Categorical Speech Representation in the Human Superior Temporal Gyrus

Edward F. Chang, Jochem W. Rieger, Keith Johnson, Mitchel S. Berger, Nicholas M. Barbaro, Robert T. Knight

SUPPLEMENTAL INFORMATION

Derivation of neuronal identification functions

Our approach to derive neuronal identification functions was inspired by a widely accepted definition of categorical perception (Harnard, 1987) which states that stimuli perceived as belonging to the same category are judged as perceptually more similar than stimuli belonging to a different category. Neuronal responses linked to categorical perception should show an analogous property: Responses to pairs of stimuli from within a perceptual category should be less distinguishable than responses to between category pairs. These considerations imply that in order to directly compare neuronal responses to categorical perception the neuronal responses must be transformed to a similarity metric as a first step. In this similarity based representation we then derived neuronal response classes and neuronal class prototypes. Finally, we used the neuronal class prototypes to calculate neuronal identification functions which were directly comparable to the psychophysically measured identification functions.

Confusion matrices

Confusion matrices provide information about the dissimilarity of neuronal responses elicited by pairs of stimuli, comparable to distance tables between cities provided in road maps.

$$\boldsymbol{D}_i = \begin{bmatrix} d_{1,1,i} & \cdots & d_{1,S,i} \\ \vdots & \ddots & \vdots \\ d_{S,1,i} & \cdots & d_{S,S,i} \end{bmatrix}$$

We used single trial classification to measure the pair wise dissimilarities $d_{s_1,s_2,i}$ at different time intervals i to fill the confusion matrices $\boldsymbol{D}_i$. In this approach, neuronal response dissimilarity is measured as the proportion of correct single trial classifications of brain responses obtained with a specific pair of stimuli. Higher entries in the confusion matrix indicate larger distances between pairs of brain responses in the dissimilarity space we aim to derive.

Classification

We estimated pair wise dissimilarities $d_{s_1,s_2,i}$ using an L1-norm regularized logistic regression classifier (Koh et al, 2007) applied to the time series data in a leave-one-trial-out cross validation procedure. L1-norm logistic regression is well suited for classification problems involving high dimensional feature spaces and relatively few examples for training because it does not overfit easily when the ratio of training data samples to feature space dimensions is low (for a discussion see e.g. Koh et al., 2007). Overfitted classifiers achieve only low prediction rates on new data in cross-validation tests and would therefore lead to unstructured dissimilarity matrices with similar entries.

We used the time series data for classification because they contain all information available in the data and hence require no prior hypothesis about what features of the neuronal response could be important for categorical neuronal coding in STG. Since we had no prior assumption about the time when categorical neuronal responses occur, we derived neuronal confusion matrices in steps of 10 ms from 40 ms long intervals. The 40 ms time window was a compromise between temporal resolution and the dimensionality of the feature space. The full feature space included up to 16002 dimensions (63 channels * 254 time samples), whereas the 40 ms included only up to 1260 dimensions (63 channels * 20 time samples).

In the cross-validation loop both feature selection and classifier training was performed on a subset of $Q = R - 1$ trials (leave-one-out cross validation). In the feature selection step we discarded those samples in the time series that did not exhibit any indication of difference between two stimulus conditions. This first univariate feature selection was done by calculating for each sample a t-value over stimulus repetitions (indicated by the '.')

$$t_{c,t,s_1 vs\ s_2} = \frac{\bar{z}_{c,t,.,s_1} - \bar{z}_{c,t,.,s_2}}{\sqrt{\frac{s^2_{z_{c,t,.,s_1}} + s^2_{z_{c,t,.,s_2}}}{Q}}}$$

and discarding those samples that did not pass a liberal criterion of $|t_{c,t,s_1 vs\ s_2}| < t_{crit}$. The criterion was chosen individually to maximize the classifier's peak classification performance. However, the optimal criterion was very similar among participants (P1: $t_{crit} = 2.5$, P2: $t_{crit} = 2.0$, P3: $t_{crit} = 2.0$, P4: $t_{crit} = 2.0$). The samples that survived this initial selection were assembled for each single trial in a feature vector $\vec{x}$ holding a total of F features (samples). These lower dimensional feature vectors were used for classification.

The classifier's logistic model is:

$$p(b|\vec{x}) = \frac{e^{b(\vec{w}^T\vec{x}+v)}}{1 + e^{b(\vec{w}^T\vec{x}+v)}}$$

Here, $b \in \{-1,1\}$ denotes the class a trial belongs to, v is an intercept, and $\vec{w}^T\vec{x} + v = 0$ defines the separating hyperplane. The optimization problem solved in L1 norm regularized logistic regression is to minimize:

$$\frac{1}{Q}\sum_{i=1}^{Q} log\ (1 + e^{-b_i(\vec{w}^T\vec{x_i}+v)}) + \lambda \sum_{f=1}^{F} |w_f|$$

The additive term controls, via the regularization parameter λ, the sparseness of the decision subspace by penalizing non-zero entries in $\vec{w}$, the normal vector on the decision hyperplane. The L1 norm optimization of the classifier's decision hyperplane sets the weights of a large number of less informative features to zero, thereby excluding them from the solution. As a consequence, the actual classification is performed in an informative subspace of much lower dimensionality than the original data space. This regularization can be understood as a second feature selection step (t-value feature selection was the first) that helps to prevent overfitting the classifier to the training set. The regularization parameter λ was held fixed at 0.1.

The single left out trial $\vec{x}$ was classified with the optimized $\vec{w}$ and v in:

$$sign(\vec{w}^T\vec{x} + v)$$

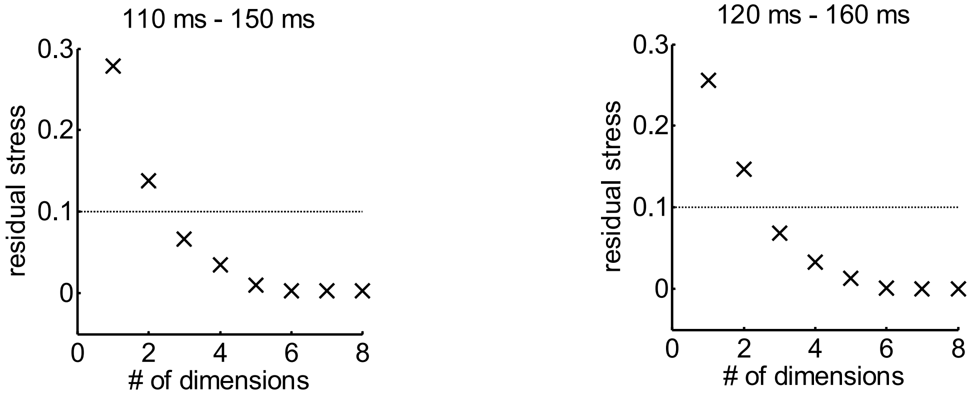
This procedure was repeated R times and the proportion of correctly classified trials was used as the estimate of the dissimilarity $d_{s_1,s_2,i}$ of the neuronal responses to a given stimulus pair in a given interval. To further increase the ratio of the number of examples to the number of features we combined the neurophysiological measurements of adjacent stimuli (e.g. 1&2; 2&3 etc.) yielding a total of $4R = 100$ trials used per dissimilarity estimate $d_{s_1,s_2,i}$. Note that labels in the figures of the main paper list only the first stimulus in these combined sets of trials. With respect to the goal of our analysis combining responses to adjacent stimuli could smooth category boundaries somewhat.

Multidimensional scaling

In the next step we applied metric multidimensional scaling (MDS) to the confusion matrices averaged over subjects. MDS is a method for embedding objects, in our case neuronal responses, in a low dimensional Euclidian space such that distances between the objects reproduce an empirical matrix of dissimilarities, in our case $\boldsymbol{D}_i$, as well as possible. MDS has been suggested to be useful in psychophysical research to analyze categorical perception (Shepard, 1980), and has subsequently been used to study e.g. face perception (Bimler & Kirkland, 2001) and the perception of phonetic stimuli (Shepard, 1980, Iverson & Kuhl, 1995). The objective in MDS is to minimize the reconstruction error measured by Kruskall Stress (Kruskall & Wish, 1978):

$$stress_i = \sqrt{\frac{\sum_{n=1}^{13}\sum_{j>n}^{13}(\delta_{n,j,i} - d_{n,j,i})^2}{\sum_{n=1}^{13}\sum_{j>n}^{13} d_{n,j,i}^2}}$$

The $d_{n,j,i}$ are the neuronal response dissimilarities classification revealed for stimulus pairs s_n, and s_j in interval i. The $\delta_{n,j,i}$ are the corresponding distances between responses in the Euclidian embedding constructed by MDS. The summation is over 13 responses because responses to adjacent stimuli were pooled, as noted above. The MDS embedding was calculated in three dimensions. This choice was based on a priori considerations of how many dimensions would be maximally required. If the neuronal responses in STG parallel the perceptual domain, comprising three phonemes, then three dimensions should be sufficient for MDS to construct a configuration of responses reflecting the pairwise dissimilarities. Moreover, in the case that neuronal responses reflect the linear physical variation of the speech stimuli a single dimension should already suffice to reconstruct the linear configuration. Supplemental figure 1 shows residual Kruskal Stress as a function of the number of dimensions in the two the two consecutive intervals 110 ms - 150 ms, and 120 ms - 160 ms after phoneme onset. The reconstruction should exceed 90% correspondence with the original dissimilarities (stress ≤ 0.1, Kruskal & Wish, 1978). This criterion is reached using three dimensions. This more formal analysis confirms that our a priori choice was adequate. The MDS reconstructions for the critical time intervals between 110 ms and 160 ms result in stress values of 0.066 (110 ms to 150 ms), and 0.068 (120 ms to 160 ms).



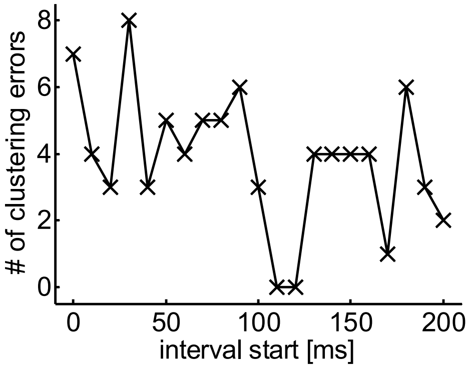
Supplemental figure 1. Residual stress at two intervals plotted as a function of the number of dimensions included in the MDS solution. Three dimensions are sufficient to obtain a solution with residual stress < 0.1.

K-means clustering

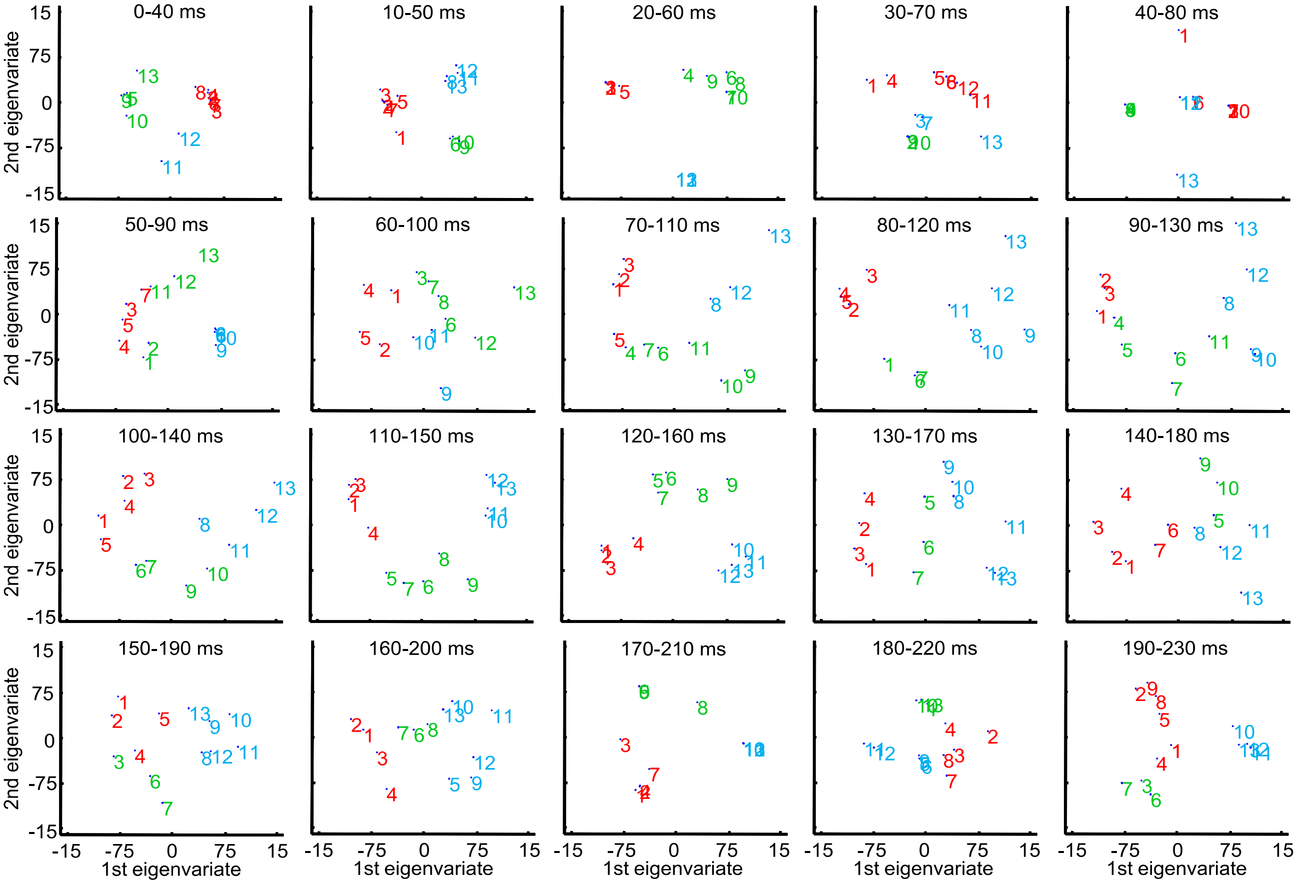
The simultaneous representation of all neuronal responses in on common similarity space allowed us to use cluster analysis (Shepard, 1980) to test when, if at all, neuronal responses group in a way that parallels perceptual grouping obtained psychophysically. We used K-means clustering because it implements the definition of categorical perception. K-means finds a partitioning of the data into k cluster that minimizes the within cluster sum of squares:

$$\sum_{i=1}^{k} \sum_{\vec{x}_j \in c_i} \left(\vec{x}_j - \mu_i\right)^T * \left(\vec{x}_j - \mu_i\right)$$

Because overall variance is constant, the between cluster variance is simultaneously maximized. Here, k is the pre-defined number of clusters, $\vec{x}_j$is the position of the neuronal response elicited by stimulus j in the MDS-representation, and μ_i is the center of gravity of cluster c_i. The obvious choice for the number of expected clusters was three, the number of perceived phonemes. We used the K-means algorithm implemented in matlab (The Mathworks Inc.) which requires as an additional restriction that each cluster must hold at least one neuronal response. Supplemental figure 2 shows the correspondence between the psychophysically obtained stimulus clusters and the cluster assignments obtained with K-means for the analysis time series. The two consecutive intervals from 110 ms to 150 ms, and from 120 ms to 160 ms, were the only where the psychophysically obtained categories and those derived from the neuronal responses were in exact correspondence. Supplemental figure 3 shows the full time series of MDS and K-means solutions in the same format as in the main manuscript.



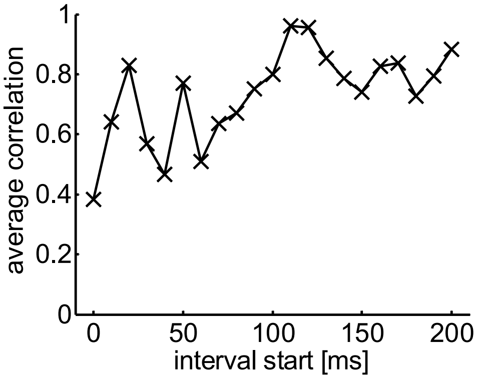
Supplemental figure 2. Only the neuronal responses in the intervals starting 110 ms and 120 ms after phoneme onset were classified exactly the way predicted by the psychophysical identification function. This is indicated by zero classification errors in these intervals. The probability for obtaining zero errors by chance is p<10^{-6}.



Supplemental figure 3. The full time series of MDS- and K-means solutions. The first two eigenvariates are shown.

Neuronal identification functions and correlation analysis

Cluster analysis is concerned with category membership but the results of it allowed us to calculate continuous neuronal identification functions analogous to psychophysical identification function. In our approach, we assume that participants compare stimulus percepts to phoneme prototypes in the identification task, and assign them to the most similar one. In this sense, the psychophysical identification function represents a measure of the distance between the phoneme prototypes and the stimulus percept. To derive the three neuronal identification functions we calculated three distance functions in MDS similarity space, one between each of the three cluster prototypes and all neuronal responses. These functions let us directly compare neuronal and psychophysical responses on a continuous scale. Supplemental figure 4 shows the time course of the mean of the correlations between the neuronal and psychophysical functions. Again, a clear peak is visible for the intervals between 110 ms to 150 ms, and 120 ms to 160 ms. During these intervals the correlations are excellent and average to 0.94 further corroborating our suggestion that stimuli along the /ba/-/da/-/ga/ continuum are categorically coded in the population response of posterior STG neurons.



Supplemental figure 4. Time course of mean correlations of neuronal and psychophysical identification functions.

Reconstruction of informative electrodes

The trained classifier’s weight vector quantifies the amount information each feature provides for classification. Highly informative features receive higher weights and features providing little or no information receive low or zero weights in the sparse classifier we employed. Features with zero entries in the weight vector do not contribute to the classification results. This can be easily seen in the decision function:

$$sign(\vec{w}^T\vec{x} + v)$$

which can be written as:

$$sign(w_1x_1 + \cdots + w_jx_j + \cdots + w_Jx_J + v)$$

The feature weights shown in Figure 3 (main manuscript) represent w_j averages over cross validation results and samples per electrode in the analysis interval. The average feature weights represent an estimate of how informative an electrode was judged by the classifier. Thus, the channels highlighted in Figure 3 can be considered as those providing most reliable information for discriminating the stimuli presented.

References:

Bimler D., and & Kirkland J. (2001) Categorical perception of facial expressions of emotion: Evidence from multidimensional scaling, Cognition & Emotion 15, 633–658

Harnad S. (ed.) (1987) Categorical Perception: The Groundwork of Cognition, Cambridge University Press, New York

Iverson P., and Kuhl P.K. (1995) Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling, J. Acoust. Soc. Am. 97, 553-562

Koh K., Kim S.-J., and Boyd S. (2007) An Interior-Point Method for Large-Scale l_1-Regularized Least Squares. Journal of Machine Learning Research 8, 1519-1555

Kruskal, J.B., and Wish M. (1978) Multidimensional Scaling. Sage Publications, Newbury Park

Shepard, R.N. (1980) Multidimensional Scaling, Tree-Fitting, and Clustering. Science 210, 390-398