Is speech special?

Casey O'Callaghan
Rice University, Philosophy Department

There is a thriving debate over what aspects of our capacity to produce and understand language are special. My concern here is a key part of this wider debate: Is speech special? In particular, my focus is on speech perception, and whether it is special. This isn't just one but a number of different questions. Too frequently, these very different questions are not clearly distinguished and kept apart. I discuss a framework for distinguishing various versions of the question, Is speech perceptually special? Focusing on a particular class of questions, I make a proposal about the sense in which speech is perceptually special. According to this account, the capacity to perceive speech is an acquired perceptual skill, and involves learning to hear language-specific types of biologically-significant sounds. This account illuminates the significance of interlocution in understanding what makes the perception of speech distinctive.

1 Is speech perceptually special?

There is a thriving debate over whether the faculty of language is special (see, e.g., Hauser et al. 2002; Pinker and Jackendoff 2005). The question is if our capacity to produce and understand language is special. A key part of this wider debate is whether speech is special (see, e.g., Liberman 1996; Trout 2001). My concern here is speech *perception*. Is speech perception special?

Of course, answering this question means answering not just one but a number of different questions. Too frequently, these different questions are neither clearly distinguished nor kept apart. My goal is to present a framework for distinguishing various versions of the question, Is speech perception special? Focusing on a particular class of questions as an illustration, I make a proposal about the sense in which speech is perceptually special. This account illuminates the significance of interlocution for understanding what makes the perception of speech distinctive.

2 Many questions

To be special requires at least a *difference*. This raises the question, How different? The debate about whether speech is special aims at a stronger claim than that perceiving speech somehow differs from other perceptual capacities. Usually, it concerns whether speech perception is *distinctive*, or can be distinguished as a distinct variety among other forms of perception. Often, the question considered is stronger yet. Is speech perception *unique*, or the only instance of its kind?

Put this way, the question relies on a comparison. The most common contrast is with *general audition*. Is speech perception to some degree different from or unique in contrast to *non-linguistic* (or "*ordinary*") *audition*? A separate contrast (too often used interchangeably with the first) is to the capacities of *non-human animals*. Are humans alone in having the capacity to perceive speech? Is speech perception uniquely human?[1]

Furthermore, a difference is a difference in some respect, and being distinctive or unique is being distinctive or unique in some way or for some reason. So the most important question is, In what respect is speech perception special? Here there are a number of candidates. I find it helpful to divide them into two broad classes.

The first class of questions (Type 1, or "What", questions) are forms of the question, *What* do we

1 Another contrast might be made between speech perception and *perception in general*. Is perceiving speech different in kind from or unique among varieties of perception? Finally, we might consider the contrast between *perceiving* speech and *understanding* speech.

perceive when we perceive speech? This first class includes questions about the *phenomenology*, *objects*, and *contents* of speech perception. Is there something special about the phenomenology, or the experience, of perceiving speech? Does speech perception have special objects? Does the perception of speech involve special contents?

The second class of questions (Type 2, or "How", questions) are forms of the question, *How* do we perceive speech? Or, What are the *mechanisms* of speech perception? This second class includes questions about the *processes*, *module*, or *modality* involved in speech perception. Does perceiving speech involve special perceptual processes? Does speech perception involve a special perceptual module? Is speech perception itself a special perceptual modality?

The question, Is speech special?, thus becomes: Is human speech perception different, distinctive, or unique, when compared to non-linguistic audition, or to the perceptual capacities of non-human animals, in respect of its phenomenology, objects, or contents, or in respect of the processes, modules, or modality it involves?

I won't tackle all of this here. Instead, I'll focus on the Type 1 questions, which concern *what* we perceive when we perceive speech. Part of the reason for this is historical. There is a history of using answers to *what* questions to ground claims about *how* we perceive speech. Liberman (see, e.g., 1996) famously argued that the objects of speech perception differ from the objects of non-linguistic audition, and used this to argue that speech perception and non-linguistic hearing involve different perceptual modalities. This approach makes sense. We need to be clear about what the task is in order to understand what is required to perform it. So, it is particularly important to get the answers to Type 1 questions right.

3 Phenomenology

Is perceiving speech phenomenologically special? Is what it's like for the subject to perceptually experience speech different or unique in relation to non-linguistic audition?

It's common to think that the perceptual experience of speech is phenomenologically distinctive. Introspectively, there is a *perceptual* phenomenological difference between the experience of listening to speech and listening to non-speech sounds. But it can be difficult to motivate a phenomenological claim. Since we are good at detecting phenomenological contrasts, that is a good way to start.

Consider the contrast between listening to non-linguistic sounds and listening to spoken language. First, take the case of a language you know. It strikes me that there is a significant qualitative difference between the experience of listening to a language I know and the experience of listening to non-linguistic environmental sounds. Sinewave speech seems to confirm this. The same stimulus first is experienced as non-speech sounds, and then it is experienced as speech. However, in this case you *comprehend* the speech. Understanding might suffice to explain the phenomenological difference in listening to speech in a language you know. You grasp meanings, so the experiential difference could be explained in terms of cognitive, rather than perceptual, phenomenology.

So, consider listening to non-speech sounds in contrast to listening to speech in a language you do not know. Is there a perceptual phenomenological difference? It *seems* to me that there is some difference. The fact that neonates perceptually prefer speech sounds from non-speech sounds hints that there's a difference, since neonates do not yet understand language. But reflection and babies are inconclusive here. Is the difference due to a difference in strictly audible characteristics of the stimulus, or does some further perceptual phenomenological difference accrue in virtue of its seeming like speech? If you could experience sinewave speech in a language you don't know either as non-speech sounds or as speech sounds, that would provide good evidence for a perceptual phenomenological difference.

But consider the contrast between listening to speech in a known language and listening to speech in an unknown language. To control for differences in the stimulus, make it the same language. So, consider either the experiences of one person listening to speech in some language before and after learning the language, or consider the experiences of two similar listeners, one who knows the language and the other who doesn't. While in each case there is a cognitive difference, that does not suffice to capture the phenomenological difference. There now also is good evidence for a perceptual phenomenological difference. While the stimulus is the same, perceptually experiencing speech in a language you know differs in a couple of respects. First, and most obviously, it has different temporal characteristics. You hear (and perhaps exaggerate) gaps and pauses, and better resolve temporal features

and contrasts. And, it has different qualitative characteristics. You detect subtle changes, contrasts, and differences in qualitative features you couldn't hear before. The stimulus *sounds* different when you recognize it as speech and know the language.

This all suggests that there is a *perceptual* (and not merely cognitive) phenomenological difference between listening to speech in a language you know and listening either to speech in a language you do not know or to non-speech. So there's good reason to think that speech perception has a distinctive phenomenology compared to non-linguistic audition.

4 Objects

The perceptual phenomenological difference might suggest that *what* you perceive when you perceive speech differs from what you hear when you listen to non-linguistic sounds. One kind of difference in what's perceived is a difference in the *objects* of perception. So, the phenomenological difference might suggest that the objects of speech perception differ from those of non-linguistic audition. In fact, researchers commonly have claimed that the objects of speech perception differ from those of audition generally. Evaluating the suggestion requires asking about the objects of speech perception and audition.

What are the intentional objects of speech perception? One primary focus of research into speech perception has been upon *phonemes*, whose patterns form the basis for recognizing and distinguishing words.

What is a phoneme? Phonemes are the minimal linguistically significant ways in which spoken words in a language differ perceptually. They're like the basic perceptible vocabulary that comprises spoken words in a language. For instance, the spoken word 'bad' includes /b/, /æ/, and /d/, while 'bat' and 'bash' differ because the former contains /t/ and the latter contains /ʃ/.

Phonemes are language-specific. *Phones*, on the other hand, include all of the perceptually discernible differences that can be linguistically significant in human languages. Phones are individuated in terms of the auditorily discernible (humanly producible) differences that could be exploited by a spoken language to signal a linguistically significant difference. Thus, phones are *types* that comprise auditorily equivalent perceptual objects. If audition's objects are sounds, phones are perceptually equivalent sound types. Phonemes, then, are classes of phones that, even though they might strictly speaking be perceptually distinguishable, are treated *as equivalent* within the context of a given spoken language. Hearers need not decide to treat strictly discernible sounds as members of a common type. When listening to a known spoken language as such, the listener cannot help but treat different phones as *allophones*, or as instances of a common phoneme. They are perceived as the same. On a natural picture, phonemes are perceptual equivalence classes of phones. Thus, the phonemes you perceive in perceiving speech are auditory equivalence classes of sounds. Phonemes are *sound types*.

Do the objects of speech perception differ from those of ordinary non-linguistic audition? Speech construed as such involves different *kinds of sounds* from non-linguistic environmental sounds, such as those of clucks, beeps, doors closing, cars backfiring, hands clapping, and dogs barking. But, according to the proposal above, the objects of speech perception and auditory perception belong to the same *ontological* kind, since both are types of *sounds*.

Nevertheless, the claim that speech perception and audition have different objects traditionally isn't just the claim that they involve hearing different *kinds of sounds*. It is the claim that perceiving speech is perceiving some object of an entirely different ontological kind from ordinary sounds. It is perceiving non-sounds. Many have argued that speech perception's objects do differ from non-linguistic audition's in this stronger sense. Three main sorts of argument are offered.

The strongest appeals to the *mismatch* between salient aspects of the experience of speech and features of the acoustic signal. The *acoustic* features that ground a perceived phoneme are highly context dependent. Not only do they vary in expected ways, with speaker, mood, and accent, but they also depend more locally upon the surrounding phonemes. Thanks to the effects of *coarticulation*, information about a given phoneme is blended with information about surrounding phonemes, and differs depending on the adjacent phonemes. While we experience /du/ and /di/ to share the sound of /d/, there is no invariant acoustic signature that corresponds to the /d/. Furthermore, while we experience a speech stream to be *segmented* into discrete phonemes and words, the acoustic information that cues /æ/ in 'dab' is present

during the articulation of both the /d/ and the /b/. There are no clear acoustic boundaries that correspond to those between experienced phonemes. In sum, there is no consistent, context-independent homomorphic mapping between experienced phonemes and straightforward features of the acoustic signal.

In light of this, Liberman and other proponents of the Motor Theory suggested that speech perception's objects are not sounds at all, but instead are aspects of the articulation of speech. The idea is that the gestures involved in the production of speech do map in a homomorphic, invariant way onto perceived speech. For instance, pronouncing /d/ involves stopping airflow by placing the tongue at the front of the palate, then releasing it while activating the vocal folds. Such gestures, and the features that combine to make them, make intelligible the perceptual individuation of speech in a way that the acoustic signal does not. The objects of speech perception and of audition thus differ in kind. The former are gestures, the latter are sounds.

This argument fails to establish that the objects of speech perception differ from the objects of non-linguistic auditory perception. On one hand, it relies on the premise that ordinary audition *does* map in an invariant and homomorphic way onto straightforward features of the acoustic stimulus. Even a simple audible quality like pitch has a complex relationship to frequency. And, especially in acoustically complex environments, the things we hear do not map straightforwardly and in an invariant way onto acoustic features. Context effects abound. For instance, the timbre of an enduring sound will differ if its attack differs. Furthermore, a 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 acoustic information (e.g., Bregman 1990). Nothing obvious in an acoustic stream signals how to distinguish the sound of a guitar from the sound of a voice in a crowded bar.

On the other hand, it relies on the premise that ordinary audition's objects *do not* map in an illuminating way onto aspects of the events and happenings that produce acoustic information. However, the sounds we hear are individuated in terms of features of their sources. This is reflected in how we talk about sounds: the sound *of the car door*, the sound *of the dog*, the sound *of scratching*. We carve up the auditory scene in large part in terms of sound sources and happenings. While the individuation of the objects of speech perception is illuminated by considering aspects of the articulatory gestures involved in speaking, this mirrors the fact that the individuation of the objects of non-linguistic auditory perception is illuminated by considering aspects of the happenings that make sounds. The function of auditory perception is to make perceptually accessible environmentally significant events. Sounds, among the objects of audition, are individuated not just in terms of the simplest physical properties of an acoustic stimulus, but in a way that facilitates awareness of sound sources. In this respect, speech perception does not differ in kind from non-linguistic audition. The *mismatch* argument fails.

Second, some argue that *cross-modal influences* in the perception of speech, such as the McGurk effect, reveal a substantial difference in the objects of speech perception and the objects of ordinary audition (see, e.g., Trout 2001 for discussion). Does the fact that visual information can impact which phoneme you experience show that, unlike sounds, which you can only hear, the objects of speech perception are available to both vision and audition? But in this respect speech is not unique. Cross-modal illusions and influences are rampant. Vision impacts non-linguistic audition (ventriloquism), ordinary audition impacts vision (sound-induced flash), vision alters touch (visual capture), touch alters audition, and so on (see, e.g., Spence and Driver 2004). Explaining each requires positing shared items among the perceptual objects of different modalities if the speech case does (O'Callaghan 2008). In fact, cross-modal effects support conceiving of audition's function as revealing the things and happenings that make sounds. Multimodality is not unique to speech.

Third, speech perception is *categorical*. That is, gradually varying a physical acoustical parameter leads to uneven perceptual variation. Where the effect is quite exaggerated, varying a physical parameter gradually might, for instance, cause one first to experience /b/ and then abruptly to experience a /d/ with little change noticed in between. In a dramatic case, an "analog" signal might cause apparently "digital" perception. Some have argued that the categoricity of phoneme perception means its objects are something other than ordinary sounds (see, e.g., Trout 2001; Pinker and Jackendoff 2005 for discussion). But, we now know that speech perception is not alone in being categorical. For instance, color perception is categorical, other forms of audition are categorical, and non-human animals show evidence of

perceiving categorically (see Harnad 1987; Cohen and Lefebvre 2005). The categoricity of phoneme perception cannot ground an argument that the perceptual objects of ordinary audition and speech perception differ in ontological kind.

If no good argument shows that the objects of speech perception and ordinary audition belong to entirely different kinds, what is the difference between perceiving speech and perceiving ordinary environmental sounds? If speech is just a variety of sounds that are perceptually individuated in terms of features of their sources, what explains the perceptual phenomenological difference between hearing speech and hearing ordinary sounds?

5 Contents

What we perceive when we perceive speech does, in another sense, differ from what we perceive when we perceive non-speech sounds. Speech perception has different *content* from non-linguistic audition. Content concerns how things are represented to be, or what features objects are perceptually attributed. One gloss on the contents of perception is in terms of the accuracy or veridicality conditions of a perceptual state. Another way to put this is in terms of what a given experience purports to be facts about the world (see Siegel 2005). My claim is that the content of speech perception differs from that of non-linguistic audition in two noteworthy respects.

First, the audible features of a stimulus perceptually experienced as speech differ from those of non-speech sounds. Experiencing speech involves experiencing certain fine-grained qualitative and temporal details that one does not normally experience when listening to even the same stimulus as non-speech. Speech experience discerns more and different *qualitative* details and contrasts than non-linguistic audition. Speech also audibly appears to have different and finer-grained *temporal* features—we hear gaps and pauses we didn't hear before. Further, we *segment* the sound stream differently over time when we hear it as speech. The individuation of sounds in time differs when we hear those sounds as speech. What formerly sounded like a continuous babble comes to sound like discrete sounds, syllables, and words. If sounds are audible individuals, we *hear different sounds* when we hear speech.

The second difference stems from the fact that human speech perception is *categorical*. While physical acoustical features vary along a continuum, phonemes are language-specific classes defined by perceptual equivalence of a certain sort. Belonging to a given phoneme category is an all-or-nothing matter, so perceiving phonemes is a kind of classificatory perception (cf. Matthen 2005). What consequences does this have for characterizing the content of speech perception? Perceiving speech is hearing sounds in a way that is consistent with their belonging to language-specific categories. Since these categories are equivalence classes of sounds, perceiving speech involves hearing sounds to stand in certain relations of similarity and difference to each other. These patterns of similarity and difference form a speech-specific (and language-specific) similarity space among sounds, in which regions correspond to particular language-specific speech sounds such as phonemes. In fact, one way to characterize these speech-specific categories is in terms of the speech-specific similarity relations among sounds. In hearing speech, the features sounds are perceived to have match this speech-specific similarity space among sounds. The content of speech perception, which grounds its difference from non-linguistic audition, reflects a distinctive pattern of similarity and difference among sounds.

6 *How* questions

What are the implications for questions about *how* we perceive speech? Type 2 questions concern the *means* or *mechanisms* involved in speech perception. Specifically, they ask whether speech perception involves special processes, a special module, or a special modality.

The evidence strongly suggests perceiving speech involves at least some special perceptual processes. Duplex perception for dichotic stimuli, developmental (especially critical period) differences, brain activity revealed by functional imaging, and dissociated disorders for speech and auditory perception all provide evidence of processes devoted to the perception of speech (see, e.g., Trout 2001).[2]

2 Nevertheless, we should take care. Rich physiological and functional connections exist between general auditory and multimodal areas and language-specific areas. So, it is not always entirely clear whether some activity or

But, perhaps we can make do with a *minimal* story about the sense in which speech perception is special without appealing to special modules or even a special modality devoted to speech perception.[3]

This story is framed in terms of our *treatment* of speech and speech sounds. It involves two main claims, which are drawn from facts that must be accommodated by any adequate contemporary account of speech perception.

First, humans have a special or differential *selectivity* or *sensitivity* for the sounds of speech, in general. The striking evidence is that neonates distinguish and prefer speech to non-speech (Vouloumanos and Werker 2007). The sounds of speech in general are special for us, and they receive different treatment from other kinds of environmental sounds.

How do infants perceive speech sounds if speech sounds comprise *language-specific* classes of sounds? Humans are not born with the capacity to perceive phonemes *as such*. Very young infants in fact discern phonological differences from all languages—they distinguish among all of the possible speech sounds their language could include. Later, between 5-9 months, infants discern only the phonetic differences relevant to their own language. The usual story is that infants prune or forget how to perceive audible differences among phones that are not significant in their own language. Humans perceptually learn to ignore differences that are irrelevant to their language. Doing so is learning to treat one's language's allophones as such while losing the ability to distinguish sounds that other languages count as distinct phonemes. Such learning alters the language-specific similarity space among sounds, so we come to perceive sounds in a way that is consistent with their belonging to the relevant language-specific equivalence classes. We learn to discern the language-specific classes of sounds that comprise our language's phonemes. So, second, humans have a propensity for learning to perceive *language-specific* sound types.

Perceiving speech sounds from a known language, according to this understanding, requires experience and learning. It is an *acquired* perceptual skill. One learns to *hear* the sounds of one's language. Thus, learning a language is not just a matter of learning a sound-meaning mapping. It involves acquiring the auditory skill of *hearing* sounds in a way consistent with their belonging to language-specific perceptual equivalence classes. Learning a language is a partly a matter of learning a perceptual skill.[4]

Since the capacity to perceive speech sounds in accordance with language-specific categories is an acquired perceptual skill, it differs (at least in degree) in this respect from the capacities to perceive individuals such as three-dimensional objects and events, persistence, and sensible qualities like color, pitch, and loudness. Arguably, these are capacities humans possess much earlier. On the other hand, speech perception may be more like our capacities to perceive things like clapping hands, dog barking, metal scraping metal, or fingernails scratching a chalkboard. These are best understood as acquired perceptual capacities.

7 Interlocution

What is the role of interlocution according to this account? Hearing speech is an acquired perceptual skill for which we have a special propensity from before birth. It involves learning to perceive language-specific types or equivalence classes of sounds, whose individuation is illuminated by

process is perceptual or extra-perceptual. So it is not entirely clear whether perceiving speech involves special *perceptual* processes.

3 While I won't argue for it here, I'm reluctant to say that speech perception involves a distinctive or unique perceptual modality independent from ordinary audition. Whether we individuate modalities in terms of their objects, phenomenology, function, or physiology, the evidence doesn't require a separate modality to deal perceptually with speech. While speech certainly may be handled differently from non-linguistic sounds by audition, colors and objects also are handled in different ways by vision. Similarly, I doubt speech perception is accomplished by a devoted perceptual module. If a process is modular only if it is informationally encapsulated, then speech perception isn't a module. Appelbaum (1998) argues convincingly against Fodor that domain general top-down influences impact the perception of speech sounds. Further, as I've argued here, audition and speech perception to a significant extent share function.

4 It is no objection to the claim that perceiving speech is acquired or learned that perceiving speech requires a special propensity, which must be innate. So does walking on two legs.

considering salient happenings in the environment: articulatory gestures and talking faces. 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 sorts of environmental sounds? Consider a family of capacities that reveal varieties of *animacy*. For instance, we might perceive a pattern of moving dots as *running*, or one dot to *chase* another dot around a display (Heider and Simmel 1944; see also Scholl and Tremoulet 2000). Here we describe the perception of inanimate things and motion in terms applicable to animate things and activities on the basis of very minimal cues, which suggests we have a special propensity to perceive animate things and activities. Perceiving speech is similar to perceiving these other special sorts of biologically significant things and activities, in that its concern is a type of *animacy* exhibited by living things to which we have special sensitivity. That is, we have differential sensitivity to certain kinds of *activity* that creatures engage in, in contrast to simple motion patterns or inanimate happenings. Furthermore, in the case of speech, this capacity is directed at members of our own species, as is the capacity to perceive *faces*.

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

References

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

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

Cohen, H. and Lefebvre, C. (2005). *Handbook of Categorization in Cognitive Science*. Elsevier, New York.

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.

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

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

O'Callaghan, C. (2008). Seeing what you hear: crossmodal illusions and perception. *Philosophical Issues*, 18:316–338.

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

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

Siegel, S. (2005). The contents of perception. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*.

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.

Casey O'Callaghan
http://ocallaghan.rice.edu

UBCWPL

University of British Columbia
Working Papers in Linguistics

-Papers for the Interlocution Workshop-
Interlocution:
Linguistic structure and human interaction

Edited by:
Anita Szakay, Connor Mayer, Beth Rogers, Bryan Gick and Joel Dunham

July 2009 **Volume 24**