Different oscillations in the superior temporal sulcus integrate faces and voices differently

Chandramouli Chandrasekaran, Asif A. Ghazanfar
Princeton University

During vocal communication (including speech), combining auditory and visual information leads to better detection, discrimination and learning. The neural mechanisms of such audiovisual integration have received considerable attention. For instance, hemodynamic studies [1] have shown that the superior temporal sulcus/gyrus is the site of enhanced responses to audiovisual stimuli relative to unisensory stimuli. In contrast, EEG and MEG studies with human subjects have consistently shown that evoked responses to audiovisual stimuli from putative sources in the superior temporal gyrus are attenuated relative to unimodal stimuli [2,3]. Our study uses a macaque monkey model system [4,5] to investigate the neural bases for these discrepant results.

We hypothesized that different components of the oscillatory hierarchy in the neocortex might integrate audiovisual information differently depending on the temporal relationship between the auditory and visual components of the vocal signal. To test this, we recorded local field potentials (LFPs) and single units from the upper bank of the superior temporal sulcus (STS) of monkeys while presenting them with face+voice, voice alone and face alone stimuli from their vocal repertoire. The monkey subjects were simply required to maintain their fixations on the screen for the duration of a given stimulus. Importantly, our stimuli possessed a natural variation in the timing of the visual component relative to the auditory component: for each call, there is a delay between the onset of the mouth movement and the onset of the voiced component (“time-to-voice”) that could range from 66ms to 331ms. We used wavelet-based spectral analyses to examine LFP responses to bimodal versus unimodal signals and as a function of the time-to-voice.

We found that in low frequency bands (alpha and theta), audiovisual responses are suppressed relative to unimodal auditory responses when the time-to-voice is greater than 200ms, but enhanced when the time-to-voice is less than 100 ms. Surprisingly, audiovisual responses in the gamma band were consistently enhanced regardless of the time-to-voice variable. The finding that the gamma band activity and the low frequency activity possess different temporal properties supports differential roles for these signals in integrating multisensory signals. Furthermore, these results explain the discrepant findings in hemodynamic versus EEG/MEG studies of audiovisual integration. The enhanced BOLD response seen in hemodynamic studies of audiovisual integration [1] is largely driven by local gamma band activity [6]. In support of this, our data show that, regardless of time-to-voice, gamma band activity is always be enhanced. With regard to the EEG/MEG studies of audiovisual integration [2,3], the experimental paradigms incorporated a very large and artificial time-to-voice component (>250ms) to eliminate the onset of the evoked response to the face in influencing their analyses. Our data show that such large delays between the visual and auditory components will almost invariably lead to suppressed responses in the low frequency bands of the LFP signal.