Natural Audiovisual Speech Encoding in the Early Stages of the Human Cortical Hierarchy

Funding Source: NIH R01DC016297 (MPI: Matthew Dye. MPI: Ed Lalor - URMC)

Seeing a speaker’s face can greatly improve one’s ability to understand what they are saying, especially under adverse hearing conditions. This phenomenon has been attributed to the multisensory integration of audio and visual speech. Decades of research suggests that such integrative mechanisms come in two forms based on “when” the audio and visual speech are integrated: 1) early integration at the acoustic level, before speech acoustics have been transformed into linguistic representations; and 2) late integration at the linguistic level, where the visual system supplies its own linguistic representations that constrain the inferences being made about audio speech. This has led some to propose that audiovisual speech integration is a multistage process.

However, the multistage models that have been proposed lack detail on how speech acoustics and visible articulations are initially transformed into linguistic representations. And, more generally, they have been vague on how visual speech constrains linguistic categorization. Progress on updating these models has been slow over the past decade, in large part because the field has overly relied on – admittedly interesting and important – paradigms involving discrete (often illusory) speech stimuli. To fully characterize early integrative processes – likely based on the correlated dynamics of visual and auditory speech – and later integrative processes – likely based on how the form of visual articulators helps with linguistic categorization – surely requires the use of ongoing, natural audiovisual speech.

In this project, we propose an updated model of multistage audiovisual speech processing. This model is built on the hypothesis that hierarchical speech processing occurs in both the auditory and visual systems – and that different hierarchical speech representations in the visual system flexibly influence different hierarchical levels of processing in the auditory system, depending on the quality of the acoustic input and the task.

We aim to test several predictions generated by this model. A first prediction is that correlated motion from visual speech enhances auditory selective attention and, thus, auditory cortical sensitivity to acoustic speech, while visual articulatory cues enhance the categorization of auditory speech into linguistic units.

A second prediction is that visual speech aids in speech recognition according to an information theoretic process at the sub-lexical (i.e., phoneme-by-phoneme) level. Finally, we predict that different visual speech representations differentially influence audio speech processing as a function of acoustic experience/quality. We will test this by characterizing such processing in people who vary in their level of acoustic speech experience, namely deaf individuals who do and do not use cochlear implants, as well as typically hearing individuals.

Ultimately, this project will provide a rigorous test of an updated model of natural audiovisual speech processing in the human brain. We also expect it to produce methods, paradigms and measures that will be valuable in future basic and clinical research. And we have assembled a team that is ideally suited to deliver on these goals.