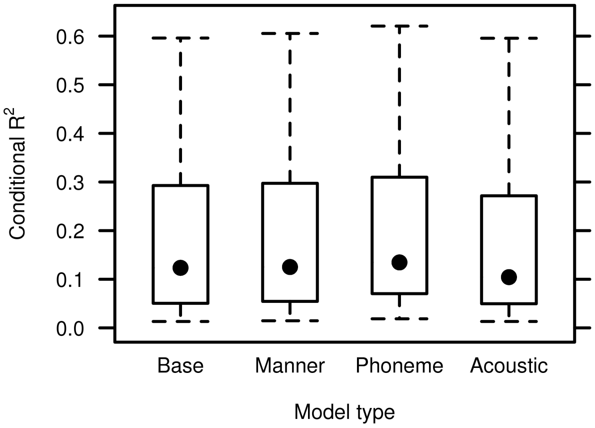
Supplement: Accounting for phonetic and acoustic tuning

The superior temporal gyrus (STG) is finely tuned to the spectrotemporal acoustic features of different speech sounds (Mesgarani et al. 2014). For example, some neural populations show strong responses to sounds with more energy in higher frequencies (characteristic of fricatives, like /s/), while other populations respond most strongly to energy in lower frequencies, characteristic of voicing (e.g., vowels). Because a large number of the electrodes in the present study recorded signals from the STG, it is important to ensure that the effects described above are not driven by acoustic- or phonetic-level properties of the stimuli.

To test this, we performed three analyses that controlled for acoustic and phonetic features in different ways. The first control (phoneme control) was the same as the above GCA model, except that it included a categorical predictor for phoneme identity (e.g. /s/, /i/, /m/) at each time point. The second control (manner control) included a categorical predictor for the manner of articulation of each phoneme, defined as one of the following categories: fricative, plosive, liquid/glide, nasal, low back vowel, high back vowel, or high front vowel. The third control (acoustic control) was identical to the original analyses described above, but prior to modeling, linear spectrotemporal receptive fields (STRFs) were estimated to describe the acoustic tuning in each electrode based on responses generated while listening to the TIMIT corpus (Garofalo Lamel et al. 1993). The STRF-predicted response was subtracted from the actual high-gamma signal, and the residual was used as the dependent variable in these models. (For more detail on this method, see Theunissen et al. 2001 and Mesgarani and Chang 2012.)

Across all electrodes, the average conditional R^2 values were as follows: base analyses – 0.177 (standard deviation 0.149); phoneme control – mean 0.193 (s.d. 0.150), manner control – mean 0.182 (s.d. 0.150), acoustic control – 0.166 (s.d. 0.143). To evaluate whether these differences were meaningful, the conditional R^2 values from each electrode (log-transformed for normality) were compared with a simple ANOVA, with main effects for subject and control type (base, phoneme, manner, or acoustic). While there was a small effect of model (F (3,1183) = 3.3708, p = 0.018), post-hoc tests (Tukey's honestly significant difference test) revealed that this was driven exclusively by the difference between the phoneme models and the acoustic model (p = 0.010). This indicates that no model taking acoustic or phonetic features into account was appreciably different from the reported base models, suggesting that our main analyses were not driven by lower-level differences in the stimuli.



Supplemental figure 1: Explanatory power of different model types. Mean and quartiles of R^2 values across all electrodes in all subjects for each model type: the basic models reported in the main text ("base"), models controlling for the manner of articulation of each segment ("manner"), models controlling for segment identity ("phoneme"), and models controlling for spectrotemporal tuning, as quantified using STRF models ("acoustic"). No model type was found to be significantly different than the base models.