Summary: The sounds of speech

When we talk about **speech perception**, we are talking about how listeners use information in the acoustic signal to identify speech sounds.

By speech sounds (or phonemes, as a linguist would say), we mean consonants and vowels. A **phoneme** is the smallest unit of language that can change the meaning of a word. For example, /b/ and /p/ are phonemes because switching from one to the other can change the meaning of a word (e.g., from *bin* to *pin*). But [p] (an unaspirated *p*, as in the word *spin*) and [p^h] (an aspirated *p*, as in the word *pin*) are not phonemes – using one or the other does not change the meaning of a word (even though they are acoustically distinct). In this example, [p] and [p^h] are called **allophones** of a single phoneme; they fall within the same phoneme category.

We can describe phonemes based on how they are articulated. Consonants are contrasted from each other by which part of the vocal apparatus is involved in making the sound (place of articulation), how air moves through the vocal tract (manner of articulation), and when the vocal folds vibrate (voicing).

We are not very good at being able to distinguish between two instances of the same phoneme – that is, two productions of /b/, though acoustically different, are relatively hard to tell apart. By contrast, it is much easier to differentiate a /b/ from a /d/. The fact that it is easier to differentiate between two categories than it is to differentiate within a phoneme category is called **categorical perception**.

Categorical perception is not unique to humans, and it is not unique to speech either. Colors are also perceived categorically – it's easier to tell apart a green from a blue than it is to tell apart two shades of green. The phenomenon of categorical perception is a striking example that our perception of the world is not veridical – the knowledge that we have dramatically affects the way we experience the world.

Indeed, there are many examples of how our prior knowledge influences how we perceive the world. The way we perceive a color, for instance, is relative to the background we see it against. Similarly, the way we perceive speech sounds is relative to their context – the same speech sound can be perceived differently depending on factors like what the surrounding phonemes are, who is talking, and what visual information accompanies the speech signal.

The fact that the same auditory information is not always perceived the same way is known as the **lack of invariance problem in speech perception**. That is, there do not appear to be invariant (unchanging) cues in the speech signal itself that tell us how acoustics map onto phonemes. Put simply, *there is no one-to-one mapping between acoustics and phonetic categories.*

The **motor theory of speech perception** claims that there *is* an invariant cue, but it is not acoustic in nature. The motor theory holds that listeners overcome the lack of invariance problem through an understanding of *how the sounds are produced in the vocal apparatus*. That is, we can tell whether a sound is a /b/ or a /p/ because we know what articulatory gesture a talker needs to make to produce each one.

One recent suggestion is that we are able to overcome the lack of invariance problem by leveraging context. We can use contextual cues (such as the visible movement of the articulators) to guide our decision about what a person has said. Critically, this context does not just guide perception in the moment; we also use contextual knowledge to guide future encounters with a talker, even if, in that second encounter, we no longer receive visual information about the talker. This latter process is referred to as **phonetic recalibration**.