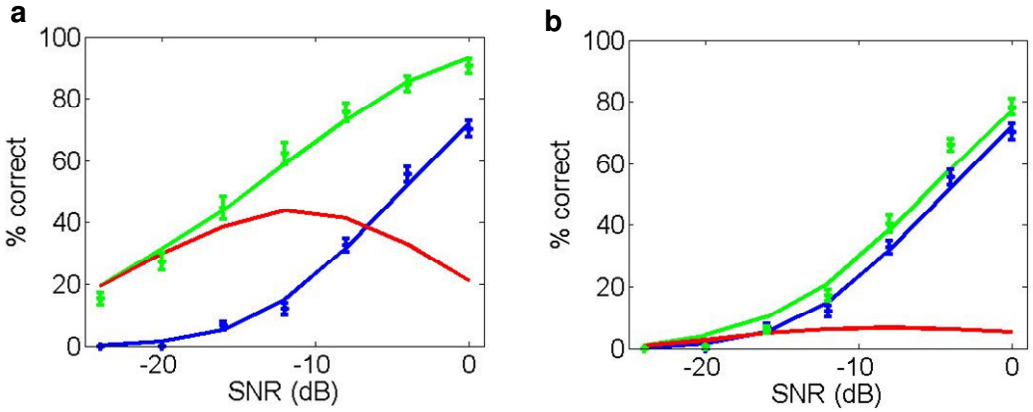
Auditory-Visual Speech Recognition is Consistent with Bayes-Optimal Cue Combination

Wei Ji Ma[1], Xiang Zhou[2], and Lucas Parra[2]

[1]University of Rochester, [2]City College of New York

Watching a speaker's facial movements can dramatically enhance our ability to comprehend words, especially if the speech sound is corrupted by noise. It was recently found in an open-set word recognition task that the improvement in percentage correct due to the use of visual information is maximal not at high but at intermediate auditory noise levels. This seems to contradict the principle of inverse effectiveness often found for multisensory cue combination (the enhancement due to one sensory modality is largest when the other modality has the lowest signal-to-noise ratio). To better characterize this apparent disagreement we reproduce the result under better controlled conditions and show that a novel visual stimulus that provides only temporal information produces a behavior conflicting with, and even opposite to inverse effectiveness.

We then present a Bayesian model of optimal weighting of auditory and visual information that can explain the data from both conditions. In this model, words are regarded as points in a continuous, multidimensional space with values that have to be inferred from the sensory cues. The high dimension is a key property for explaining the maximal performance enhancement at intermediate noise levels. In contrast, inverse effectiveness results from the Bayesian model in low dimensions. These results suggest that auditory-visual speech perception is consistent with the same optimality properties of multisensory cue combination that were previously observed only on very simple stimuli. Several predictions are obtained for multisensory cue combination with complex stimuli.



Auditory-alone performance (blue symbols) and auditory-visual performance (green symbols) in speech recognition as a function of auditory SNR. Behavioral data are fitted well by a realistic Bayes-optimal model of speech perception (lines). Red line: multisensory enhancement. **a:** Full visual information. **b:** Impoverished visual information.

Reference

[1] Ross LA, Saint-Amour D, Leavitt VN, Javitt DC, Foxe JJ (2007) Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17:1147-1153.