Are the objects of speech perception auditory qualities or articulatory gestures? This course shows that this is not a parochial question, of interest only to phoneticians and phonologists, but belongs to a much larger debate in psychology, cognitive science, neuroscience, and philosophy about the nature of percepts, which can itself be cast in the form of an apparently simple as well as ancient question: are percepts in the world or in the head? In its modern form, this question is about whether and how cognition is embodied and extended rather than computed.

To listeners, sounds appear to occur in the world, where the sound-producing event occurs, e.g. speech sounds appear to come from the speaker's mouth. This phenomenological characteristic has motivated many philosophers to argue that auditory percepts are these external events or the objects that they produce in the world, and not the sensations they produce in the listener. The external alternative is also encouraged by evidence that percepts are grounded in how the perceiver's body interacts with the world, and that they may apprehend information arrays that are wholly outside the perceiver. In this course, I contrast this external or embodied and extended account with the internal alternative (which I advocate), that auditory percepts must be in the head, because they are produced by considerable mental computation, including sophisticated inferences, and that the apparent external phenomenological character of sounds is a projection into the world of the outputs of these computations and inferences.

The choice here is between a theory about the overt, public, or phenomenological aspects of auditory percepts versus one about the covert, private, and internal processes that produce those aspects. Regardless of whether the listeners is perceiving speech or some other kind of sound, I argue that the theory should first account for the covert internal processes and then use their mechanisms and outputs to account for the overt characteristics of percepts. For example, it should acknowledge that listeners separate concurrent vowels perceptually and produce the cocktail party effect by the vowels' spectral prominences rather than their entire spectra, by their perceived pitches rather than such waveform properties as misaligned harmonics or asynchronous pitch periods, and via glimpses of one vowel's formants during brief intervals when they are more prominent than the other's rather than via the gestalt principle of good continuation. Even though I argue that percepts are first in the head and only eventually in the world, the course's ultimate goal is to reconcile and integrate evidence from empirical studies of listeners' behavior with philosophical and psychological arguments based on their phenomenological experience.

How do the competing accounts of the objects of speech perception represent competing sides in this larger debate? According to the motor theory, listeners recognize speech sounds by internally emulating the articulation of the sound they have just heard and matching the internal acoustic simulacrum produced by the emulation to the heard sound's acoustics. By relying on emulation of the speaker's articulations, the motor theory instantiates embodied cognition. According to the direct realist theory, no emulation is necessary because a speech sound's articulation so structures the signal's acoustic properties that they provide all the information needed to identify the responsible articulation. Because this information is in the world and its source can be recognized by inverting the transformation of articulations into acoustics, direct realism instantiates extended cognition. Both the motor and direct realist theories also easily accommodate the integration of the visual information obtained from watching the speaker's face with the auditory information conveyed by the signal's acoustic properties. According to the auditory alternative, emulation is unnecessary because non-human listeners respond like human listeners to speech sounds, inversion is impossible because the relationship between articulations and acoustics is many-to-many rather than one-to-one, and visual-auditory integration must be late rather than immediate. Auditory transformations of speech sounds' acoustic properties make it easier for the listener to distinguish one speech sound from another by integrating psychoacoustically similar properties and also easier to parse the signal into its constituent sounds by exaggerating the perceived difference between successive sounds. The auditory alternative also accounts for similarities between listeners' responses to non-speech analogues and their responses to the original speech sounds.

This brief and selective review shows that the debate remains unresolved because the competing sides appeal to different kinds of evidence. The auditory account relies on covert aspects of listeners' responses to speech signals, the inaccessibility of articulations and internal computations, unconscious analogies, and inferences, while the articulatory alternative relies on overt aspects of those responses, what the perceiver sees as well as what they hear and thus links of percepts to actions, how listening to speech is described, and how it differs from listening to non-speech.

The course first reviews the six decades of research on speech perception and its relationships to speech production since WWII, then situates speech perception within research in philosophy, cognitive science, neuroscience, and artificial intelligence into sound perception, and ends by discussing new experimental studies of decisive predictions of the competing accounts.