Speech Perception

Casey O'Callaghan, Rice University

Oxford Handbook of the Philosophy of Perception, ed., Mohan Matthen

Author biographic information

Casey O'Callaghan's research aims at an empirically informed philosophical understanding of perception that is driven by thinking about non-visual modalities and the relationships among perceptual modalities. His research has focused upon auditory perception, speech perception, and the theoretical import of multimodality and cross-modal perceptual illusions. O'Callaghan is author of *Sounds: A Philosophical Theory* (Oxford, 2007) and co-editor, with Matthew Nudds, of *Sounds and Perception: New Philosophical Essays* (Oxford, 2009). O'Callaghan received a B.A. in Philosophy and Cognitive Science from Rutgers University and a Ph.D. in Philosophy from Princeton University. He is Associate Professor of Philosophy at Rice University.

Abstract

Is speech special? This chapter evaluates the evidence that speech perception is distinctive when compared with non-linguistic auditory perception. It addresses the phenomenology, contents, objects, and mechanisms involved in the perception of spoken language. According to the account it proposes, the capacity to perceive speech in a manner that enables understanding is an acquired perceptual skill. It involves learning to hear language-specific types of ethologically significant sounds. According to this account, the contents of perceptual experience when listening to familiar speech are of a variety that is distinctive to hearing spoken utterances. However, perceiving speech involves neither novel perceptual objects nor a unique perceptual modality. Much of what makes speech special stems from our interest in it.

Keywords

speech perception, auditory perception, content, objects of perception, McGurk, aphasia, pure word deafness, phoneme, motor theory

Philosophers have devoted tremendous effort to explicating what it takes to *understand* language. The answers focus on things such as possessing concepts, mastering grammar, and grasping meanings and truth conditions. The answers thereby focus on extra-perceptual cognition. Understanding spoken language, however, also involves *perception*—grasping a spoken utterance requires hearing or seeing it. Perception's role in understanding spoken language has received far less philosophical attention. According to a simple view, understanding speech is just a matter of assigning meaning to the sounds you hear or to the gestures you see. If so, what perception contributes to understanding spoken language is not distinctive to the case of spoken utterances. Against this, however, is the prospect that speech is special. In this chapter, I present and evaluate the evidence that speech perception differs from non-linguistic auditory perception. In particular, I discuss the phenomenology, contents, objects, and mechanisms of speech perception. I make proposals about the ways in which speech is and is not perceptually special. According to the account I offer, the capacity to perceive speech in a manner that enables understanding is an acquired perceptual skill. It involves learning to hear language-specific types of ethologically significant sounds. According to this account, while the contents of perceptual experience when listening to familiar speech are of a variety that is distinctive to hearing spoken utterances, perceiving speech involves neither novel perceptual objects nor a unique perceptual modality. Much of what makes speech special stems from our fierce interest in it.

1 Is speech perceptually special?

There is a thriving debate about whether the human capacity to use and understand language is special (see, e.g., Hauser, Chomsky, and Fitch, 2002; Pinker and Jackendoff, 2005). A key part of this wider debate is whether the capacity to speak and understand speech is special (see, e.g., Liberman, 1996; Trout, 2001; Mole, 2009). My concern here is with speech perception. Is the human capacity to perceive spoken language special?

To be special requires a difference. However, the debate about whether speech is special is not just about whether speech perception in some respect differs from other forms of perception. It concerns whether speech perception should be distinguished as a distinctive or a unique perceptual capacity. Put in this way, the question relies on a comparison. The most common contrast is with general audition. The question thus is whether speech perception differs or is a distinct perceptual capacity when compared with *non-linguistic auditory perception*. A separate contrast is with the capacities of *non-human animals*. Is speech perception uniquely

human? The contrast between human and non-human responses to spoken language frequently is used to illuminate the contrast between human speech perception and non-linguistic audition.

A difference is a difference in some respect, and being distinctive or unique is being distinctive or unique in some way, for some reason. In what respects is speech special? It is helpful to divide the candidates into four broad classes.

The first concerns the *phenomenology* of speech perception. Does what it is like to perceptually experience spoken utterances contrast with what it is like to perceptually experience non-linguistic sounds and events? One way to make progress on this question is to ask whether the perceptual experience of hearing speech in a language you know differs phenomenologically from that of hearing speech in an unfamiliar language.

The second concerns the *contents* of speech perception. Does the perceptual experience of speech involve contents absent from non-linguistic auditory experience? Does understanding a language affect which properties perceptual experiences represent spoken utterances to have?

The third concerns the *objects* of speech perception. Are the objects of speech perception distinct from the objects of non-linguistic audition? Does speech perception share objects with non-linguistic audition?

The fourth concerns the *mechanisms* of speech perception. Does perceiving speech involve perceptual *processes* that differ from those involved in perceiving non-linguistic sounds and events? Does speech perception involve a special perceptual *module*? Is speech perception the work of a distinct perceptual *modality*?

Answering the question, "Is speech special?" thus means addressing a number of different questions. This chapter focuses on the contrast between speech perception and human non-linguistic auditory perception. I distinguish the various respects in which speech might be special when compared with non-linguistic audition. I assess the evidence and advance proposals about the respects in which speech perception is special.

2 Phenomenology

Is perceiving speech phenomenologically special? Is what it's like, for the subject, to perceptually experience speech different, distinctive, or unique when compared with non-linguistic audition?

It is natural to think that the perceptual experience of listening to spoken language differs phenomenologically from the perceptual experience of listening to non-linguistic sounds, simply because speech sounds and non-linguistic sounds differ acoustically. Hearing the sound of a drop

of water differs phenomenologically from hearing the sound of the spoken word 'drop' because the sounds differ in their basic audible qualities.

However, the perceptual experience of spoken language also may involve distinctive phenomenological features that are absent from non-linguistic auditory experience. Start with the experiential contrast between listening to non-linguistic sounds and listening to spoken language. Begin with the case of a language you know. The experience of listening to speech in a language you know differs noticeably from the experience of listening to ordinary, non-linguistic environmental sounds, even once we eliminate acoustical differences. The phenomenological shifts associated with sinewave speech support this claim. Sinewave speech is an artificial signal in which an acoustically complex human voice is replaced by several sine waves that vary in frequency and amplitude with the primary formants of the original speech signal, while removing acoustical energy at other frequencies (Remez et al., 1981). At first, it is difficult to recognize the sounds of sinewave speech as speech sounds. Instead, they just sound like computer-generated noises. However, after hearing the original human speech from which the sinewave speech is derived, it is easy to hear what the sinewave speech says. The same stimulus first is experienced as non-speech sounds, and then it is experienced as speech. And this change is accompanied by a dramatic phenomenological shift.

In the case just described, you come to *comprehend* the speech. Thus, understanding might suffice to explain the phenomenological difference when you are listening to speech in a language you know. You grasp meanings, so the experiential difference could in principle be explained in terms of cognitive, rather than perceptual, phenomenology. (This explanation is unavailable if you reject that extra-perceptual cognition has proprietary phenomenology.)

To control for any contribution from understanding, consider the experiential contrast between listening to non-speech sounds and listening to speech in a language you do not know. Is there any phenomenological difference? It is possible reliably to discriminate speech in a language you do not understand from ordinary environmental sounds. Neonates prefer speech sounds to non-speech sounds though they do not understand language. In addition, sinewave speech in a language you do not know may appear first as non-speech sounds and then as speech. Thus, we have evidence that perceptually experiencing a stimulus as speech rather than as non-speech sounds makes a phenomenological difference that does not depend on understanding.

Understanding spoken utterances need not, however, contribute exclusively to the phenomenology of extra-perceptual cognition. Knowing a language also may impact the phenomenal character of perceptual experience. Consider the phenomenological contrast

between the perceptual experience of listening to speech in a language you know and of listening to speech in an unfamiliar language. Of course, languages differ acoustically in ways that affect how they sound. For instance, whether or not you know Hindi, it sounds different from German. To control for acoustical differences that affect phenomenology, fix the language. Contrast the experience of a person who knows the language with that of a person who does not know the language when faced with the same spoken utterance. Or, consider a person's experience prior to and after learning the language. Many philosophers agree that knowing the language affects the phenomenological character of perceptual experience, even while they disagree about the diagnosis (see O'Callaghan, 2011a, pp. 4-5).

What is the source of the difference? Speech in a language you know differs perceptually in several respects. Most obviously, your perceptual experience of its temporal characteristics differs. When you know the language, audible speech does not seem like an unbroken stream of sounds. It seems instead to include discernible gaps, pauses, and other boundaries between words, clauses, and sentences, and you are able perceptually to resolve qualitative features and contrasts at a much finer temporal grain. Familiar speech also appears in other respects to differ qualitatively from unfamiliar speech. For instance, when you have mastered a spoken language, you are able to detect subtle qualitative features and their contrasts, such as the difference between 's' and 'z', or the dropped 'g' or 't' of certain accents. The stimulus sounds different and more detailed when you recognize it as speech and you know the language.

The argument of the last paragraph, unlike the argument from sinewave speech, requires comparing phenomenology across subjects or across long stretches of time. Thus, it is more controversial. An alternative way to establish the point is to compare the shift that occurs with sinewave speech in a language you know with the shift that occurs with sinewave speech in a language you do not know. In each case, recognizing the sounds as speech leads to a shift in phenomenal character. The change, however, is far more dramatic when you know the language. The difference between the two phenomenological contrasts is the difference that accrues thanks to knowing the language.

These arguments indicate that one's perceptual experiences may differ phenomenologically when listening to speech in a known language, when listening to speech in an unfamiliar language, and when listening to non-speech sounds. Moreover, such phenomenological differences can be evoked even when we have controlled for acoustical differences. This supports the following two claims: knowing a language impacts the phenomenal character of perceptual experience when listening to spoken utterances; and, speech

perception has phenomenal features that are distinctive when compared with non-linguistic audition.

3 Contents

Content concerns how things are represented to be. Content thus concerns things that are perceptually experienced and the features they are perceptually experienced to have. One way to characterize the contents of perceptual experiences appeals to their accuracy or veridicality conditions. Some prefer to speak of what a given perceptual experience purports to be facts about the world, or of how things seem or appear perceptually. Some philosophers hold that perceptual experiences differ phenomenologically only if they differ in how they represent things as being. Some also hold that there is a variety of content such that perceptual experiences differ in content only if they differ phenomenologically. In either case, a difference in content may help to explain the sorts of phenomenological differences mentioned in section 2. What we perceive when we perceive speech may, in this sense, differ from what we perceive when we perceive non-speech sounds. Speech perception may involve contents that are special or distinctive when compared with non-linguistic audition.

In what respects does the content of speech perception differ from that of non-linguistic audition? The characteristic sounds of human vocalization differ acoustically from the sounds of non-linguistic happenings such as blowing leaves, backfiring automobiles, and violins. The perceptual experience of speech reflects this. Such evident qualitative differences, which are underpinned by acoustical differences, are part of why sinewave speech at first sounds like meaningless computer noise, and why artificial speech often sounds inhuman. Perhaps, then, the perceptual experience of speech differs phenomenologically from the perceptual experience of non-linguistic sounds and happenings because its perceptually apparent features differ in a way that is recognizable and distinctive to spoken language.

This is compatible with an *austere* view of the types of features that one perceptually experiences when listening to speech or to non-speech sounds. The phenomenological difference between perceptually experiencing speech and non-speech may just stem from a difference in the patterns of *low-level* properties that the perceptual experiences represent. For instance, it may just stem from a difference in the apparent pattern of pitch, timbre, and loudness of a sound stream over time. Any further experiential differences may result from extra-perceptual cognition, such as thought or imagination.

This austere picture also suggests an attractive account of how perceptually experiencing

speech in an unfamiliar language differs phenomenologically from perceptually experiencing speech in a language you know. As discussed in section 2, the audibly apparent *temporal* and *qualitative* features of spoken utterances in a language you know generally differ from those of speech in a language that is unfamiliar to you. Foreign language may sound like a continuous stream of indistinct babble, but familiar speech perceptually appears to be chunked into units that correspond to words and phrases and to include discernible gaps, pauses, and boundaries that distinguish such units from each other. Hearing familiar language also involves the capacity to perceptually experience sublexical features at a finer temporal grain, and to discern linguistically significant qualitative details and contrasts that you could not make out before. Conversely, it also involves failing to discern other qualitative contrasts that are linguistically irrelevant. Thus, in these ways, differences in the perceptually apparent pattern of individual sounds and low-level audible qualities such as pitch, timbre, and loudness over time may explain the phenomenological difference that knowing a language makes.

Nevertheless, such an austere account might not suffice. Some philosophers have claimed that grasping meanings or semantic properties contributes in a constitutive rather than merely causal manner to the phenomenal character of perceptual experience. They argue therefore that listening to spoken utterances when you know the language involves perceptually experiencing *meanings* or *semantic properties* (e.g., McDowell, 1998; Siegel, 2006; Bayne, 2009). According to such an account, perceptual experiences may represent or involve awareness not just as of *low-level* sensible features, such as pitch, timbre, loudness, and timing, but also as of *high-level* features, including semantic properties. Such an account supports a *liberal* view about what types of properties may be represented by episodes of perceptual experience (see, e.g., Siegel, 2006; Bayne, 2009).

The liberal view of speech perception's contents faces an objection if it also must explain the phenomenological difference between the perceptual experience of listening to speech in a familiar language and of listening to speech in an unfamiliar language. The account requires that, for an utterance you understand, there is something distinctive it is like for you to perceptually experience its specific meaning. That is because nothing suggests you could not hear foreign utterances as *meaningful* if that does not require hearing specific meanings. Hearing meaningfulness, if not specific meanings, for instance, could help to explain the phenomenological difference between hearing speech in an unfamiliar language and hearing non-linguistic sounds. However, perceptually experiencing specific meanings also can account for the difference between hearing familiar and unfamiliar speech. Suppose, therefore, that you

perceptually experience specific meanings, rather than just meaningfulness. Thus, differences in apparent meaning should impact the phenomenal character of perceptual experience for utterances in a known language. But consider homophonic utterances, which share pronunciation but not meaning. Homophonic utterances do not apparently cause perceptual experiences that differ in phenomenal character. For instance, even when they are embedded appropriately in meaningful linguistic contexts, the perceptual experience of hearing an utterance of 'to' does not clearly differ in phenomenal character from the perceptual experience of hearing an utterance of 'too' or 'two' (the same holds for homographic homophones). Complete sentences present a similar problem. Utterances of structurally ambiguous statements, such as, 'Visiting relatives can be boring,' and those with scope ambiguities, such as, 'Everyone chose someone,' may not, under their differing interpretations, lead to perceptual experiences that differ phenomenologically. The argument from homophones thus casts doubt on the claim that specific meanings make a distinctive difference to the phenomenal character of perceptual experience (O'Callaghan, 2011a).

A moderate account denies that the perceptual experience of speech includes awareness as of meanings or high-level semantic properties. It nevertheless explains the phenomenological difference that accrues thanks to knowing a language using resources beyond the austere account's low-level acoustical features. According to one such account, listening to speech in a familiar language involves the perceptual experience of *language-specific* but non-semantic properties of spoken utterances.

Phonological features, such as *phones* and *phonemes*, form the basis for recognizing and distinguishing spoken words. Phonological features in general are respects of discernible non-semantic similarity and difference among utterances that may make a semantic difference. Phonological features are like the basic perceptible vocabulary or "building blocks" of spoken language.[1] To illustrate, consider utterances of 'bad', 'imbue', and 'glob'. In one respect, these utterances are perceptibly similar. Each apparently shares with the others the 'b' sound—[b] in phonological notation. Next consider utterances of 'lab' and 'lash'. They perceptibly match, except that the former contains the 'b' sound and the latter contains the 'sh' sound—[ʃ] in

[1] Here I am alluding to but not endorsing the notorious "beads on a string" analogy. I do not accept that characterization of phonological attributes, because I believe neither that they are items or individuals nor that they occur in neat, discrete sequences. Instead, I believe they are properties whose instances overlap. Further discussion in section 4.

phonological notion. The *phones* [b] and [ʃ] are examples of features which may be shared among distinct spoken utterances, which may differ among otherwise indistinguishable utterances, and which may make a semantic difference. Distinct phones are distinguished by a perceptible difference that is linguistically significant in some human language. A phone thus is usefully understood in terms of a type whose members make a common linguistic contribution to any given language. One phone is distinguished from another by some perceptually discernible difference that is or may be exploited by some spoken language to signal a semantically significant difference. Since phones are the minimal perceptible features that make a linguistic difference in some world language, they are in this sense the perceptible "building blocks" of spoken language.

Specific spoken languages do not all make use of this basic stock of building blocks in the same manner. Some spoken languages, for instance, include clicks and buzzes, while others do not. Moreover, spoken languages may, even when they make use of the same basic stock, differ in which classes of utterances they treat as linguistically equivalent and in which classes of utterances they treat as distinct. For example, spoken English distinguishes [l] from [r][2], but Japanese does not. Thus, the phones [l] and [r] correspond to distinct English phonemes, /l/ and /r/, but are *allophones* or linguistically equivalent variations of a single Japanese phoneme. Another example is that [p] and [p^h] are allophones of the English phoneme, /p/, but Mandarin Chinese treats them as distinct phonemes, /p/ and /p^h/. The difference between [p] and [p^h] suffices to distinguish Chinese but not English words. So, some languages treat [p] and [p^h] as allophones of a single phoneme, while others treat them as distinct phonemes that may suffice for a semantic difference.

Phonemes thus may usefully be understood in terms of language-specific classes whose members are treated as linguistically equivalent, or as *allophonic*, within the context of that spoken language, even if under certain conditions its members may be perceptually distinguishable. A language's phonemes are distinguished from one another by perceptually discernible differences that are semantically significant. The lesson is that certain utterance pairs are treated as linguistically equivalent by some languages but as linguistically distinct by others. Thus, spoken languages yield differing families of equivalence classes of utterances that make a common semantic contribution. So, the way in which a basic stock of speech sounds, which have

[2] For readability, I use the upright rather than inverted 'r' for the alveolar approximant. The upright 'r' standardly (in the International Phonetic Alphabet) is used for the trill.

the potential to signal semantic difference, in fact is utilized by a particular language is specific to that language. A language's stock of linguistically significant sound types is distinctive to that language.

Since phonemes differ across languages, discerning a language's phonemes requires substantial exposure and learning. That such features may be perceptually experienced nonetheless helps to explain patterns of similarity and difference among utterances that are apparent to users of a given language. The capacity perceptually to discern such similarities and differences is critical to understanding spoken language. It is not, however, explained by the perceptual experience of low-level audible attributes alone.

What is noteworthy is that users of a given language commonly treat certain crucial pairs of sounds or utterances as perceptibly equivalent, while those who do not know that language treat them as perceptibly distinct. For example, auditory perceptual discrimination tasks in linguistic contexts reveal that the sounds corresponding to 't' in utterances of 'ton' and 'stun' auditorily appear virtually the same to fluent monolingual English users, but appear noticeably to differ to fluent monolingual users of Chinese. Spoken utterances of 'bed' and 'bad' in linguistic contexts differ audibly to English speakers but not to Dutch speakers. Speakers of one language may discern a common linguistic sound across utterances that differ acoustically while speakers of another language do not. So, suppose we have two groups of language users. Suppose all are attentively listening, and that each is presented with two sounds uttered by the same talker in a linguistic context. Those in the first group do not notice a difference between the speech sounds. They judge that they are audibly equivalent, and they behave as if the sounds are equivalent. Those in the other group do notice a difference between the speech sounds. They judge that they audibly differ, and they behave as if the sounds are not audibly equivalent. In this case, for at least one of the speech sounds, it is plausible to say that the perceptual experience of a language listener from the first group differs phenomenologically from the perceptual experience of a listener from the second group. If so, then for a large class of linguistic sounds, the perceptual experience of someone who knows a given language may differ from the perceptual experience of someone who does not. If only those who know a spoken language perceptually experience its language-specific phonological attributes, such as its phonemes, then this provides an attractive explanation for the difference. For instance, having a perceptual experience that represents the English phoneme /l/, rather than /r/, may explain why hearing an utterance of 'law' differs phenomenally from hearing an utterance of 'raw'. Having perceptual experiences as of a single English phoneme explains a monolingual English speaker's failure to perceptually distinguish

utterances of distinct Chinese words. A central part of the phenomenological difference that accrues thanks to knowing a language thus stems from the perceptual experience of attributes whose linguistic significance is specific to that language.

The perceptual experience of language-specific features explains apparent patterns of similarity and difference that to a noteworthy degree are independent from lower-level audible attributes, such as pitch, timbre, and loudness over time. For instance, the low-level audible qualities of an utterance of /p/ vary across phonological contexts, speakers, moods, and social contexts. The perceptual experience of a single phoneme explains this kind of perceptually apparent sameness in the face of differing lower-level audible qualities. On the other hand, the same acoustical signal may appear as a /p/ in some contexts and as a /b/ or /k/ in another. In different contexts, distinct apparent phonemes may accompany matching low-level audible qualities.

A moderate account of this sort finds converging support from three sources of evidence. First, *developmental* evidence shows that young infants discern a wide variety of phonetic differences that are linguistically significant in various languages. However, between five and twelve months, infants cease to discern phonetic differences that are not linguistically significant in the languages to which they have been exposed. Babies in Pittsburgh stop distinguishing utterances that differ with respect to [p] and [p^h], and babies in Madrid stop distinguishing utterances that differ with respect to [s] and [z]. Such pruning requires regular exposure to the language, and it is part of learning to become perceptually responsive to the features that are distinctive to a spoken language. Children thus learn to hear the sounds of their language (see, e.g., Eimas et al., 1971; Jusczyk, 1997).

Second, adult perception of certain critical speech sounds, such as stop consonants, is *categorical* (see Chapter XX, this volume; Harnad, 1987). This means that, in critical cases, such as the perception of stop consonants, gradually varying the value of a diagnostic physical parameter leads to uneven perceptual variation. For example, suppose we start with a stimulus experienced as /ba/ and gradually increase its voice onset time. At first, this makes little difference. At some point, however, the stimulus abruptly appears to shift to a /pa/. In a dramatic case of categorical perception, the change seems perfectly abrupt. Thus, given a boundary that is diagnostic for a perceptual category, stimuli that differ by a certain physical magnitude may differ only slightly in perceptual appearance when each falls within that boundary; however, stimuli that differ by that same physical magnitude may differ greatly in perceptual appearance when one but not the other falls within the boundary.

Patterns of categorical perception in fact vary accordingly. Adult categorical perception of speech sounds corresponds to language-specific phonological categories, generally those of the listener's first language (though there is some flexibility). Perceptual awareness of phonological features thus helps to explain both perceptually apparent patterns of similarity and difference among utterances within a language and variation in patterns of apparent similarity and difference across speakers of different languages.

Third, evidence from *aphasias*, language-related disorders, suggests that the capacity to understand spoken language normally requires the capacity to perceive language-specific attributes of speech that are not meanings. Moreover, the latter capacity affects the phenomenal character of auditory perceptual experience. Individuals with *transcortical sensory aphasia* (TSA) have a severely impaired capacity to grasp and to understand linguistic meanings, but they retain the capacities to hear, to generate, and to repeat spoken utterances. They commonly are unaware of their disorder. In contrast, individuals with *pure word deafness* (PWD) have intact semantic capacities but lack the capacity to perceive spoken language as such. Individuals with PWD are unable to hear sounds or utterances as spoken words or linguistic units. Their deficit is limited to auditory language perception. They may learn to use sign language or even read lips. And their hearing otherwise remains normal. They can hear and recognize barking dogs, cars, and even the sounds of familiar voices. Individuals with PWD say, however, that words fail to "come up" and describe the auditory experience of spoken language as like hearing garbled sound or foreign language (see, especially, Poeppel 2001, p. 681). These descriptions of TSA and PWD suggest that there is an important phenomenological difference in perceptual experience that stems from being able to discern and to recognize language-specific features but that does not require the capacity to discern and to recognize the meanings of spoken utterances. Auditorily experiencing language-specific features other than meanings therefore plausibly captures this difference. Phonological and other structural features of spoken utterances are good candidates.[3]

Appealing to the content of perceptual experience thus helps to explain what is distinctive about the perceptual experience of listening to speech. In particular, two sorts of features help to account for the difference between the perceptual experience of listening to unfamiliar speech and of listening to speech in a language you know. When you know a language, the patterns of

[3] Indeed, individuals with PWD perform poorly on tasks that require categorical perception for language-specific attributes. Thanks to Bob Slevc for discussion.

determinate low-level audible attributes you perceptually experience differ from when you do not know the language. This difference concerns the specific arrangement of low-level qualitative and temporal attributes, each of which you could, in principle, perceptually experience even non-linguistic sounds to bear. However, understanding speech also involves perceptually experiencing spoken utterances to bear language-specific attributes, including phonological properties such as phonemes. Developing the capacity to perceptually experience such language-specific features requires exposure and perceptual learning. Its exercise is part of any adequate explanation for the experiential difference that accrues thanks to knowing a language. While I have expressed doubt that meanings and high-level semantic properties are represented by perceptual experiences, I leave open whether and which additional language-specific features are among the contents of perceptual experience when listening to speech. For instance, you may perceptually experience morphemes, lexemes, or even grammatical properties when you listen to speech in a language you understand. Greater attention to the ways such features affect the phenomenal character of perceptual experience will inform broader debates about the richness of perceptual experience—that is, about the types of features awareness of which constitutively shapes the phenomenal character of perceptual experience. This, in turn, should impact how we understand the interface of perception with cognition.

4 Objects

The previous section argued that the perceptual experience of speech differs in content from non-linguistic audition. This section concerns whether the *objects* of speech perception differ from those of non-linguistic audition. There are two ways to understand the objects of perception. Construed broadly, the objects of perception simply are targets of perception, and may include particular individuals, their attributes, happenings, or states of affairs. In this broad sense, to be an object of perception is just to be perceived. According to some accounts, objects of perception in the broad sense are the components of content. In section 3, I proposed that the perceptual experience of speech involves awareness as of language-specific features. So, in the broad sense, the objects of speech perception are special when compared with those of non-linguistic audition.

Construed more narrowly, however, the objects of perception are the *individuals* that bear perceptible attributes. In this narrow sense, vision's objects might include ordinary material objects that look to have attributes such as shape and color, and audition's objects plausibly include individual sounds that have pitch, timbre, and loudness. Further philosophical debates concern the natures of the objects of perception, including whether they are public or private.

The phenomenological differences between speech perception and non-linguistic audition, especially since they are dramatic, might be taken to suggest that the objects of speech perception in this sense differ from those of non-linguistic audition. This discussion concerns whether the objects of speech perception are special in the narrow sense that includes only individuals.

In one respect, it is trivial that the objects of speech perception differ from those of non-linguistic audition. One case involves perceiving speech, and the other involves perceiving non-speech. At the very least, perceiving speech involves perceiving sounds of a kind to which non-linguistic sounds do not belong, and vice versa. Speech sounds and non-linguistic sounds differ in their causes, their sources, and their effects, as well as in their semantic and other linguistic properties.

The claim that speech perception and general audition have different objects typically is not just the claim that they involve hearing different kinds of sounds or sounds with distinctive features. Speech perception researchers have claimed that the objects of speech perception are not sounds at all. This is a claim about the sorts of individuals perceived when one perceives speech. In particular, it is the claim that while the individuals you perceive in non-linguistic auditory perception are sounds, the individuals that you perceive when you listen to speech are not sounds. The objects of speech perception instead are individuals of a wholly different sort.

Three main sorts of argument are offered. The first type of argument appeals to the *mismatch* between salient features of the objects of speech perception and features of the acoustic signal. We can reconstruct the argument in the following way. The objects of non-linguistic audition are sounds. The perceptible features of sounds correspond to aspects of the acoustic signal. But, the perceptible features of speech do not correspond to aspects of the acoustical signal. The perceptible features of speech thus are not perceptible features of sounds. So, the objects of speech perception differ from those of non-linguistic audition.

This argument can be illustrated using the case of apparent phonological features, such as *phones* or *phonemes*. The *acoustic* attributes that correspond to a perceived phonological feature vary greatly depending upon setting and context. Not only do they vary in expected ways, with speaker, mood, and accent, but they also depend locally upon the surrounding linguistic context. For example, phonological features are not uttered in discrete, isolated units. Instead, they are articulated in a continuous stream that flows gradually from one to the next. This has two noteworthy consequences. First, information about one phoneme is blended with information about surrounding phonemes. Because distinct speech sounds are *coarticulated*, when I utter

'imbue', the fact that the /i/ is followed by /m/ shapes how I pronounce the /i/. This differs from how I pronounce the /i/ when it is followed by /d/, as in 'idiom'. In fact, no clear *invariant* acoustic signature corresponds to an utterance of a given phoneme in all of its perceptible instances. And a given acoustical configuration might contribute to distinct apparent phonemes in different contexts. Second, some have been inclined to say that perceptible speech appears to be *segmented* into discrete phonemes. However, the acoustic information by which you to discern the presence of a given phoneme is present during the utterance of surrounding phonemes. For instance, the acoustical information corresponding to /æ/ in an utterance of 'dab' is present during the articulation of both the /d/ and the /b/ (and vice versa). Thus, no clear acoustic boundaries correspond to any segmentation that is apparent between adjacent phonemes. Therefore, there exists no consistent, context-independent, homomorphic mapping between apparent phonemes and straightforward features of the acoustic signal (see, e.g., Appelbaum, 1999; Remez and Trout, 2009).[4] This point should be evident to anyone who has labored with speech recognition software. It leads some philosophers to anti-realism about phonological features. Rey (2007), for instance, holds that phonemes are *intentional inexistents* (see also Smith, 2009).

In light of this, Liberman et al. (1967) and other early proponents of the Motor Theory famously proposed that the objects of speech perception are not sounds at all, but instead are something involved in the pronunciation of speech (see also the papers collected in Liberman, 1996). The core idea is that features of perceived speech do map in a homomorphic, invariant way onto types of *gestures* involved in the production of speech. For instance, pronouncing an instance of /d/ involves stopping airflow by placing the tongue at the front of the palate behind the teeth and then releasing it while activating the vocal folds. Pronouncing /b/ involves a voiced release of air from pursed lips. Such *articulatory gestures*, and the component configurations and movements they comprise, make the manner in which speech is perceptually experienced intelligible in a way that attention to the acoustic signal does not, since such gestures and their descriptions are less sensitive to context.[5] The claim was that the acoustical signal encodes

[4] Early text-to-speech methods failed to appreciate this context dependence, and thus failed. Early attempts assigned each letter a sound and played the sounds assigned to specific letters in sequences that mirrored written texts. The results were unintelligible.

[5] One complication is that due to coarticulation the gestures pronounced in normal speaking also exhibit some lack of invariance. Liberman and Mattingly (1985) revised the Motor Theory to

information about articulatory gestures and their features. If articulatory gestures and their features rather than sounds and their attributes are the best candidates for what we perceive when we are perceptually aware of instances of phonemes, then articulatory gestures are the objects of speech perception. Thus, the objects of speech perception and of non-linguistic audition differ in kind. The former are articulatory gestures with phonological characteristics, and the latter are sounds with audible attributes.

These arguments do not establish that the bearers of phonological features are not bearers of non-linguistic audible attributes. Thus, they do not establish that the objects of speech perception include individuals of a wholly different kind from the objects of non-linguistic audition. On one hand, the mismatch argument relies on the presumption that ordinary auditory awareness *does* map in an invariant, homomorphic way onto features of the acoustic stimulus. However, even pitch, an apparently simple audible quality, has a complex relationship to frequency. In addition, context effects abound. For instance, varying the attack of a sound affects its timbre, and the apparent duration of a tone is affected by the duration of a tone presented earlier or even later. More generally, the apparent objects of auditory awareness in acoustically complex environments do not map clearly and in invariant ways onto straightforward features of the acoustic signal. Nothing obvious in an acoustical stream signals how to distinguish the sound of a guitar from the sound of a voice in a crowded bar. The central lesson of work on *auditory scene analysis* is that ordinary sounds are individuated—they are distinguished from each other at a time, and they are tracked and segmented over time—in the face of highly complex, interwoven acoustic information (Bregman, 1990).

On the other hand, the argument also relies on the presumption that non-linguistic audition's objects *do not* map in an illuminating way onto the events that produce acoustic information. However, audition's vital function is to provide perceptual access to events in the environment. Accordingly, human audition carves up the acoustical scene in a way that is predicated upon an interest in identifying sound sources. In fact, the way in which sounds are individuated suggests that the objects of non-linguistic auditory perception include sound sources rather than mere acoustical events or sound streams. In the face of complex, entangled acoustical

---

claim that *intended motor commands* are the objects of speech perception. See Mole (2009) for a convincing critique of the revised account. Fowler's (1986) Direct Realism maintains that articulatory gestures are the objects of speech perception but rejects that gestural events differ in kind from the objects of non-linguistic audition.

information, you distinguish the sound of the guitar from the sound of the voice because they have distinct sources. We attend to and identify sounds relative to sources, and this is reflected in our thought and talk about sounds, which concern, for instance, the sound *of the car door*, the sound *of the dog*, the sound *of scratching*. Many descriptive sound words are source oriented: *rattle*, *bang*, *crack*. So, just as articulatory gestures illuminate the manner in which the objects of speech perception are individuated and classified (see Matthen, 2005), considering the environmental happenings that make sounds illuminates the manner in which the objects of non-linguistic auditory perception are individuated and classified (see, e.g., Nudds, 2010). Audition's objects thus fail to map in an invariant, homomorphic manner onto simple physical properties of an acoustic stimulus, and sound sources help to explain the manner in which audition's objects are individuated and classified. In these respects, non-linguistic audition does not differ from speech perception. The *mismatch* argument fails.

The second type of argument is that *cross-modal influences* in the perception of speech reveal that the objects of speech perception differ in kind from the objects of non-linguistic audition (see, e.g., Trout, 2001, for discussion). The McGurk effect is one powerful example (McGurk and Macdonald, 1976). Subjects presented with audio of an utterance of the velar /ga/ along with video of a speaker uttering the bilabial /ba/ regularly report perceptually experiencing the alveolar /ga/. Seeing the speaker impacts which phoneme perceptually appears to be uttered. In fact, visual information systematically affects which phoneme you perceptually experience, so both vision and audition provide information about the objects of speech perception. Moreover, Gick and Derrick (2009) demonstrate tactile influences on speech perception. The objects of speech perception are multi-modally accessible. Sounds, however, are neither visible nor multi-modally accessible. Therefore, since sounds are the objects of ordinary non-linguistic audition, the argument concludes that that the objects of speech perception and non-linguistic audition must differ.

One objection stems from the reply to the first argument. If audition's objects include sound sources, and sound sources are ordinary happenings like collisions and vibrations, then audition's objects might include things that are visible. The other objection is that speech perception is not unique in being subject to influence from multiple senses. Cross-modal recalibrations and illusions are rampant. The ventriloquist illusion shows that vision impacts non-linguistic audition. The motion bounce effect and the sound-induced flash illusion show that non-linguistic audition alters visual experience. Visual capture and the rubber hand illusion show that vision affects touch and proprioception. And the touch-induced flash shows that touch alters

vision. The examples multiply (for references and discussion, see, e.g., Spence and Driver, 2004; O'Callaghan, 2011b; Chapter XX, this volume). In many such cases, the best explanation for some cross-modal effect is that perceptual modalities share common objects (O'Callaghan, 2008; 2011b). Consider the sound-induced flash illusion. When presented with a single flash accompanied by two beeps, many subjects illusorily visually experience two flashes instead of one as a result of the two sounds. This illusion occurs because an apparent conflict between vision and audition is resolved in audition's favor. Since even apparent conflict requires the assumption of a common subject matter, perceptual processes unfold as if a common environmental source produces both the visual and the auditory stimulation. Since, under such conditions, audition is more reliable for temporal features, the overall perceptual experience that results is as of two events rather than one. If, therefore, cross-modal effects support the claim that multimodal speech perception targets common objects of perception, cross-modal effects may support the claim that there are common objects of perception in multi-modal cases that do not involve speech. Such cross-modal effects thus offer additional support for the claim that non-linguistic audition reveals the sources of sounds, which also are visible. Multimodality is not unique to speech.

The third type of argument stems from the received view that speech perception is *categorical*. Some have argued that the categorical nature of phoneme perception (see section 3) shows that its objects are not ordinary sounds, since ordinary sounds need not be perceived categorically (for discussion, see, e.g., Trout, 2001; Pinker and Jackendoff, 2005; for a critical perspective, see, e.g., Diehl et al., 2004). It is true that some attributes of sounds, such as loudness or pitch height (cf. pitch chroma), are not perceived categorically. Nevertheless, there are several lines of response to the argument from categorical perception. First, categorical perception may be limited to certain types of phonemes, such as stop consonants, so not all phoneme perception is categorical. Second, non-linguistic audition may involve categorical perception if speech perception does. Third, non-linguistic creatures, such as quail and monkeys, perceive some speech sounds categorically (see, e.g., Diehl, Lotto, and Holt, 2004, p. 177). Finally, color perception commonly is regarded as categorical, but this does not establish that the objects of color vision differ from the objects of ordinary vision. Categorical perception for selected phonemes therefore does not show that the objects of speech perception and the objects of non-linguistic audition differ in kind.

Arguments from mismatch, cross-modal influence, and categorical perception thus do not show that the objects of speech perception differ in nature from the objects of ordinary audition.

Sounds are among the objects of auditory perception. But to deny that the objects of speech perception include sounds would require denying that spoken utterances may perceptually appear to have pitch, timbre, and loudness. Nonetheless, the considerations discussed above do support the claim that the objects of speech perception include events or happenings beyond sounds, such as the articulatory gestures of speakers. However, I have maintained that environmental happenings that make or have sounds also are among the objects of non-linguistic auditory perception. For instance, while you hear the crashing sound, you also may hear the collision that makes it. Thus, in speech perception and in general audition, both sounds and sound sources plausibly are among the objects of perceptual awareness.

Suppose one held that phonological features of perceptible speech, such as phones and phonemes, themselves were the objects of speech perception. Since phonological features are not individual sounds, one might be tempted to hold that the objects of speech perception differ from the objects of non-linguistic audition.

This would be a mistake. It conflates the broad and the narrow ways to understand the objects of perception. I have been discussing the narrow understanding of the objects of perception as individuals that bear perceptible attributes. Phonological features as I have characterized them may be among the objects of perception in the broad sense, but they are not objects of perception in the narrow sense.

The account I have offered denies that phones and phonemes are novel perceptible *objects*, understood as *items* or *individuals*, wholly distinct from audible sounds and articulatory events. It maintains instead that phonological features, including specific phones and phonemes, are perceptible *properties* or *attributes* of audible and multimodally perceptible objects, such as sounds and articulatory events. Thus, for instance, a stream of utterances may perceptually appear to have, to bear, or to instantiate phonological attributes, such as *[d]* or */d/*. Such perceptible linguistic features may be complex properties, and they may have complex relationships to simple acoustical, physical, or physiological properties. They may be common sensibles. One important virtue of this account is that it allows us to abandon the troublesome "beads on a string" model of perceptible phonemes and to accommodate coarticulation. It does so because continuous sound streams or gestural events may perceptually appear at certain moments to instantiate multiple phonological attributes. Rather than perceptually appearing as discrete perceptible items or individuals arranged in a neatly segmented series (like typed letters in a written word), phonological properties of continuously unfolding spoken utterances may instead appear to be instantiated in connected, blended, or overlapping sequences by a common

perceptible individual.

The objects of speech perception thus need not be wholly distinct from the objects of non-linguistic audition. Each may include sounds and happenings in the environment that ordinarily are understood to be the sources of sounds. In the specific case of speech, the objects of perception may include sounds of speech and gestures used to articulate spoken language. In a broad sense, they also may include phonological features.

5 Processes

What are the implications for questions about *how* humans perceive speech—about the *means* or *mechanisms* involved in speech perception? Does the perception of speech involve special processes, a special module, or perhaps even a special perceptual modality?

There is evidence that perceiving speech sounds does involve distinctive perceptual processes beyond those involved in hearing non-linguistic sounds. Duplex perception for dichotic stimuli shows that a single stimulus presented to one ear can, in conjunction with information presented to the other ear, contribute simultaneously to the perceptual experience as of both a non-linguistic sound and an apparently distinct speech sound (Rand, 1974). The same acoustic cue is integrated into two distinct percepts. Duplex perception is thought by some to provide evidence for a special system or mode of listening for speech. That is because, under similar experimental conditions with only non-speech tones, masking rather than integration takes place. However, duplex perception does occur for complex non-linguistic sounds, such as slamming doors, so others have responded that speech perception does not involve dedicated perceptual processes distinct from general audition (Fowler and Rosenblum, 1990). Nevertheless, the capacity to perceive non-linguistic sounds does differ developmentally from the capacity to perceive speech. Notably, for instance, the timing of critical periods for the development of linguistic and non-linguistic perceptual capacities differs. In addition, functional neuroimaging establishes that the patterns of brain activity associated with the perception of speech sounds do not match those associated with the perception of non-linguistic sounds. Most tellingly, however, perceptual capacities and disorders related to speech may dissociate from those related to non-linguistic audition. The example of pure word deafness discussed above puts this into relief. Individuals with PWD have intact abilities to hear and to recognize ordinary sounds but are unable to hear and recognize speech sounds as such. In addition, auditory agnosia concerning environmental sounds may leave linguistic capacities intact (Saygin et al., 2010). This shows that one could auditorily perceive speech while lacking other commonplace auditory

capacities. Thus, there is evidence to support the claim that there exist perceptual resources and processes devoted to the perception of speech.

Some have held on such grounds that, when compared with general, non-linguistic audition, speech perception is special in that it is *modular* (e.g., Fodor, 1983). Others even have claimed that it involves a special perceptual *modality* (Liberman, 1996). I am reluctant to accept the strong view that speech perception involves a dedicated perceptual modality that is distinct from general audition and vision. Audition and vision may treat speech sounds and spoken utterances in a manner that differs from non-linguistic sounds and events, but this does not show that speech perception is a novel perceptual modality. Vision, for instance, devotes special resources and deals in different ways with the perception of objects, color, motion, and shape. Still, there is considerable debate concerning how to count and individuate perceptual modalities. We might identify modalities by their distinctive objects, stimuli, physiology, function, or phenomenology, or by some combination of these criteria. In the case of the classic sense modalities, at least, the criteria tend to align. Some have maintained that we should be pluralists when individuating and counting sense modalities (Macpherson, 2011). Maintaining that speech perception involves a novel perceptual modality nevertheless requires appealing to one or more of the criteria. None of these criteria, however, warrants positing a modality devoted to the perception of speech that is distinct from but on a par with the familiar examples of vision, hearing, smell, taste, and touch. For instance, speech perception does not involve awareness of novel perceptual objects, and it lacks proper sensibles inaccessible to other modalities. Speech perception lacks a distinguishing kind of proximal stimulus, and it lacks a dedicated sense organ and receptors. Its functional relations do not clearly mark it off as a wholly distinct way or manner of perceiving independent from audition or vision. And it is not apparent that its phenomenology has the type of proprietary, internally unified qualitative character that is distinctive to other perceptual modalities. For instance, while the phenomenology of other sensory modalities doubly dissociates, speech perception requires auditory or visual phenomenology. Despite these indications, however, a more satisfactory theoretical understanding of the modalities of sensory perception will help to make progress on this question (see, e.g., Matthen, this volume).

The weaker claim is that speech perception is modular. But good reasons also exist to doubt that a devoted perceptual module is responsible for the perception of speech. Appelbaum (1998), for instance, argues forcefully against Fodor that domain general, top-down influences impact the perception of speech sounds. If a process is modular only if it is informationally

encapsulated, then speech perception is not modular.

Perhaps it is possible to make do with a *minimal* story about the sense in which the processes associated with speech perception are special without appealing to a perceptual modality or even a perceptual module devoted to the perception of spoken language. Such a story may be framed in terms of our perceptual *treatment* of speech and speech sounds. Humans do have a special or differential selectivity or sensitivity for the sounds of speech. The striking evidence is that even neonates distinguish and prefer speech to non-speech sounds (Vouloumanos and Werker, 2007). The sounds of spoken utterances are of special *interest* to us, relative to other kinds of environmental sounds and events.

Humans are not, however, born able to perceive all of the attributes that are distinctive to specific languages. Infants must prune and cease to perceive audible differences that are not linguistically significant in their own languages. They also must learn perceptually to discern linguistic sameness in the face of variation across speakers, moods, and contexts. This is learning perceptually to ignore irrelevant differences and to attend to crucial similarities, and it alters the language-specific perceptual similarity space involving speech sounds. Understanding a language, as it is spoken in a variety of contexts, demands such learning. In coming to know a spoken language, we begin to perceive the relevant language-specific features of sounds and utterances. Humans thus have a *propensity* for learning perceptually to discern the appropriate *language-specific* types to which spoken utterances belong.

6 What makes speech special?

Perceiving the attributes that are distinctive to the speech sounds of a given language, I have argued, requires experience and learning. Learning a language thus is not simply a matter of learning a sound-meaning mapping. It involves acquiring the capacity perceptually to discern language-specific attributes of spoken utterances. In this sense, you learn to hear the sounds of your language. Learning a language is a partly a matter of acquiring a perceptual skill.

Humans have a special propensity to learn to perceive language-specific attributes of speech sounds from birth, but this capacity develops later than other familiar perceptual capacities. For instance, young infants perceive individual objects and events, persistence, and sensible qualities, including color, pitch, and loudness, prior to perceptually discerning types of sounds that are specific to a particular language. Perceptual awareness of spoken language therefore may be more like perceptual awareness of clapping of hands, barking dogs, or fingernails scratching a chalkboard, each of which involves acquired perceptual skills.

As with other auditory phenomena, the manner in which language-specific sounds are perceptually individuated and classified is illuminated by taking into account the environmental happenings that generate sounds. In particular, articulatory gestures and talking faces make sense of why users of a given language discern and treat various speech sounds as standing in relations of similarity and difference that do not stem in straightforward ways from acoustical characteristics. Considered as such, perceiving speech is a matter of detecting and discerning biologically significant kinds of sounds and happenings, rather than just detecting abstract features of an acoustic signal.

How does perceiving speech differ from perceiving other biologically significant kinds of environmental sounds? Consider a family of perceptual capacities attuned to varieties of *animacy*. For instance, humans sometimes may perceptually experience a pattern of moving dots as *running*, or seem to be aware of one dot *chasing* another dot around a display (Heider and Simmel, 1944; see Scholl and Tremoulet, 2000; Gao et al., 2009). Here we describe the perception of inanimate things and motion in terms applicable to animate creatures and activities. Since such effects require only very minimal cues, this suggests humans have a special propensity to perceive aspects of animate creatures and their activities. That is, we have differential sensitivity to certain kinds of *activity* that creatures engage in, in contrast to simple mechanical patterns of motion traced by inanimate things. Perceiving speech is similar to such perceptual capacities in that its concern is a type of *animacy* exhibited by living things to which we have special sensitivity. In the case of speech (as in the case of *faces*) this perceptual capacity is directed predominantly at members of our own species.

Speech perception belongs to an even more special subclass. Speech sounds are generated by *communicative intentions* of other humans. Like some facial expressions and non-linguistic vocalic sounds, the sounds of spoken utterances are caused by and thus have the potential to reveal the communicative intentions of their animate sources. Speech perception is among a class of ethologically significant perceptual phenomena that serve to disclose intentional activities involved in communication. Perceiving speech is detecting and discerning language-specific kinds of biologically significant events: those generated by communicative intentions of fellow human talkers. We hear people talking. We hear them as interlocutors.

Acknowledgements

I have learned a great deal about the philosophical issues raised by speech perception from Matthen (2005), Rey (2007), Mole (2009), Remez and Trout (2009), and Smith (2009). These

works, and conversations with their authors, drew me from my more general concern with sounds, audition, and multimodality to the philosophically and empirically rich subject matter whose focus is the perception of spoken language. I gratefully acknowledge their influence upon my approach to this topic. Thanks to Mohan Matthen for helpful comments on this chapter.

References

Appelbaum, I. (1998). Fodor, modularity, and speech perception. *Philosophical Psychology*, 11(3):317–330.

Appelbaum, I. (1999). The dogma of isomorphism: A case study from speech perception. *Philosophy of Science*, 66:S250–S259.

Bayne, T. (2009). Perception and the reach of phenomenal content. *The Philosophical Quarterly*, 59(236):385–404.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, 55:149–179.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968):303–306.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14:3–28.

Fowler, C. A. and Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):742–754.

Gao, T., Newman, G. E., and Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59:154–179.

Gick, B. and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, 462(7272):502–504.

Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition*. Cambridge University Press, Cambridge, UK.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298:1569–1579.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259.

Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. MIT Press, Cambridge, MA.

Liberman, A. M. (1996). *Speech: A Special Code*. MIT Press, Cambridge, MA.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21:1–36.

Macpherson, F. (2011). Taxonomising the senses. *Philosophical Studies*, 153(1):123–142.

Matthen, M. (2005). *Seeing, Doing, and Knowing: A Philosophical Theory of Sense Perception*. Oxford University Press, Oxford.

McDowell, J. (1998). *Meaning, Knowledge, and Reality*. Harvard University Press, Cambridge, MA.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

Mole, C. (2009). The Motor Theory of speech perception. In Nudds, M. and O'Callaghan, C., editors, *Sounds and Perception: New Philosophical Essays*, chapter 10, pages 211–233. Oxford University Press, Oxford.

Nudds, M. (2010). What are auditory objects? *Review of Philosophy and Psychology*, 1(1):105–122.

O'Callaghan, C. (2008). Seeing what you hear: Cross-modal illusions and perception. *Philosophical Issues: A Supplement to Noûs*, 18:316–338.

O'Callaghan, C. (2011a). Against hearing meanings. *Philosophical Quarterly*, 61:783–807.

O'Callaghan, C. (2011b). Perception and multimodality. In Margolis, E., Samuels, R., and Stich, S., editors, *Oxford Handbook of Philosophy and Cognitive Science*. Oxford University Press, Oxford.

Pinker, S. and Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, 95:201–236.

Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science*, 25:679–693.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, 55:678–680.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carell, T. D. (1981). Speech perception without traditional speech cues. *Science*, (212):947–950.

Remez, R. E. and Trout, J. D. (2009). Philosophical messages in the medium of spoken language. In Nudds, M. and O'Callaghan, C., editors, *Sounds and Perception: New Philosophical Essays*, chapter 11, pages 234–264. Oxford University Press, Oxford.

Rey, G. (2007). Externalism and inexistence in early content. In Schantz, R., editor, *Prospects for Meaning*, volume 3 of *Current Issues in Theoretical Philosophy*. de Gruyter, New York.

Saygin, A. P., Leech, R., and Dick, F. (2010). Nonverbal auditory agnosia with lesion to Wernicke's area. *Neuropsychologia*, 48:107–113.

Scholl, B. and Tremoulet, P. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8):299–309.

Siegel, S. (2006). Which properties are represented in perception? In Gendler, T. S. and Hawthorne, J., editors, *Percepual Experience*, pages 481–503. Oxford University Press, New York.

Smith, B. (2009). Speech sounds and the direct meeting of minds. In Nudds, M. and O'Callaghan, C., editors, *Sounds and Perception: New Philosophical Essays*, chapter 9, pages 183–210. Oxford University Press, Oxford.

Spence, C. and Driver, J., editors (2004). *Crossmodal Space and Crossmodal Attention*. Oxford University Press, Oxford.

Trout, J. D. (2001). The biological basis of speech: What to infer from talking to the animals. *Psychological Review*, 108(3):523–549.

Vouloumanos, A. and Werker, J. F. (2007). Listening to language at birth: evidence for a bias for speech in neonates. *Developmental Science*, 10(2):159–164.