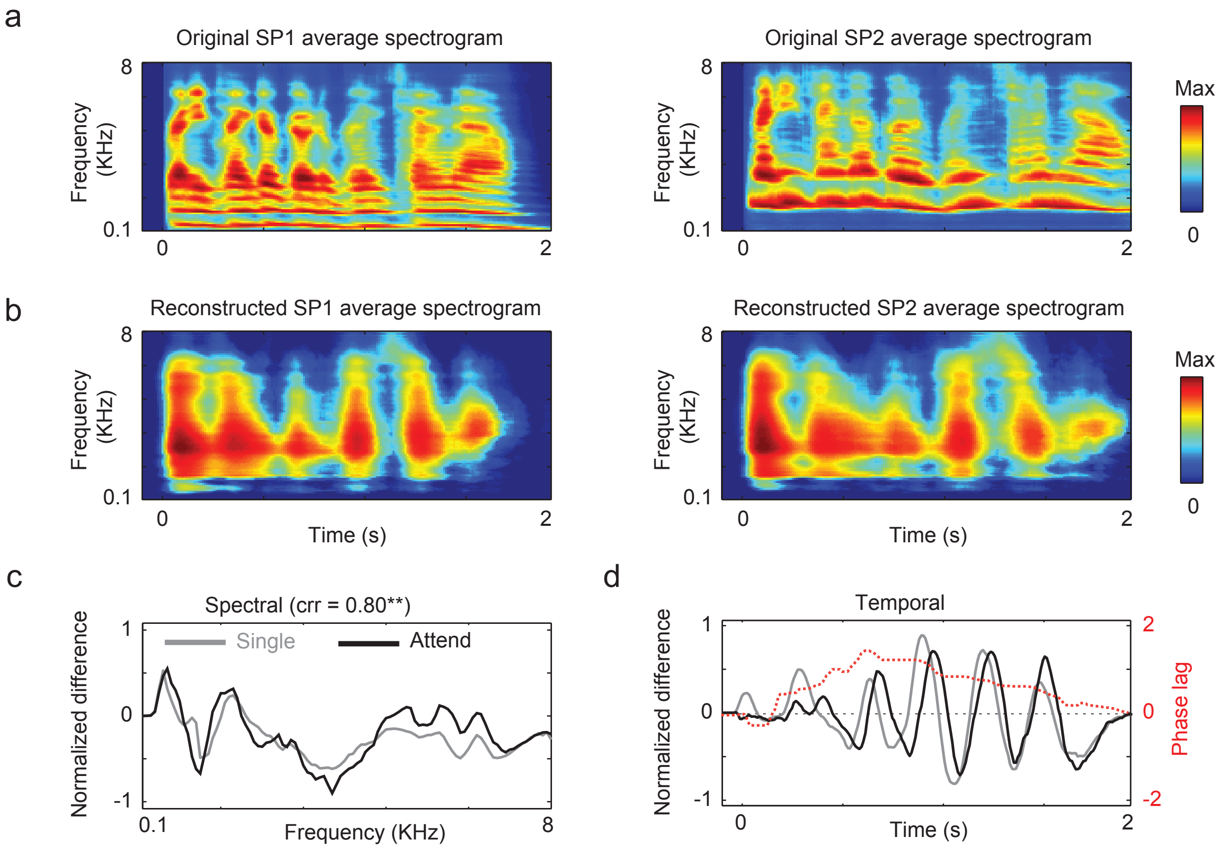
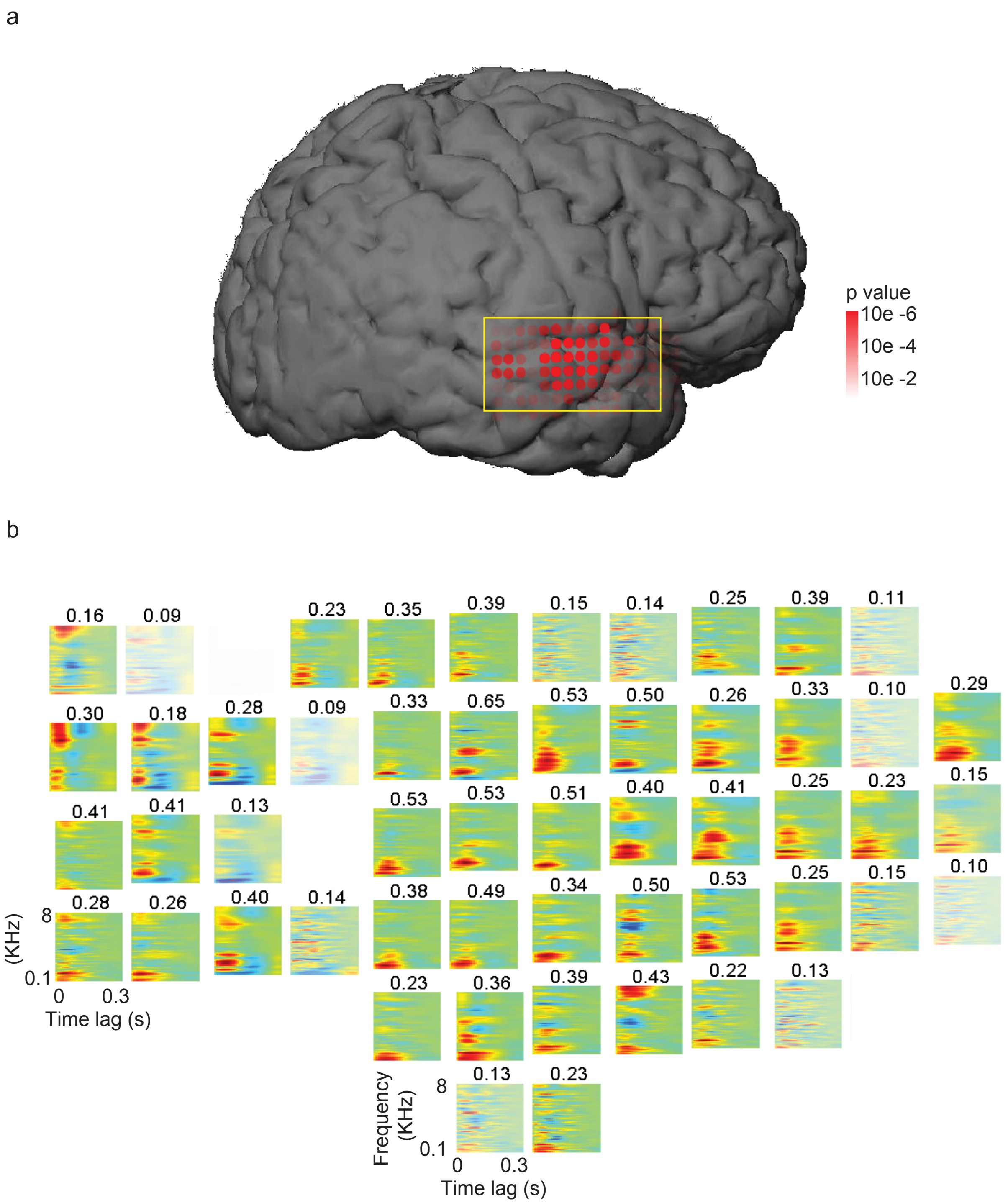
doi:10.1038/nature11020



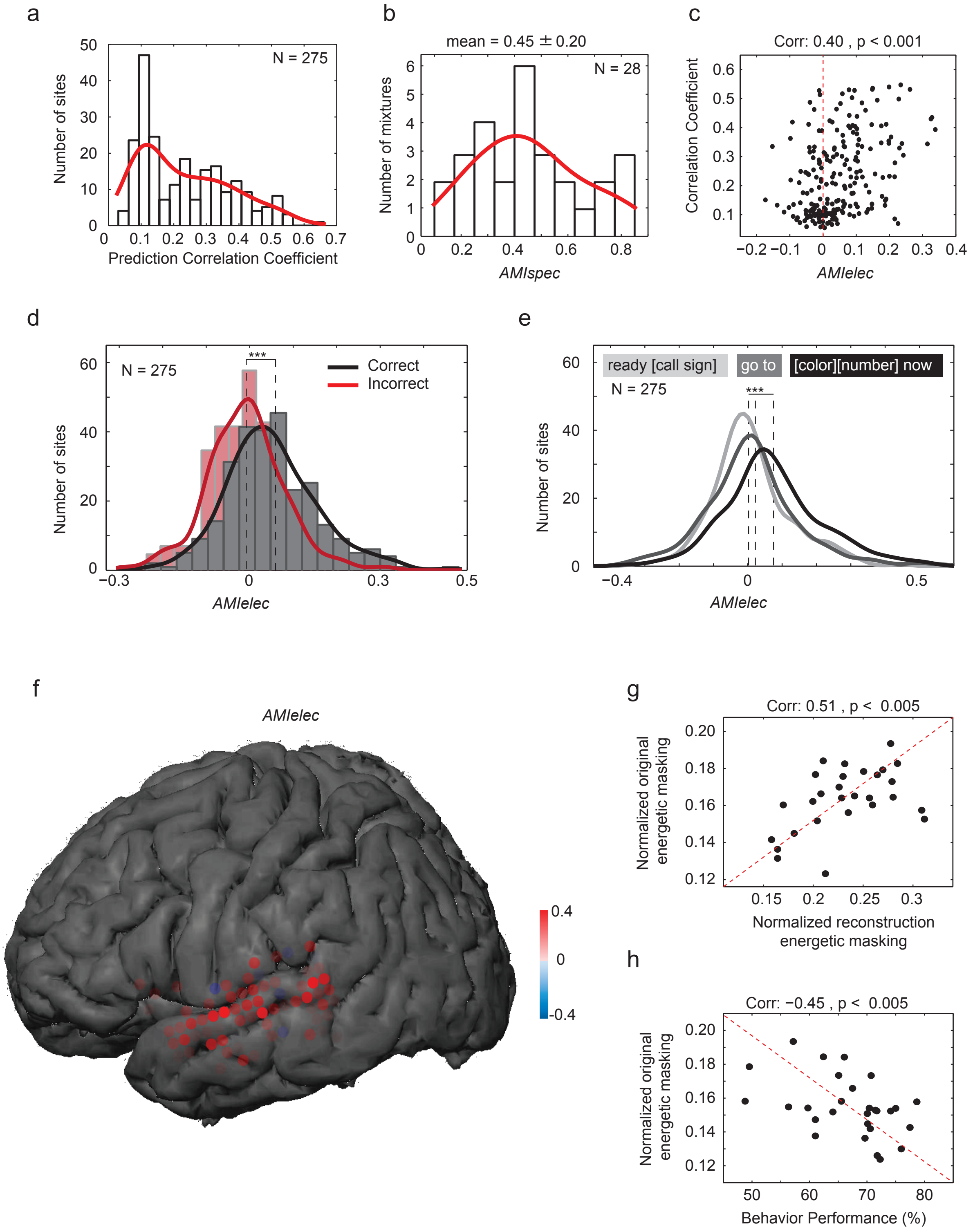
Supplementary Figure 1.

(a) Average original acoustic spectrograms of speakers one and two. (b) Average reconstructed spectrograms of the two speakers show the same spectrotemporal energy distribution as in the acoustic spectrograms. (c) Reconstructed spectral difference of speakers one and two alone (gray) and in attended mixture (black) estimated by averaging the temporal dimension of Figs. 2e,f. The similarity of the difference spectral profiles in single and attended mixture (corr = 0.80, $p < 0.01$, t-test) suggests an accurate restoration of discriminating spectral features of the two speakers induced by attentional modulation. (d) The difference in speaking rate of the two speakers, yet stereotyped structure of the carrier phrases, results in specific average temporal modulation profiles for each speaker (Fig. 2e,f). Reconstructed temporal difference of speakers one and two alone (gray) and in attended mixture (black) shows a gradually enhanced synchrony between temporal envelopes. The average instantaneous Hilbert phase difference between the two plots decreased in time, indicating a gradually reduced time delay (increased synchrony) between the two temporal envelopes. Therefore, selective attention enhanced the encoding of both spectral aspects and temporal characteristics of the attended speaker.



Supplementary Figure 2.

(a) Electrodes with significant difference between responses to silence and speech sounds ($p<0.01$, t-test). (b) Spectrotemporal receptive fields (STRF) for responsive sites measured from passive listening to TIMIT speech corpus. The receptive fields are plotted at the location of their corresponding electrode on the brain and their opacity is proportional to the corresponding correlation value between STRF predicted and actual neural responses, also plotted on top of each STRF.



Supplementary Figure 3.

(a) Histogram of correlation values between STRF predicted and actual neural responses, for all 275 electrode sites. (b) Histogram of *AMIspec* values for all the individual mixture sounds (N = 28) estimated from the correct trials. The *AMIspec* values are all positive ranging from 0.1 to 0.8, showing a varied degree of shift toward the target speaker. (c) Attentional modulation index (*AMIelec*, equation 2 in Methods) of all 275 sites plotted against their STRF prediction correlation coefficient. The significant correlation between the two (Corr = 0.40, p<0.001, t-test) suggests that sites with better STRF predictions are generally more modulated by attention. The STRF prediction values are influenced by factors such as linearity of the response and neural variability (noise). Separating linearity versus noise is intrinsically difficult given the limited repetitions of each stimulus. (d) Histogram *AMIelec* of 275 responsive electrode sites in correct and incorrect trials, fitted with a non-parametric curve (Gaussian kernels). Difference between the mean of two groups is 0.072 (p<10-5, Kruskal-Wallis test), showing varying degree of bias toward the attended speaker responses across the population. (e) Attentional modulation for correct trials at different word positions shows a gradual time -dependent population tuning shift towards to the attended speaker after the call sign ended (difference between means = 0.031, 0.083, p<10-e-5, Kruskal-Wallis test). (f) Distribution of AMI for one example subject appeared to be well distributed over responsive sites. There was no clear topographical organization or localization of modulated sites. (g) Normalized energetic masking is measured by calculating the degree of overlap between two spectrograms in each mixture. Energetic masking in each of the 28 mixtures is significantly correlated with the difference between reconstructed mixture spectrograms in the two attended conditions (Corr = 0.51, p<0.005, t-test) suggesting that an overlap in the acoustic domain results in reduced separability of the neural responses (h) Energetic masking is also significantly correlated with the perceptual performance of subjects for each mixture sound (Corr = -0.45, p<0.005, t-test).