in oral/laryngeal cavities (including tongue; Fig. 2, Panel 2; see Fig. 1b for an enlarged view), nasal/pharyngeal cavities (Fig. 2, Panel 7), and frontal sinus (Fig. 2, Panel 15) coexist with artifacts in ventromedial cerebral regions and brain stem. Common spatial properties, e.g., a characteristic radiant stripe pattern close to the caudate and thalamus, are often observed for these components; in addition, a single component may sometimes contain multiple sources of origin; thereby indicating a common underlying mechanism – presumably a susceptibility distortion effect (Barch et al., 1999; Kemeny et al., 2005) caused by the vibrations of air-tissue interfaces.

Most of the articulator motion components can be robustly detected by a spatial feature derived from the dual-mask method – out-of-brain ratio. These artifacts may also show scattered/interspersed intensity patterns in both brain tissues and extracerebral soft tissues, possibly caused by image warping. Additionally, for images collected using interleaved sagittal slices, the jaw motion artifacts only affect lateral slices, whereas the artifacts caused by susceptibility distortion primarily affect medial slices; and both may be identifiable from an interleaved pattern in axial or coronal views (Fig. 2, Panel 11).

To identify the susceptibility distortion components near the frontal sinus, sometimes lack all above features, a complementary template matching method (Fig. 2, Panel 15) may be necessary.

These components might be contributed by both articulation and head motion, based on the observation of temporal correlations with both types of motion measures, i.e., speech envelope and rigid-body alignment parameters. Their relationship to articulation might be driven by the vibration of the sinus air cavities during vowel nasalization (Pruthi et al., 2007). Meanwhile, the spatial displacement of these air cavities along with bulk head motion may introduce an effect also called susceptibility-by-motion interaction (Andersson et al., 2001; Wu et al., 1997).

S2. Partitioning of temporal variance based on mechanistic classification

First of all, it is important to mention that although the variance decompositions in sICA and its PCA preprocessing step are computed spatially, here we only report the partitioning of temporal variance, as this is more relevant to the time series analyses commonly applied to fMRI data. The amount of variance for each noise category is defined as the sum of time course variances of all components belonging to that category. In addition, the variances were measured separately for the data acquired in speech production and speech comprehension tasks. The latter served as a low-noise control condition that should contain significant less amounts of articulator and head motion related variances by our prediction. In order to obtain a proper measurement of the task-related variance (i.e., neural signal), the time course segment for each task includes both their task blocks and the succeeding rest intervals.

On average across datasets, the identified noise components explained more than sixty percent of the temporal variance in speech production tasks (IS Fig. 1a). Nearly half of the noise variance was accounted for by head motion. A smaller portion of variance appeared to be due to articulator motion, but these artifacts might be more detrimental in the detection of functional activity because they are more focally distributed. Compared with the large proportion of noise variance, the temporal variance in the neural signal components only account for a fairly small percentage. This is not surprising as it is generally known that fMRI has a rather low signal-to-noise ratio (Caparelli, 2005). When comparing noise variances between production and comprehension (IS Fig. 1b), we indeed found that articulator motion showed the most significant difference followed by head motion, whereas no significant differences were observed for other noise categories.

The denoised datasets contain two other types of variance in addition to the variance explained by neural signal components (IS Fig. 1a). Unlike a direct reconstruction algorithm which operates by summing up the signal components only (Kochiyama et al., 2005), our technique adopts a noise component removal approach, which retains the temporal variances removed in the two preprocessing steps of sICA: spatial mean centering and PCA data reduction.

Spatial mean centering removed the global mean fluctuation caused by scanner drift, bulk respiratory motion (Glover et al., 2000), and CO_2-related hemodynamic changes (Birn et al., 2006). However, due to their heterogeneous amplitude distribution across regions, these effects are better to be removed by a voxel-wise regression method (Macey et al., 2004) rather than subtracting a uniform mean value across voxels.

PCA data reduction also removed a significant amount of temporal variance, which might contain random thermal noise as well as a mixture of unmodeled structured noise and neural signal. This portion of variance should therefore not be removed to avoid “throwing the baby out with the bathwater”. Equally important, PCA residual variance contains the majority of the temporal degrees of freedom in the original fMRI time series, which are crucial for maintaining the statistical power of individual-level inferences or group analyses based on hierarchical models (Hodges and Sargent, 2001).

The amount of PCA residual variance is determined by the dimensionality (i.e., model order) of the brain-masked decomposition (Fig. S3). Increasing the order of dimensionality will cause a shrinkage of the PCA residual variance, which may lead to an increase of variance explained by both signal and noise components. By varying the dimensionality with an incremental multiplier sequence of the original MDL estimate, we found that the residual variance shrinkage primarily contributes to the variance explained by noise components. However, this increase of noise variance is systematically reduced as the multiplier increases above one. These result support the idea that MDL estimate is a reasonable trade-off between the amount of noise variance removal and the model order (which in turns determines the degree-of-freedom cost as well as the amount of computational time).

S3. Comparisons with existing fMRI methods for imaging overt speech production

The effectiveness of our technique was further demonstrated by comparing it to existing BOLD fMRI methods that have been used for imaging overt speech production. First, most of the existing methods are restricted to the use of slow event-related designs or short block durations with relatively low detection power (Birn et al., 2002) because the confounding temporal correlation between overt speech artifacts and hemodynamic responses increases with block duration (Birn et al., 2004; Soltysik and Hyde, 2006). Second, there are experimental paradigms specially designed to work around the severe motion during overt speaking. For example, the most commonly used method in the field employs sparse image acquisition (here we called *sparse*), in which images can only be collected during a period of silence following speech (Gracco et al., 2005). An alternative method allows continuous acquisition, but images collected during speech are simply discarded (Birn et al., 2004; here we called *discard*). The underlying assumption of both methods is that there is a brief non-overlapping period between the lagged hemodynamic response and the overt speech artifacts, which can be utilized to obtain “clean” images. However, this makes them suffer from the similar limitations of the scrubbing method that were mentioned earlier in the introduction, particularly the inability to obtain continuous fMRI time series. Therefore, developing an effective denoising technique is critical for overcoming these limitations, especially in a context of ecologically valid language research that is focused on connected speech production (Braun et al., 2001).

The methodological comparisons in this section were based on an ROI (BA 45) that was used previously for evaluating the impact of artifacts on activity detection. Both sICA denoising and the *discard* method (by the exclusion or “censoring” of images during all task blocks plus one image immediately following each block, which was applied only in the individual-level GLM stage after all the preprocessing steps, along with the removal of the corresponding parts of design matrix and serial correlation matrix derived from the uncensored data; for more details of this method, see Birn et al., 2004) were used to analyze data acquired in both short (10 s) and longer (30 s) blocks. Hence, the effects of these two different techniques, block duration, as well as their interactions can be examined in a single model (IS Fig. 2a). In the analysis of uncorrected data, the artifactual deactivation of the ROI was more severe in the 30 s production than in the 10 s production (t_{16} = 4.08, P_{adj} = 0.0044), as predicted by the differential temporal correlations between task periods (which contain artifacts) and hemodynamic responses in IS Fig. 2b. Notably, the ROI still showed apparent deactivation, though to a lesser degree, in the 10 s production blocks, which was nevertheless the opposite of the positive activation observed in PET (see Fig. 6d). These findings indicate that optimizing block duration alone cannot fundamentally solve the artifact problem. After sICA denoising however, both 10 s and 30 s production showed positive activations (10 s: t_{16} = 2.47, P = 0.0254; 30 s: t_{16} = 3.39, P = 0.0037). The activation of the 30 s production task was even (nonsignificantly) higher, probably due to a difference in detection power (Birn et al., 2002) between the two types of designs.

When compared to the direct analysis of uncorrected data, the *discard* method eliminated the artifactual deactivations for both 10 s and 30 s production blocks. However, its performance was inferior to the sICA method in terms of the magnitude of positive activation ($F_{1, 16} = 5.63$, $P = 0.0305$, pooled across 10 s and 30 s; IS Fig. 2a, the 2nd and 3rd groups of yellow and red bars). Furthermore, the combination of these two methods ("discard+sICA") showed no significant difference in percent signal change than estimates obtained using the sICA method alone. What is surprising is that the sICA method further increased the signals of the production tasks even after the noisy images acquired during task periods were discarded ($F_{1, 16} = 10.46$, $P = 0.0052$; IS Fig. 2a, the 3rd and 4th groups of yellow and red bars). If this is due to persistent motion artifacts observed in the post-task periods, then the basic assumption of both the *sparse* and *discard* methods may be incorrect.

To resolve this question, we obtained the FD and DVARS measures during and after each narrative block using the methods proposed by Power et al. (2012). Both FD and DVARS revealed a very striking motion "aftereffect" during the post-task periods of narrative production (IS Fig. 2c/d), i.e., both bulk head motion and noise-induced intensity fluctuations failed to fall back to the baseline level immediately after the task. The results of DVARS also showed a transient burst at approximately two seconds after production, which was even higher than the mean measurement during task ($t_{166} = 4.33$, $P_{\text{adj}} = 0.0033$, pooled across 10 s and 30 s). However, after sICA denoising, the DVARS curves for production appeared to be flat, with their means lowered approximately to (and slightly lower than) the range observed during comprehension tasks. This indicates that the sICA method can reduce the artifacts of all acquired images – during and after speech production – to at least a level similar to that seen during conventional low-noise tasks.

The exact mechanisms underlying the above post-task motion are still not clear to us. They are likely caused by deep breathing and/or swallowing after a period of continuous speaking, both of which may result in bulk head motion (Seto et al., 2001) as well as image intensity changes (Birn et al., 1998; Glover et al., 2000). Nevertheless, these measurements may indeed explain the suboptimal performance of the *discard* method.

In addition, although the *sparse* method was not formally investigated in this study, we expect it may suffer from the same problem when subjects are required to maintain speaking in separate but consecutive task epochs (Gracco et al., 2005), especially since sparse images are typically acquired in 2-3 second gaps between these epochs, during which deep breathing and swallowing are most likely to occur. Hence, the FD and DVARS profiles observed here may argue against the benefits of using this paradigm as well.

Taken together, the above evidence suggests that the denoising performance of sICA is not constrained by block duration, nor does it require the discarding of noise-contaminated images. Importantly, the ability to use a task period of at least 30 seconds is critical for obtaining a more natural narrative structure in discourse-level production. Besides, the flexibility for subjects to choose variable speaking durations is a necessity in studies of conversation (Scott et al., 2009), required by the normal turn-taking behavior during such interactions. Moreover, retaining all of the consecutive images within task blocks is mandatory for evaluating hemodynamic fluctuations that correlate with cognitive-behavioral events occurring during longer production tasks. All of these may have significant implications for unraveling the neural bases of a variety of clinical disorders including autism, stuttering, schizophrenia, aphasia, traumatic brain injury and Alzheimer's disease, in which symptoms may emerge only in the above contexts (Bloom et al., 1994).

S4. A brief comparison with prospective motion correction

In parallel to the research on retrospective artifact removal methods (including our sICA-based denoising technique), recent advances in prospective motion correction have made it possible for correcting rigid-body head motion by real-time adjustment of the imaging pulse sequence based on pose information obtained from external tracking mechanisms (Maclaren et al., 2013). There are clear advantages for prospective correction as it directly improves the quality of collected data by preventing certain artifacts from being generated, e.g., spin-history effects (Friston et al., 1996). In our opinion, the prospective and retrospective methods are complementary rather than competitive because they have quite different applications. While prospective correction provides unique benefits for reducing the effects of rigid-body motion, our denoising technique can deal with a much broader range of artifacts including complex and non-rigid motion (e.g., artifacts generated by articulator or eye motion), physiological noise, and residual head motion effects caused by the perturbation or susceptibility distortion of the magnetic field. Moreover, there are still numerous technical challenges for the wide adoption of prospective correction as an essential tool of MRI (Maclaren et al., 2013). Our denoising technique, besides being more broadly applicable, is on the contrary immediately available and can be applied to data that have already been collected, e.g., the huge fcMRI database of the Human Connectome Project.

Supplementary Data

Fig. S1. Dimensionality estimation for the head-masked sICA.

Line plots depict the relationships between the spatial correlation of head-masked (HM) and brain-masked (BM) independent component pairs (ICPs), the classification scores of the automated independent component classifier (AICC), and the dimensionality multiplier between the HM and BM sICA. The classification specificity (black) of AICC was crucially affected by the dimensionality multiplier: at a multiplier of one, the specificity is under 0.85, with a significant number of signal components misidentified as “noise” (i.e., false positives); the specificity increases with the multiplier and reaches a plateau at a critical point near 2.5. The classification sensitivity (gray) decreases only very slightly as the value of the multiplier increases. The relationship between the multiplier and AICC specificity was best predicted by the minimal spatial correlation between the HM and BM ICPs for signal components (red). When the ground truth classification scores established by human experts are not available, an optimal multiplier can be predicted by the mean spatial correlation of the “signal” ICPs identified by AICC (blue), whereas the mean spatial correlation of all ICPs (yellow) may lead to an underestimate. The optimal multiplier value may vary under different image acquisition settings. A reduced extracerebral coverage may decrease this value, whereas severe Nyquist ghost artifacts may increase this value. However, given the relationships between the multiplier and the classification sensitivity/specificity, using a higher value as a conservative guess can usually give fairly reliable classification results at the cost of computational efficiency.

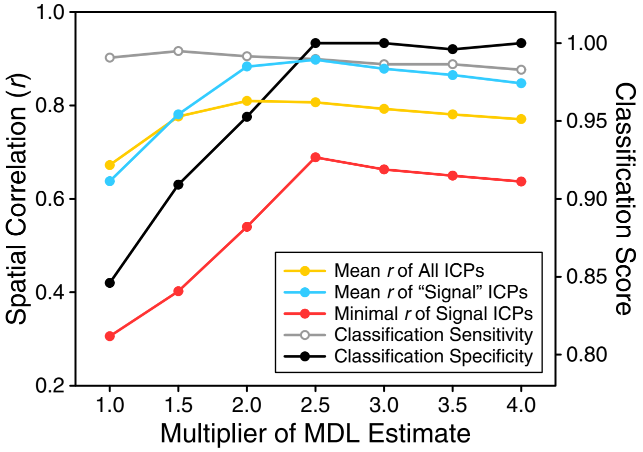

Spatial Correlation (*r*)
1.0
0.8
0.6
0.4
0.2
Classification Score
1.00
0.95
0.90
0.85
0.80
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Multiplier of MDL Estimate
Mean *r* of All ICPs
Mean *r* of "Signal" ICPs
Minimal *r* of Signal ICPs
Classification Sensitivity
Classification Specificity

Fig. S2. Effects of multiple runs for the Head-Masked sICA.

Line plots depict the increases of spatial correlation between matched HM and BM components with the number of runs. Data points indicate mean spatial correlation across all ICPs. After an initial steep increase, the mean spatial correlation increases more slowly with the number of runs. In addition, ICASSO can slightly increase the mean spatial correlation if matching is confined into the most stable run as compared to a random single run. But the benefit appears to be much less than that provided by matching across runs since the goal is to maximize the matching between BM and HM components rather than the stability of HM decomposition itself. Across a given number of runs, the convergence of mean spatial correlation is actually slightly slower if the runs are sorted by the stability metric (gray), as compared to randomly ordered runs (black). This is because there is less random variability for the components contained in runs with higher stability metrics than those with lower stability metrics.

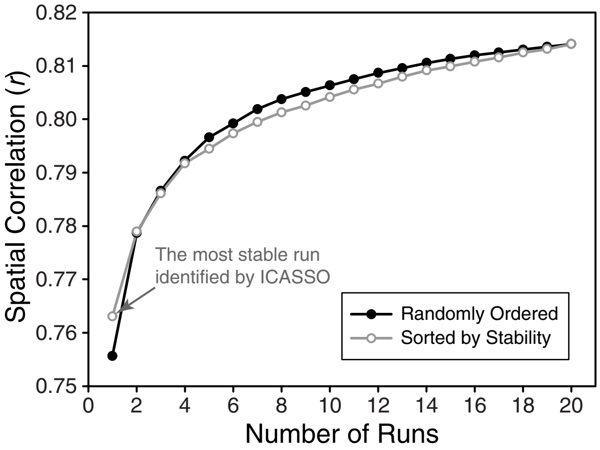


Fig. S3. Effects of source dimensionality on denoising.

Stacked bar charts depict the variation of temporal variance partitions across different orders of source dimensionality. The dimensionality of brain-masked sICA was varied systematically by applying a linearly incremental multiplier sequence of the original MDL estimate. Bars of each color indicate the mean proportions of variances across 18 datasets. Note that the total stacked variance of each multiplier group is slightly smaller than one. This is due to the existence of small temporal covariance between signal and noise components.

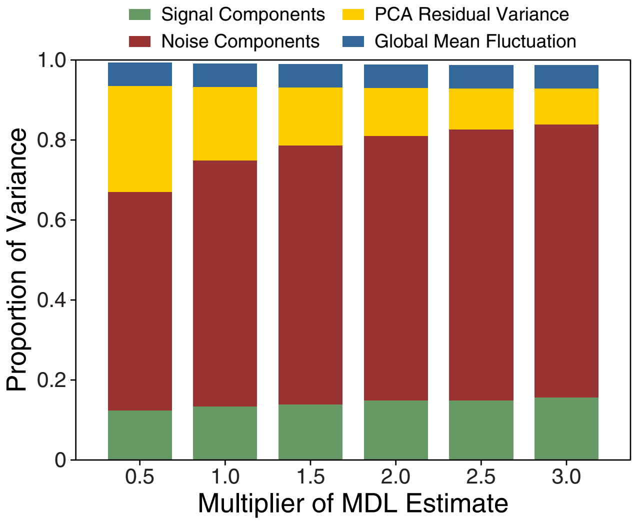


Fig. S4. In-plane and off-plane correlations of head motion components.

Bar charts depict the means and standard errors of Fisher’s z'-transformed correlation coefficients between the time courses of head motion components and the motion parameters. The number of components (*N*) identified for each type of head motion is annotated under the corresponding label. The set of in-plane parameters include the rigid-body alignment parameters and their first-order Volterra expansion in anterior-posterior, inferior-superior and pitch directions. The set of off-plane parameters are derived from the left-right, roll and yaw directions. There are 12 time series in each set for each fMRI run. The correlation coefficient of each component time course with each set of motion parameters is represented by the maximum absolute value among all the 12 correlations and across four runs. Asterisks indicate the significance levels of Tukey-Kramer honestly significant difference test: * $P_{adj} < 0.05$.

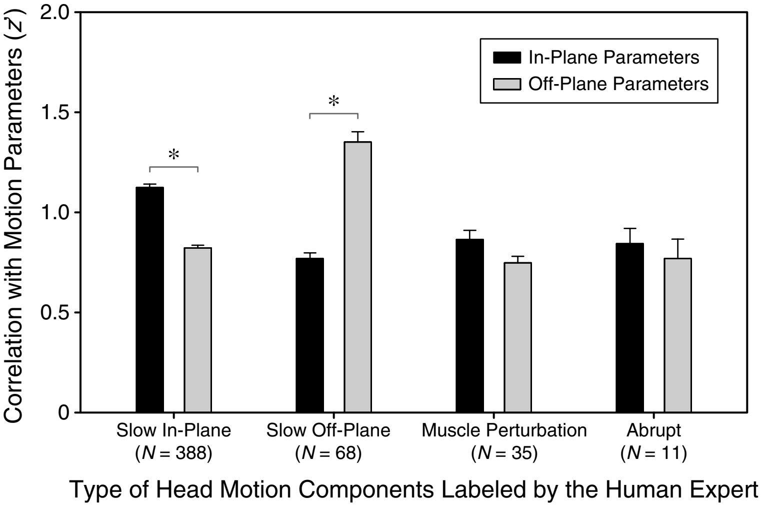
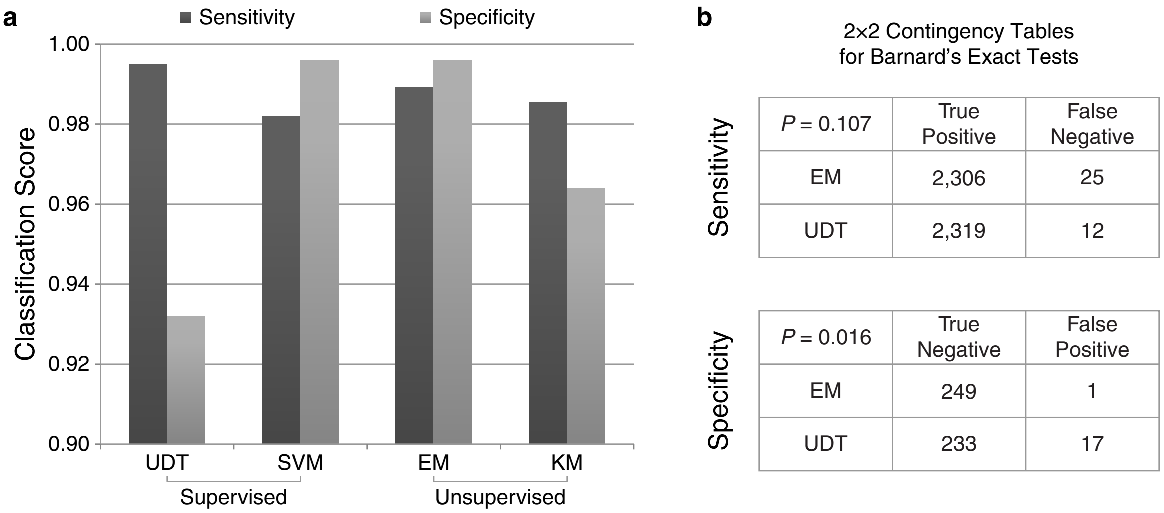


Fig. S5. Generalizability of different machine learning algorithms to novel datasets.

(a) Bar charts for the classification scores tested on 2,581 independent components decomposed from 22 resting-state datasets (Power et al., 2012). All the four classifiers utilized the same set of spatial features selected by our performance criteria. The two supervised classifiers, based on univariate decision tree (UDT) and support vector machine (SVM) respectively, were trained on all the components of our speech production datasets and then applied on these testing datasets. The two unsupervised classifiers, based on expectation maximization (EM) and k-means (KM) respectively, were applied on the testing datasets directly.

(b) Exemplar 2×2 contingency tables of sensitivity and specificity for comparing the performance of different leaning algorithms using Barnard's exact tests. The classification scores of the EM algorithm employed in our technique were compared with the other three algorithms. The specificity of UDT is significantly lower than EM. No significant differences were observed for the other tests. However, the overall performance of EM appeared to be the most optimal for a balanced consideration of both sensitivity and specificity. The performance of SVM is close to EM, but EM is more advantageous in that it is completely automatic and independent of pre-labeled data.



$P = 0.107$	True Positive	False Negative
EM	2,306	25
UDT	2,319	12

$P = 0.016$	True Negative	False Positive
EM	249	1
UDT	233	17

Fig. S6. Region of interest (ROI) defined from PET activation.

A conjunction map between production and comprehension was computed by taking the minimal *t*-values (Nichols et al., 2005) from the group-level contrast maps comparing narrative vs. pseudoword. A watershed transform (Meyer, 1994) was applied to segment the large left perisylvian cluster (encompassing frontal and temporal lobes) seen in the conjunction map into discrete regional clusters. The ROI (red) selected to evaluate the effectiveness and specificity of denoising was an inferior frontal cluster centered in BA 45.

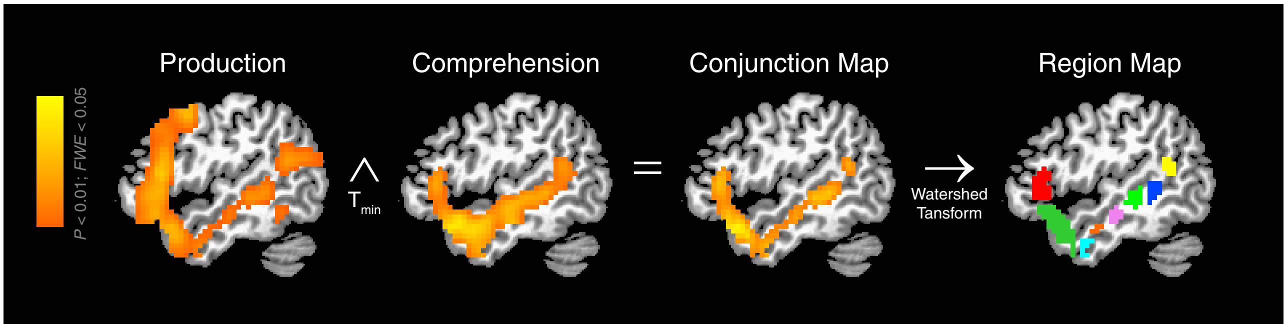


Table S1. Distributions of components for different noise categories and spatial features.

Values indicate the number of components above the threshold for each feature. There is a large redundancy between the four features so that a single noise component may meet the thresholds of more than one feature. The “unique” count in the table indicates the number of components that only meet the threshold of one specific feature. Based on both the “total” and “unique” counts, the *out-of-brain ratio* is the most effective feature for detecting articulator and eye motion; the *scattering degree* is the most effective for detecting both head motion and physiological noise; whereas most components in the “other” category need to captured by the template matching method.

	Components Identified by Each Feature (Total / Unique)				Total Identified / Human Expert
	Out-of-Brain Ratio	Scattering Degree	Slice-Wise Variation	Template Match	
Articulator Motion	235 / 9	229 / 3	187 / 2	21 / 4	255 / 255
Head Motion	224 / 31	471 / 209	90 / 0	15 / 1	504 / 507
Physiological Motion	202 / 1	300 / 31	234 / 6	1 / 1	311 / 316
Eye Motion	37 / 1	26 / 0	26 / 0	19 / 0	38 / 38
Other Structured Noise	1 / 1	0 / 0	1 / 0	47 / 46	48 / 50
All Noise Components	699 / 43	1026 / 243	538 / 8	103 / 52	1156 /1166

Table S2. Performance measures of AICC spatial features on resting-state datasets.

The ground truth classification (2,331 noise components, 250 signal components) for computing the sensitivity index and identified noise component counts (total, unique, false positive) is a set of human expert ratings obtained by detailed examination on each component using our mechanistic classification scheme. The identified noise component counts were derived from the classification results of AICC.

	Sensitivity Index	Bimodal Coefficient	Total Identified	Uniquely Identified	False Positives
Out-of-Brain Ratio	1.60	0.607	1,070	46	0
Scattering Degree	4.66	0.812	2,182	710	0
Slice-Wise Variation	1.41	0.588	908	7	0
Template Match	0.14	0.706	74	26	1
Total	–	–	2,306	–	1

Table S3. Noise components matched to flexible templates.

The same set of flexible templates, created by a secondary sICA on the components decomposed from the speech datasets, were applied in the classification of the resting state datasets. The total component count was the sum of all matched noise components for each type of template across datasets. The unique count was the sum of all matched noise components that were missed by the other spatial features. The ground truth count was the sum of the components labeled by a human expert as "dural venous sinuses" or "ventricles" during mechanistic classification. The detection rate was computed by the percentage of components conjointly identified by the template and the human expert among the ground truth count.

Templates	Speech Datasets				Resting-State Datasets			
	Total	Unique	Ground Truth	Detection Rate	Total	Unique	Ground Truth	Detection Rate
Dural Venous Sinuses	18	17	18	100%	18	12	18	88.9%
Ventricles	11	11	12	91.7%	16	7	13	100%
Other Templates	56	6	–	–	20	3	–	–

Supplementary Appendices

A. A reference measure for validating the removal of overt speech artifacts

Overt speech artifacts have imposed significant limitations on the application of BOLD fMRI to the research of language production (Price, 2010). For this reason, a majority of fMRI studies in this field have used covert speech, i.e., silent speech without actual movement of articulators. A covert speech production task is not a good reference for the purpose of validation in our study because there are distinct neural correlates between overt and covert speech (Barch et al., 1999; Huang et al., 2002).

The overt production of single words or brief individual sentences have been investigated previously by using slow event-related designs (Birn et al., 1999; Gopinath et al., 2009), sparse image acquisitions (Abrahams et al., 2003; Gracco et al., 2005), or by discarding images during short blocks (Birn et al., 2004). However, there are no validated BOLD imaging methods for working around the severe artifacts during continuous (or “connected”) overt speech production. Recent fMRI studies based on arterial spin labeling (ASL) perfusion contrast (Kemeny et al., 2005; Troiani et al., 2008) appeared promising, but may still be limited by the low intrinsic signal-to-noise ratio of this technique (Calamante et al., 1999).

Therefore, questions related to continuous overt speech production have been conventionally studied using positron emission tomography (PET; Awad et al., 2007; Blank et al., 2002; Braun et al., 2001; Brownsett and Wise, 2010), which is still considered the gold standard up to this point (Horwitz and Simonyan, 2014). In fact, the lack of susceptibility-related artifacts in PET not only makes it a convenient reference for validating the functional activity in language production, but also provides certain benefits for other cognitive tasks (Devlin et al., 2000; Frey and Petrides, 2002;

Schacter and Wagner, 1999), especially in the requirement of imaging regions near air-tissue interfaces (Ojemann et al., 1997).

However, the fMRI and PET images derived from individual-level analyses are not directly comparable even though in both cases, values have been converted into percent signal change. This is due to an intrinsic difference in image contrast, where PET is much higher than fMRI in terms of both signal range and standard deviation (Ramsey et al., 1996). In order to obtain a quantitative comparison, we converted image intensities into a standardized normal variate, called *standardized signal change* (Z_s). Since the original percent signal change images had global means very close to zero (due to global signal removal for fMRI, and proportional scaling with baseline subtraction for PET), this procedure essentially applied a common scale factor (*SF*) to all images within each imaging modality, i.e.,

$$Z_s = I \times SF = I/PSD \approx I/I_{rms}^{G},$$

where *SF* was the inverse of the pooled standard deviation (*PSD*) of image intensities across subjects and tasks (for fMRI, only including tasks with 30 s block length). Again, due to a near zero global mean, *PSD* is approximately equal to the grand root mean square of image intensities (I_{rms}^{G}) across all voxels and images for all subjects and tasks. For fMRI, this computation was also pooled across denoised and uncorrected datasets.

The above Z_s method for standardizing image intensities shares the same goal but has a different formula from the z_t method proposed by Ramsey et al. (1996). The major difference is in the computation of *PSD*. The z_t method requires two calculation steps – first computing the voxel-wise standard deviations across images then pooling across voxels to compute *PSD*. The Z_s method only involves a single step of pooled calculation.

Although we emphasize on the importance of using PET as a reference measure for cross-modal validation, it is by no means a “better” imaging technique. Obtaining reliable fMRI measurements can provide significant advantages that are absent in PET, e.g. the ability to perform event-related analysis and functional connectivity analysis. This is the exact reason for the paramount importance of advancing fMRI methods for imaging language production.

B. Reliability benefits of multiple sICA runs

Given the stochastic nature of ICA algorithms (Himberg et al., 2004), every single run of decomposition performed on the same dataset under different random initial conditions produces a slightly different set of components (note that this stochasticity in decomposition should not be confused with the deterministic source mechanism represented by each resulting component). Although the fMRI time series reconstructed after denoising are usually minimally affected by the minor variations across ICA runs, optimizing the stability of source decomposition is still preferable in order to reduce the likelihood of picking up occasional outlier components. This is analogous to take the median value of other quantitative measures to reduce random errors.

The most stable run containing the BM components was selected from among 20 runs by a modified ICASSO algorithm implemented in GIFT. The original algorithm (Himberg et al., 2004) determines the final component maps using the centroids of component clusters across multiple runs. However, the computation of the mixing matrix can occasionally lead to erroneous results due to multicollinearity when dimensionality is high, as is the normal case in sICA-based denoising applied on individual subject data. The modified algorithm still uses the results of a single run, which is considered “stable” based on the maximization of a metric computed by summing the similarity (absolute value of the Pearson's correlation coefficient) between each component map and the centroid of its belonging component cluster across runs.

For each BM component chosen, a best-matched HM component was selected from a repertory of all HM components pooled across 10 runs. Because the best-matched HM component derived from each individual run varies slightly according to the random initial conditions, the degree of spatial matching can always be enhanced by selection from an increased number of matched candidates across multiple runs. The optimal run number was determined by a balance between the degree of matching and computational efficiency (see Fig. S2).

C. Computational theory of automated component classification

C.1. Spatial features

Among the various types of objective measures for automated component classification, spatial features provide the best generalizability in the following two aspects. First, classifiers based on spatial features are most robustly generalizable to different experimental designs – either resting-state or blocked/event-related task-based designs – due to their complete independence from the temporal structure of these paradigms. Second, classifiers based on spatial features should also be generalizable to different image acquisition parameters and field strength. This is because sICA detects consistent and statistically independent spatial patterns embedded in fMRI images, which are essentially driven by the coherent temporal relationships between voxels located in the same functional networks (Calhoun et al., 2008) or contaminated by the same source of artifacts (Turner and Twieg, 2005). Hence, the spatial distributions of the resulting components are minimally affected by the geometric parameters of imaging voxels (De Martino et al., 2011), but depend on the dimensionality of decomposition (Turner and Twieg, 2005).

In contrast, classifiers based on temporal correlation with motion measures (computed by a univariate general linear model, GLM; Rummel et al., 2013) may be effective for resting-state or event-related designs, but most likely inaccurate for block designs, during which motion and task-related effects may be significantly correlated (Johnstone et al., 2006; cf. Fig. 1c and Table 1). Kochiyama et al. (2005) proposed an alternative method to univariate GLM based on the difference in heteroscedasticity of temporal variance between motion- and task-related components. However, this method was only tested on head motion engineered by a pneumatic system in a block design, and may be problematic for naturally occurring head motion in real tasks, especially in event-related designs. In addition, spectral features derived from component time courses are also very limited in their practical use. Unless the fMRI time series are collected with a very fast temporal resolution by imaging only a few slices (Thomas et al., 2002), the physiological noise components are heavily aliased (cf. Fig. 4a/b); and in general there is no single frequency band in which the aliasing happens (Lund et al., 2006). On the other hand, components related to transient event-related neural responses may be falsely identified as high-frequency noise on the basis of spectral features (De Martino et al., 2007). In addition, separate feature design and training process might be required by different experimental designs when temporal or spectral features were utilized (Tohka et al., 2008), thus significantly compromising the generalizability of the classifier.

Due to the above reasons, the automated independent component classifier in our technique employs only spatial features. These measures are based on either the head-masked (HM) or the brain-masked (BM) component map, both derived from the dual-mask sICA method. Because the source signals from ICA are in arbitrary units, the HM and BM component maps were first scaled into *Z*-scores with the standard deviation of in-mask voxel intensities as unit.

Since these maps are mean-centered, the measurement of intensity distributions should not be based on the first order statistics, i.e., means (Tohka et al., 2008). Instead, we used sums of squared intensities for our primary features, ensuring that each measure (as a ratio between the intensities in a spatial partition and the overall intensity distribution) is normalized into a range between 0 and 1. In contrast, the ratio between two variances (i.e., sums of squares adjusted by their degrees of freedom; Tohka et al., 2008) is not normalized. It may vary significantly across datasets, and hence adversely affect the generalizability of the classifier. Furthermore, the ratio of sums of squared intensities should provide a more accurate measure of relative intensity distributions than the ratio of suprathreshold voxel counts (Bhaganagarapu et al., 2013) because the latter measure is not weighted by the magnitude of intensities.

Out-of-brain ratio measures the ratio of the sum of squared intensities, within major suprathreshold clusters, between the portion of voxels outside the brain and the entire portion of voxels in these clusters:

$$X_1 = \frac{\sum_{v \in MC \cap \overline{v \in B}} I_v^2}{\sum_{v \in MC} I_v^2},$$

where *Iv* represents voxel intensity in the head-masked (HM) component map; *MC* is the set of voxels belonging to major suprathreshold clusters; *B* is the set of voxels inside the brain. The purpose of selecting major suprathreshold clusters is to identify the primary sites of activations.

Instead of using an arbitrary extent threshold, the major suprathreshold clusters are identified in a heuristic manner: 1) all voxels in the HM were thresholded at $|Z| > 1$; 2) suprathreshold positive and negative voxels are clustered respectively based on a 6-connected neighborhood; 3) the voxels counts of all positive and negative clusters identified were further clustered into two classes, major and minor, using a one-dimensional k-means algorithm (Lloyd, 1982).

There are two seemingly related but different measures used in previous methods. Tohka et al. (2008) proposed a spatial feature that measures the relative intensity distributions between the set of voxels in the brain boundary and the set of voxels inside the brain. Bhaganagarapu et al. (2013)

measured the spatial extent of suprathreshold clusters overlapping with the brain boundary as relative to the volume of the boundary. An important advantage of our measure is that it utilizes the extracerebral spatial information provided by the dual-mask method.

Scattering degree measures the sum of squared intensities, within major suprathreshold clusters, between the portion of voxels with interspersed positive/negative values and the entire portion of voxels in these clusters:

$$X_2 = \frac{\sum_{v \in MC \cap v \in S} I_v^2}{\sum_{v \in MC} I_v^2},$$

where *S* is the set of voxels with at least 8-connected neighbors that have the opposite sign. This measure is computed from the HM component maps in order to capture the scattering patterns both inside and outside the brain.

This novel feature is different from the *degree* of *clustering* (De Martino et al., 2007) that measures the number of voxels belonging to the major clusters divided by the total number of suprathreshold voxels. Our scattering degree measure is more sensitive to artifactual intensity fluctuations due to image warping or in-plane head motion, in which the conventional clustering method often fails to separate interspersed but still spatially connected positive or negative voxels into small clusters.

Slice-wise variation measures the absolute difference of the sum of squared intensities between the voxels belonging to the odd (P_{odd}) and even (P_{even}) slices, divided by the sum of squared intensities of all voxels within the brain mask (*B*); a square-root transform is applied to reduce the skewness of the resulting distribution:

$$X_3 = \sqrt{\frac{\left|\sum_{v \in B \cap v \in P_{odd}} I_v^2 - \sum_{v \in B \cap v \in P_{even}} I_v^2\right|}{\sum_{v \in B} I_v^2}}.$$

This feature is sensitive to transient physical or physiological motion artifacts that only affect a single or a few interleaved slices. It was originally proposed by Tohka et al. (2008) with a slightly different formula. The measurement of slice-wise variation can be applied to either the HM or BM data. The latter was selected to improve the efficiency of computation.

Template match measures the maximal spatial correlation coefficient between a BM component map and a set of predefined noise templates:

$$X_4 = \max_{1 \le k \le N} \{\mathrm{corr}[I_c(v), I_t(v,k)]\}, (v \in B) \text{ ,}$$

where k is the index of N templates (N = 8 for the current study) including medial and lateral frontal air sinus (three templates), brain-edge signals for off-plane head motion, CSF signals at the cistern of great cerebral vein, dural venous sinuses, ventricles, and brain mask boundary (see Fig. 2, Panel 15-20); I_c contains the voxel intensity values of the component map smoothed by a Gaussian kernel of 3×3×3 voxels with a standard deviation of 0.65; I_t contains the voxel intensity values of the templates.

The above templates need to meet two common criteria. First, they are highly spatially clustered and can be consistent identified across datasets. Second, their spatial patterns have minimal overlap with the distributions of known functional neural networks. These templates, except for the brain mask boundary, were defined with the following procedure: 1) all BM component maps for each subject were transformed into a standard brain space by applying the spatial normalization parameters of the structural image and smoothed with a 6 mm FWHM Gaussian kernel; 2) the resulting maps for all subjects were concatenated and entered into a low-dimensional, second-level sICA (with an estimated source dimension of 30 in the current study) in order to identify common spatial patterns across the first-level components; 3) target noise components were selected from the second-level component maps based on visual inspection; 4) the selected component maps were inversely transformed to each subject's native brain space to create individualized templates.

Template matching methods have been used in previous studies for automated identification of signal components in individual-level sICA (Greicius et al., 2004), and noise components in group sICA (Sui et al., 2009). A unique property of our procedure is that the templates are defined in a data-driven manner to improve their spatial matching with the actual components. In addition, in the case of requirement for new templates, they can be generated on demand with the same procedure from new learning datasets and appended to the current set, which further increases the flexibility and general applicability of our classifier.

C.2. Feature selection criteria

Feature selection is critical for the optimal performance of an automated component classifier. De Martino et al. (2007) examined the percent loss of overall classifier performance after the removal of each individual feature. However, this measure is affected by the performance of other features in the classifiers and can also be biased by the variability of thresholds determined by different learning algorithms. Here we propose an individual feature selection scheme for the binary signal/noise classification of sICA components. This scheme is based on two criteria that evaluate the performance of each feature independent of the learning algorithm and other features included in the classifier.

The first criteria is called sensitivity index (*SI*), which measures the separability of signal and noise, in a set of components labeled by our mechanistic classification, according to their distributions of a given feature value:

$$SI = \frac{|m_S - m_N|}{\sqrt{(s_S^2 + s_N^2)/2}},$$

where m_{S} and m_{N} are the means of labeled signal and noise components respectively; s_{S} and s_{N} are their standard deviations respectively. *SI* is closely related to the Fisher criterion score ($FCS = (m_1 - m_2)^2/(s_1^2 + s_2^2)$) that has been widely used for feature selection in other classification problems, e.g., microarray classification in genetics (Guyon et al., 2002). Compared to FCS and its several variants (Furey et al., 2000; Golub et al., 1999; Maldonado et al., 2011), the value of *SI* has a more direct mathematical relationship with the separability of signal/noise distributions, i.e., the distance between the two distributions (expressed as the absolute difference of means) as relative to their average width (expressed as the root mean square of standard deviations).

Note that the pooling between two sample standard deviations is not weighted by their sample sizes, unlike what is normally used in the statistical measures of effect size (e.g., Cohen's *d*; see Olejnik and Algina, 2000). This is in order to balance the effects between signal and noise because the latter usually contains a much larger number of components. Additionally, contrary to the usual requirement of Gaussian distributions by effect size measures, the use of *SI* should not be limited by the shapes of signal/noise distributions. This is analogous to a nonparametric form of detection sensitivity (often called d-prime, *d*', an essential measure in signal detection theory) derived from rating data (Simpson and Fitter, 1973).

The second criteria, *bimodality coefficient* (*BC*; Freeman and Dale, 2013; Pfister et al., 2013), is employed in our feature selection scheme for estimating the reliability of threshold detection that can be achieved for a given feature:

$$BC = \frac{m_3^2 + 1}{m_4 + 3 \cdot \frac{(n-1)^2}{(n-2)(n-3)}},$$

where n is the number of components in the sample set, m_3 is the sample skewness (i.e., the normalized third central moment) and m_4 is the sample excess kurtosis (i.e., the normalized fourth central moment minus 3). The value of *BC* lies between 0 and 1. A critical value for *BC* is 5/9 (≈ 0.555) that would be expected from a uniform distribution. Higher values indicate bimodality whereas lower values indicate unimodality. Based on the above formula, a heavy-tailed unimodal distribution (e.g., the distribution of template match values as illustrated in Fig. 5) is expected to have a large value of *BC* due to its high skewness. This is not surprising in that a heavy-tailed unimodal distribution is very close to a bimodal distribution with severely asymmetric peak amplitudes.

The bimodality of feature distributions directly impacts the reliability of unsupervised classification algorithms such as Gaussian mixture modeling and k-means clustering (Hellwig et al., 2010; Monti et al., 2003) since the existence of a sharp decision boundary provides an important practical advantage for binary clustering. For supervised classification algorithms, using bimodally distributed features can reduce the ambiguity of training. For example, the detection of a maximum-margin hyperplane by a support vector machine should be facilitated when bimodality is present in all or most of its features.

For each of the two performance criteria, we empirically define the acceptable ranges to guide their practical use in feature selection. For *SI*: $SI < 1$, low; $1 \leq SI \leq 2$, moderate; $SI > 2$, high. For *BC*: $BC < 0.5$, low; $0.5 \leq BC \leq 0.6$, moderate; $BC > 0.6$, high. In our scheme, a feature can only be selected if at least one of the criteria falls into the high range, or both criteria falls into the moderate range.

C.3. Machine learning algorithms

The automated independent component classifier (AICC) of our denoising technique is based on an unsupervised expectation maximization (EM) algorithm (Dempster et al., 1977) by fitting the distribution of each spatial feature (called a *mixture distribution*) to a Gaussian mixture model (GMM; see Fig. 5). The threshold of each feature was determined by the crossover point of posterior probability density (also called *responsibility*) functions of the two fitted Gaussian distributions, corresponding to the hidden distributions of the actual data. After applying the threshold, the ICA components were automatically clustered into two classes, one of which contained only noise components. This procedure can be performed on the components either within each individual dataset or across multiple datasets acquired using closely matched imaging parameters. The latter was used by the current study due to a better accuracy of threshold detection given the increased sample size.

Before running the EM algorithm, either separate or shared covariance matrices can be specified for the hidden distributions of GMM. When the mixture distribution contains two Gaussian-shaped and clearly distinguishable peaks, using separate covariance matrices usually detects the location of threshold more accurately. However, the threshold detection can be biased or unstable if the mixture distribution is highly skewed (e.g., in the case of a heavy-tailed uniform distribution) or lacks of a sharp decision boundary. This is likely due to violation of the Gaussian assumption for hidden distributions. For example, the assumption may be violated by the clipping of values at either end of the mixture distribution since the values of our features are all bounded between 0 and 1. In such cases (empirically defined by skewness > 1 or bimodality coefficient ≤ 0.6), a shared covariance matrix is used to obtain a more conservative estimate.

An alternative way for unsupervised binary clustering is to use the k-means (KM) algorithm, which has been utilized by a previous study on automated component classification (Bhaganagarapu et al., 2013). KM can be considered as a nonprobabilitic and limited version of the EM algorithm (Bishop, 2006). It is equivalent to EM when setting the covariance matrix to a diagonal matrix with equal, close-to-zero diagonal elements. An important advantage of EM is that it allows clusters to have different shapes; whereas KM is less accurate when the clusters are significantly unequal in size.

Besides the above unsupervised algorithms, we also investigated two commonly used supervised algorithms, which require a training procedure performed on a set of pre-labeled components. These include a univariate decision tree (UDT) algorithm based on the classification and regression trees (CART) technique (Breiman et al., 1984) and a multivariate support vector machine (SVM) algorithm based on the sequential minimal optimization (SMO) technique (Platt, 1999).

Unlike the fixed-structure decision tree used by our unsupervised classification method, the nodes in a UDT are automatically determined by the training process. While the decision rule of each node is still represented by a single feature (i.e., univariate), each feature can be used by multiple nodes with different cut-off thresholds. In order to prevent the induced classification tree from being over-complicated, which affects the generalizability of the classifier, a post-hoc pruning procedure was applied to control the depth of the tree. An alternative method to UDT, called global decision tree (GDT), was used in a previous study on automated component classification (Tohka et al., 2008). GDT is a multivariate, fixed-structure decision tree induction algorithm. That is, the decision rule of each node is represented by the combination of multiple features as a subset. However, such combinations are arbitrarily defined a priori and may not be suitable for our technique that employs only spatial features.

SVM is a multivariate classification method that separates the data into two different classes (in the case of binary classification) with a learned hyperplane in a multidimensional space represented by the combination of all features. The hyperplane, known as the *maximum-margin hyperplane*, is determined from the training data by maximizing the distances to the nearest data points (i.e., the *support vectors*) in each class. A variant of SVM, called least squares support vector machine (LS-SVM), was used in a previous study on automated component classification (De Martino et al., 2007).

Supplementary References

Abrahams, S., Goldstein, L.H., Simmons, A., et al., 2003. Functional magnetic resonance imaging of verbal fluency and confrontation naming using compressed image acquisition to permit overt responses. Hum. Brain Mapp. 20, 29-40.

Birn, R.M., Bandettini, P.A., Cox, R.W., et al., 1999. Event-related fMRI of tasks involving brief motion. Hum. Brain Mapp. 7, 106-114.

Birn, R.M., Cox, R.W., Bandettini, P.A., 2002. Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing. Neuroimage 15, 252-264.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer, New York.

Blank, S.C., Scott, S.K., Murphy, K., et al., 2002. Speech production: Wernicke, Broca and beyond. Brain 125, 1829-1838.

Bloom, R.L., Obler, L.K., De Santi, S., et al. (Eds.), 1994. Discourse analysis and applications: Studies in adult clinical populations. L. Erlbaum Associates, Hillsdale, NJ.

Breiman, L., Friedman, J., Stone, C., et al., 1984. Classification and regression trees. Chapman and Hall/CRC, Boca Raton, FL.

Brownsett, S.L., Wise, R.J., 2010. The contribution of the parietal lobes to speaking and writing. Cereb. Cortex 20, 517-523.

Calamante, F., Thomas, D.L., Pell, G.S., et al., 1999. Measuring cerebral blood flow using magnetic resonance imaging techniques. J. Cereb. Blood Flow Metab. 19, 701-735.

De Martino, F., Gentile, F., Esposito, F., et al., 2007. Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. Neuroimage 34, 177-194.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39, 1-38.

Devlin, J.T., Russell, R.P., Davis, M.H., et al., 2000. Susceptibility-induced loss of signal: comparing PET and fMRI on a semantic task. Neuroimage 11, 589-600.

Freeman, J.B., Dale, R., 2013. Assessing bimodality to detect the presence of a dual cognitive process. Behav. Res. Methods 45, 83-97.

Frey, S., Petrides, M., 2002. Orbitofrontal cortex and memory formation. Neuron 36, 171-176.

Furey, T.S., Cristianini, N., Duffy, N., et al., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906-914.

Golub, T.R., Slonim, D.K., Tamayo, P., et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537.

Gopinath, K., Crosson, B., McGregor, K., et al., 2009. Selective detrending method for reducing task-correlated motion artifact during speech in event-related FMRI. Hum. Brain Mapp. 30, 1105-1119.

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

Greicius, M.D., Srivastava, G., Reiss, A.L., et al., 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proc. Natl. Acad. Sci. U. S. A. 101, 4637-4642.

Guyon, I., Weston, J., Barnhill, S., et al., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389-422.

Hellwig, B., Hengstler, J.G., Schmidt, M., et al., 2010. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. BMC bioinformatics 11, 276.

Himberg, J., Hyvarinen, A., Esposito, F., 2004. Validating the independent components of neuroimaging time series via clustering and visualization. Neuroimage 22, 1214-1222.

Hodges, J.S., Sargent, D.J., 2001. Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika 88, 367-379.

Horwitz, B., Simonyan, K., 2014. PET neuroimaging: Plenty of studies still need to be performed: Comment on Cumming: "PET Neuroimaging: The White Elephant Packs His Trunk?". Neuroimage 84, 1101-1103.

Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129-137.

Maclaren, J., Herbst, M., Speck, O., et al., 2013. Prospective motion correction in brain imaging: a review. Magn. Reson. Med. 69, 621-636.

Maldonado, S., Weber, R., Basak, J., 2011. Simultaneous feature selection and classification using kernel-penalized support vector machines. Inform. Sciences 181, 115-128.

Meyer, F., 1994. Topographic distance and watershed lines. Signal Processing 38, 113-125.

Monti, S., Tamayo, P., Mesirov, J., et al., 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. 52, 91-118.

Nichols, T., Brett, M., Andersson, J., et al., 2005. Valid conjunction inference with the minimum statistic. Neuroimage 25, 653-660.

Ojemann, J.G., Akbudak, E., Snyder, A.Z., et al., 1997. Anatomic localization and quantitative analysis of gradient refocused echo-planar fMRI susceptibility artifacts. Neuroimage 6, 156-167.

Olejnik, S., Algina, J., 2000. Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations. Contemp. Educ. Psychol. 25, 241-286.

Pfister, R., Schwarz, K.A., Janczyk, M., et al., 2013. Good things peak in pairs: a note on the bimodality coefficient. Front. Psychol. 4, 700.

Platt, J., 1999. Fast training of support vector machines using sequential minimal optimization. In: Burges, C., Scholkopf, B., Smola, A. (Eds.), Advances in kernel methods: Support vector learning. MIT press, Cambridge, MA, pp. 185-208.

Price, C.J., 2010. The anatomy of language: a review of 100 fMRI studies published in 2009. Ann. N. Y. Acad. Sci. 1191, 62-88.

Pruthi, T., Espy-Wilson, C.Y., Story, B.H., 2007. Simulation and analysis of nasalized vowels based on magnetic resonance imaging data. J. Acoust. Soc. Am. 121, 3858-3873.

Ramsey, N.F., Kirkby, B.S., Van Gelderen, P., et al., 1996. Functional mapping of human sensorimotor cortex with 3D BOLD fMRI correlates highly with H2(15)O PET rCBF. J. Cereb. Blood Flow Metab. 16, 755-764.

Schacter, D.L., Wagner, A.D., 1999. Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. Hippocampus 9, 7-24.

Scott, S.K., McGettigan, C., Eisner, F., 2009. A little more conversation, a little less action--candidate roles for the motor cortex in speech perception. Nat. Rev. Neurosci. 10, 295-302.

Seto, E., Sela, G., McIlroy, W.E., et al., 2001. Quantifying head motion associated with motor tasks used in fMRI. Neuroimage 14, 284-297.

Simpson, A.J., Fitter, M.J., 1973. What is the best index of detectability? Psychol. Bull. 80, 481-488.

Soltysik, D.A., Hyde, J.S., 2006. Strategies for block-design fMRI experiments during task-related motion of structures of the oral cavity. Neuroimage 29, 1260-1271.

Sui, J., Adali, T., Pearlson, G.D., et al., 2009. An ICA-based method for the identification of optimal FMRI features and components using combined group-discriminative techniques. Neuroimage 46, 73-86.

Tohka, J., Foerde, K., Aron, A.R., et al., 2008. Automatic independent component labeling for artifact removal in fMRI. Neuroimage 39, 1227-1245.

Troiani, V., Fernandez-Seara, M.A., Wang, Z., et al., 2008. Narrative speech production: an fMRI study using continuous arterial spin labeling. Neuroimage 40, 932-939.

Turner, G.H., Twieg, D.B., 2005. Study of temporal stationarity and spatial consistency of fMRI noise using independent component analysis. IEEE Trans. Med. Imaging 24, 712-718.