9

Speech Sounds and the Direct Meeting of Minds[1]

BARRY C. SMITH

Philosophers often claim that it is through speech that we make knowledge of our minds available to one another, and that it is through the medium of a shared language that we achieve a genuine meeting of minds. When combined with a conception of linguistic understanding as the direct perception of meaning in people's words, the view suggests that there is no barrier to knowing the minds of others. Certainly, when listening to a language we understand, we do not hear the acoustic speech signal as just a sequence of sounds: we hear what is being said. As a phenomenological observation, the claim is impeccable, but it is mistaken epistemologically to assume that when we hear words *as* meaningful it is because we hear meanings *in* the sounds, perceived as immediately present on the surface of speech. As I shall argue, talk of the surface of speech and the location of sounds is misplaced. What we directly perceive are the *sources* of sounds, and the source of speech sounds is the human voice. The experience of listening to speech gives us non-linguistic information about a voice as source, while the meanings we hear the voiced sounds to have are the meanings we as listeners have attached to those words: the meanings they have for us. Contrary to expectations, this inner model of linguistic understanding can still accommodate knowledge of what others are saying, and so presents no obstacle to our knowing what others have in mind.

[1] Versions of this chapter were given at the Second Dubrovnik Workshop in the Philosophy of Linguistics, in September 2006, and the Second Workshop in Philosophy of Language and Linguistics of the Irish Network of Philosophy of Language in University College Dublin in December 2006. I am grateful for comments from the audiences on both occasions, and for further discussion and comment to Ophelia Deroy, Jim Higginbotham, Guy Longworth, and Matthew Nudds, Georges Rey, Paul Pietroski, and Katerina Von Krikstein.

1. Introduction

Noise, or mere acoustic signals in the environment, can be heard as sounds in so far as there are agents of auditory perception. Perceivers can hear all manner of things: the sound of a taxi in the street, the sound of a door closing, of a dog barking, of a clock ticking, and so on. But there are also specific sounds that are heard not as the sound of objects, but as the sounds of subjects: most importantly, speech sounds. But not only these; groaning, laughing, and crying, like talking, all involve the voice in the production of what is a public display of a subject's inner experience. In what follows, I will concentrate on speech sounds and consider what enables us to hear them as the sounds of subjects, rather than objects. In particular, I shall argue that the basis on which an immediate connection between the mind of the speaker and listener is established is the experience of hearing *someone* talk rather than hearing *what they say*.

What enables us to hear some sounds as meaningful speech? The question is important, since philosophers have taken meaningful speech as the best guide to what another is thinking and the firmest evidence that they are thinking. But to understand the role that language understanding plays in the epistemology of mind, we must first give an account of the epistemology of understanding.

From an everyday perspective, language understanding is not problematic. People speak in a language we understand, and we immediately recognize what was said. It may not be obvious why they said it or what they mean in saying it, but we can recognize the words and sentences they uttered. In this respect, listening to speech in a language one understands is unlike hearing mere noise, although one can recreate that experience when listening to speakers of an utterly foreign language. In the foreign language case, we know people are saying something, or we think we do, though we cannot tell where one word begins or ends, or say which range of sounds makes up a complete phrase or sentence. As far as we are concerned, it sounds like babble, while to each of them their speech sounds are heard as sharing of intimacies, the idle talk of a moment, or requests for information. What distinguishes the two cases? Is it something in the sounds themselves, something in their reception by listeners, or some relation between the speaker and listener?

At first pass, it is easy to think that while listening to a foreign language in which one hears nothing of discernible significance, the speakers of that language are somehow adding meaning as an accompaniment to what is otherwise noise, whether anyone understands it or not. This gives the impression

that the speaker merely emits noise in the hope that others will attach some significance to it. But neither the speaker nor her listener views things in just this way. The speaker does not take herself to be merely producing sounds: she feels her mind to be fully revealed and on show in the choice of meaningful words she puts into the public sphere. And to a listener who can understand her, she is not heard as first making sounds that subsequently need decoding. Rather, she is heard as saying things, imparting information, asking questions, making demands. And even listeners who cannot understand what she is saying take the speech sounds she emits to be more than mere noise. They may feel unable to pick up what is going on in her speech, but they notice that others can immediately react to it.

Despite the ease of talking to people in a familiar language a puzzle remains. Just how do we understand one another's speech so easily? How do we make the contents of our minds publicly available to one another through our talk? In listening to speech in a language we understand, we seem to have direct access to what someone is saying. But how is such knowledge possible from the mere fact of listening to particular sounds? How, from these unpromising materials, can we be put in touch with other people's intended meanings? The fact that, environmentally speaking, we are only presented with sounds is evident when reflecting on the case of an utterly foreign language. In such circumstances one hears, not words, but a continuous sound stream interrupted when the speaker pauses for an intake of breath. The difficulty becomes clear when we realize that the only publicly available evidence we have to go on in understanding one another is observable behavior. Strictly speaking, all we can show to one another is a sequence of sounds, gestures, and facial expressions. So how do we succeed in communicating something with a precise linguistic significance on the basis of these unpromising materials? How can the noises and movements we make convey something of semantic significance to others?

If we contemplate the surface of observable behavior in a way consistent with the intuition that all that can be shown there are sounds and movements, we will probably adopt a description of linguistic behavior cast in the restricted, physicalist terms W. V. Quine recommends. And if we then try to locate meaning among these observable facts, we will be forced to reconstruct the content of speech in terms of mere patterns in verbal behavior. The sounds speakers produce will be conceived as a response to a stimulus. The range of a speaker's assents and dissents to a sound in observable circumstances will fix the *stimulus meaning* for these vocables (Quine 1960: ch. 2). This drastically impoverished notion of linguistic meaning results from the attempt to meet a publicity of meaning requirement: the obligation to show how meaning can

be publicly available to others as a matter of observed behavior in observable circumstances. The requirement must be met, according to Quine, since what people display in their behavior is all we have to go on in interpreting their utterances. However, the notion of stimulus meaning fails to square with the rich phenomenological experience we enjoy in listening to speakers talk in a language we understand. As noted already, we hear such speakers as saying things more determinate in meaning than is suggested by the limitations of stimulus meaning. And if we respect this latter intuition, and adhere to our everyday and common-sense experience of the meaningfulness of speech, we will need to find some other way to accommodate linguistically communicated meanings.

At this point, some will be tempted to locate meaning not on the surface of linguistic behavior but behind it, in the mind of the speaker. According to such a view, the speaker's intended meaning is hidden and becomes a matter of hypothesis for others. This new picture accepts, along with the previous one, that all we can show in speech behavior is a sequence of sounds and movements, and further accepts that these materials are insufficient to reconstruct the rich notion of linguistic content we are all familiar with as language users. It then attempts to save the fullness of linguistic meaning by locating it behind the surface of linguistic behavior in the mind of the speaker. The price to pay for respecting our commitment to determinate linguistic meanings is to abandon the publicity of meaning. But a view according to which 'assigning a meaning to an utterance by a speaker of one's language is forming a hypothesis about something concealed behind the surface of his linguistic behavior' (McDowell 1998b: 252) is unacceptable, according to John McDowell, Michael Dummett, and others, because it makes our understanding of one another 'a mere matter of guesswork as to how things are in a private sphere concealed behind their behavior'. And as McDowell points out, such a position distorts our immediate recognition of the meaningfulness of speech in a language we understand. As has been stressed already, in such cases hearers do not find themselves listening to uninterpreted sounds. McDowell's phenomenological insight is clearly right. Listening to speech in a language we understand is not a matter of first hearing noises and then going on to infer what they must signify. Rather, as McDowell says: 'Our attention is indeed drawn to … something present in the words—something capable of being heard or seen in the words by those who understand the language' (1998a: 99). The content of others' speech is not hidden beneath the surface of overt behavior. We hear people not merely as producing sounds but as saying something. It is part of this phenomenological insight that we cannot turn what we hear people saying back into sounds.

McDowell now looks for a middle way between the two unacceptable positions just described, and attempts, on the basis of his insight about the phenomenology of understanding, to establish a credible epistemology of understanding and metaphysics of meaning that can accommodate the publicity thesis. He looks for:

> [a] construal of the thesis that meaning can be fully overt in linguistic behavior: a construal according to which whenever someone who is competent in a language speaks, so long as he speaks correctly, audibly, and so forth, he makes knowledge of his meaning available—to an audience who understands the language he is speaking. (McDowell 1998a: 352–3)

According to McDowell, the two options first considered are wrong because they force us to choose between Behaviorism and Cartesianism about the mind. The mental is either reduced to patterns in behavior or retreats beneath the surface of behavior into an utterly private realm. But why should we take these to be the only options? We want our understanding of people's speech to engage with their inner lives, but at the same time we want what they say to be outwardly revealed to us. The dilemma we find ourselves in, according to McDowell, of being pulled in one direction or the other, is due to their sharing of a common assumption, which must be discharged to avoid the horns of the dilemma.[2] The common assumption is the thought that all people can outwardly present to us when they speak is a sequence of sounds and gestures; mere bits of behavior described in meaning-free terms. From there we seem forced to choose between finding meaning in meaning-free behavior—reducing meaning to patterns in otherwise uninterpreted verbal behavior—or to finding meaning preserved in the mind of the speaker, hidden from view as part of a private, inaccessible realm of inner items—what Quine called the museum myth. The way to discharge this assumption is to come to see that there is no sharp divide between inner and outer, between the intentional mind and outward behavior. By ceasing to dichotomize the mind and the body in terms of inner and outer, we leave room to find the mind fully exhibited in behavior rather than hidden behind it, screened off by uninterpreted sounds and movements. Just as the involvement of the mind in intentional actions goes right to the ends of our finger tips, so it reaches right out into the sounds we publicly articulate. The mind's involvement in action, linguistic and otherwise, does not stop short of the full outward display of intentional agency. That is why, from the perspective of the listener, 'the

[2] The dialectic here should be familiar to readers of McDowell, who would usually recognize these two unacceptable options as Scylla and Charybdis.

understanding of a language … consists in awareness of … unproblematically detectable facts' (McDowell 1998a: 331). '[T]he significance of utterances in a language must, in general, lie open to view, in publicly available facts about linguistic behaviour in its circumstances' (314). Otherwise understanding would consist in 'hypotheses about inner states of the speaker lying behind the behaviour' (331).

In many respects, McDowell is more Quinean than Cartesian. He regards Quine's commitment to the publicity of meaning as wholly admirable. Where Quine goes wrong, in McDowell's view, is in insisting that publicly observable behavior be characterized in meaning-free terms—in particular, in the terms the natural scientist would recognize. There is no need to insist on such limitations. Moreover, were we to adhere to such scientific—or, as McDowell and his followers like to say, scientistic—scruples, there would be no way to capture what takes place in linguistic behavior, no correct characterization of what we hear in one another's speech. We need to recognize the deliberate and purposeful behavior speakers give rise to for what it is—the intentional production of meaningful speech—and there is no way of doing so save by presupposing the meanings of the words whose use by a speaker we are describing. We must appeal to the significance these bits of behavior have in the linguistic practices of the speech community we belong to when reporting what members of that community are up to:

> [S]hared membership in a linguistic community is not just a matter of matching in aspects of an exterior that we present to anyone whatsoever, but equips us to make our minds available to one another by confronting one another with a different exterior from that which we present to outsiders. (McDowell 1998b: 253)

The use of language cannot be rendered faithfully without presupposing the language in question in our description of that use. It is shared possession of a language that makes it possible for us to reveal the contents of our minds to one another on the surface of our speech behavior: '[a] linguistic community is conceived as bound together, not by a match in mere externals (facts available to just anyone) but by a capacity for a meeting of minds' (McDowell 1998b: 253). Shared command of a language equips us to know one another's meaning without needing to arrive at that knowledge by interpretation, because it equips us to hear someone else's meaning in his words.

Quine was wrong to think the surface of speech could be described in meaning-free, physicalistic terms. Such materials cannot support descriptions of what people are up to in acts of uttering the meaningful words and sentences we immediately take them to be uttering. According to McDowell's picture, meaning is no longer 'conceived as behind the surface of linguistic

behavior but as residing on its surface' when that surface is located properly and not characterized in the shallow way Quine insists upon. The overt surface we display to one another can only be recognized when it is seen as activity characterized in normative and meaningful terms. When we encounter speech sounds made by members of our linguistic community, the meanings we hear in their words lie open to view on the surface of their practice: '... the outward aspect of linguistic behaviour is essentially content-involving, so that the mind's role in speech is, as it were, on the surface' (McDowell 1998a: 100).

What the phenomenological datum about hearing meaning in people's words is now meant to show us is how utterly misguided it would be, as part of the epistemology of understanding, to suppose that our minds engaged with a surface comprising anything less than meaningful speech. To recognize speech for what it is, its surface must be characterized richly in terms that show how meanings can be fully available to us in the experience of listening to one another: 'the senses of utterances are not hidden behind them, but lie open to view' (McDowell 1998a: 99).[3]

In what follows, I will point to overwhelming evidence that the rich texture of our linguistic experience in listening to speech cannot be found on the surface of that speech, but at this stage I am concerned with McDowell's reasons for thinking that it can.

Notice that, for McDowell, the richness of that surface, and what it makes available to us, is not available to just anyone. The outward aspect that matters can only be presented to those who understand the language: 'one hears more, in speech in a language, when one has learned the language' (McDowell 1998a: 333). Few, if any, of the linguistic features of that surface will be detectable by outsiders, as we can appreciate when listening to a foreign language. Whether one hears these sounds as meaningful or merely as noise, depends, we are told, on whether one possesses knowledge of the language. But it is the nature of the dependence of what we can hear on our knowledge of language that we need to be told more about. How does linguistic knowledge make the outward and significant aspect of speech available to us? How does it enable us to hear what is there on the surface? On these issues McDowell has little to offer,

[3] It may seem as if we could never be mistaken if we accept McDowell's view, but that is not his claim. We are fallible in our epistemology of understanding because we may think we are hearing meaningful speech, enjoying a genuine meeting of minds, hearing the meanings that are there of the surface, and yet be subject to an illusion or auditory hallucination. Nonetheless, either we are directly perceiving real speech or just getting counterfeit coin. What doesn't explain our fallibility is the idea that we are always engaging with something less than meaningful speech or less than full evidence of a mind on show, in linguistic behaviour from which we make at best risky inferences about what goes on at the real locus of mind and meaning.

and he is even less forthcoming on how we come to acquire the linguistic knowledge that gives us this capacity. He tells us the difficulty lies in having to answer the question: 'How can drilling in a behavioural repertoire [effecting a change in one's external behavior] stretch one's perceptual capacities—cause one to be directly aware of facts of which one would otherwise not have been aware?' (McDowell 1998a: 333).

One's natural inclination is to say that it can't, and that there is simply no answer to this question. It is the wrong question. Any plausible account of how we acquire the capacity to experience (certain) speech sounds as meaningful must begin elsewhere. We need to look at what linguistics and psychology tell us about the acquisition of language and the perception of speech.

2. Fodor versus McDowell on the Epistemology of Understanding

First of all, the phenomenological datum that we hear more in the speech sounds of a language we understand cannot by itself support McDowell's conclusions about the metaphysics and epistemology of speech. Further argument is needed. For the very same phenomenological insights are produced by Jerry Fodor to support the claims that speech perception must be the result of unconscious and automatic modular processes: 'You can't help hearing an utterance of a sentence (in a language you know) as an utterance of a sentence… You can't hear speech as noise *even if you would prefer to*' (Fodor 1983: 52–3). Thus: '"I couldn't help hearing what you said" is one of those clichés which, often enough, expresses a literal truth; and it is what is *said* that one can't help hearing, not just what is *uttered*' (55; emphasis in original). For Fodor: '… understanding an utterance involves establishing its analysis at several different levels of representation: phonetic, phonological, lexical, syntactic, and so forth' (64). This is the work of fast, dedicated, and mandatory cognitive processes that perform inference-like computations on their domain-specific representations. Here, Fodor cites William Marslen-Wilson and Lorraine Tyler's work on word-recognition, who tell us that:

> … even when subjects are asked to focus their attention on the acoustic-phonetic properties of the input, they do not seem able to avoid identifying the words involved… This implies that the kind of processing operations observable in spoken-word recognition are mediated by automatic processes which are obligatorily applied… (Marslen-Wilson and Tyler 1981: 327; as quoted by Fodor 1983: 53)

The automatic and obligatory character of such processes, in contrast to the voluntary and reflective process of conscious deliberation, is the hallmark of the sub-personal. And this is the picture Fodor offers us of our response to speech sounds in a familiar language. Language comprehension is accomplished by means of an input module: an informationally encapsulated cognitive mechanism that responds selectively to certain informational inputs, and delivers its outputs to the central (thought) processes. The fast and automatic way in which we hear what is said, rather than merely appreciating the acoustic properties of the sounds, is evidence, for Fodor, of the workings of a sub-personal linguistic system whose products are delivered to consciousness but whose workings are cognitively inscrutable.[4] Thus, the phenomenological datum about what is consciously accessible when listening to speech in a language one understands settles nothing about the locus of linguistic significance, nor whether linguistic comprehension is a matter of direct perception or unconscious inference.

The experience of speech sounds is richer for those who understand the language than for those who don't. But it is also poorer, in that there is good evidence that the auditory processing of environmental noise and speech sounds may proceed in parallel; the result of experiencing speech sounds may, however, inhibit the auditory processing of non-speech sounds. To hear speech is not to listen to the sounds. But it is, as I shall argue, to listen to their sources: the voices of those who are producing them.

3. Phenomenology as Epistemology: Taking Experience at Face Value

What further considerations can McDowell offer for taking his observation about perceived speech not just as a phenomenological but as an epistemological claim? There are, I think, three considerations. First, we are invited to take the experience at face value, as our listening to the meanings that are present in people's words, for unless we view our experience of speech in this way—and assuming that we accept there is determinate meaning to what people say—there would be no knowing for sure what someone else meant

[4] Fodor stresses the way in which information about the acoustic properties of speech is lost when comprehension takes place. We know how Swedish and Chinese sounds, but do we know how English sounds? We fail to notice the absence of certain acoustic properties, as when phonemes have been spliced out of the middle of a recording of someone uttering a word, and yet those listening still hear the whole word (the so-called phoneme restoration effect).

in uttering a sentence. Secondly, there is no *encounter* with anything less than meaningful speech from which to construct a meaning for the sounds we hear people utter. Thus, hearing speech in a language we understand is a matter of direct confrontation with the meaningful surface of other people's linguistic behavior. Thirdly, there is no way to recognize the activity of speech for what it is without presupposing a command of the meanings of the words in the language spoken.

There is something to the second and third of McDowell's points, though they both need careful qualification. It is true that *we* do not encounter anything less than meaningful speech from which we assemble the meaning of people's utterances, but that is because our *speech processors* make contact with features at the *sub-personal* level. However, McDowell provides an argument in support of his second point to the effect that even if we could slow down the processing and examine what goes on, as we can in the case of our fast recognition of written words on the basis of the letters that compose them, there would be nothing corresponding to letters and rules from which we could assemble word meanings that could provide an explicit grounding for our understanding of them. I agree with McDowell that even if we do not go through explicit reasoning, and we suppose that meaning recognition occurs quickly and sub-personally, or as McDowell would say by means of a 'cognitive short-cut', there is no set of meaning-free items from which to assemble lexical meanings.[5] Like McDowell, I think we recognize word meanings as a whole and that this occurs at the conscious personal level and not in our sub-personal linguistic systems. But to say this is not to say that such word meanings are to be found in words present on the surface of someone's speech.

McDowell may also be right to say that we must presuppose the hearer's knowledge of the meaning of words in order to credit him with the capacity to recognize other people's linguistic behavior as meaningful speech. But, once again, there is no reason to suppose that the word meanings he has knowledge of are located externally in outward aspects of speech behavior, parts of a publicly shared language. I shall argue that meanings reside in the minds of speakers and hearers, and that the meanings we hear in people's words are the meanings we take their words to have. The words uttered, when recognized, are heard with the meanings they have for the hearer.[6] Thus, I have to reject

[5] See McDowell's (1998a: 117–18) argument by analogy with reading letters. In fact, the empirical evidence suggests a dual-route model of reading that involves both whole-word recognition and letter-by-letter spelling out. The data from language pathologies shows evidence of double dissociation, where patients can lose one capacity while retaining the other.

[6] Note that McDowell (1998a: 282), too, claims that 'command of a meaning is wholly a matter of how it is with someone's mind' and 'that a speaker means what she does … must be constituted at least

McDowell's first point, and show there is another way to secure knowledge of what someone else is saying that does not presuppose taking our experience of perceived speech at face value as an encounter with the external surface of linguistic behavior.

In seeking another route by which to secure knowledge of what someone else is saying, I am not simply engaged in a philosophical exercise of looking for alternative accounts of the epistemology of understanding. Nor is my alternative account solely motivated by qualms about the metaphysical extravagance of McDowell's picture, according to which meanings lie on the surface of speech episodes.[7] Rather, the motivation derives in part from the existence of conclusive empirical arguments against the possibility of locating linguistic properties in the sounds speakers produce. A credible philosophical picture of how we understand one another's meanings that is compatible with the best findings of the linguistic and speech sciences is better than one that is not. Thus, despite the subtlety, attractiveness, and phenomenological acuity of McDowell's view, there are many reasons to think it is wrong about the epistemology of understanding and wholly mistaken about the locus of linguistic significance. What is more, there are further phenomenological grounds for thinking that he overlooks the real basis for a meeting of minds. I will briefly state the empirical findings that put pressure on McDowell's view and offer further philosophical considerations against the account based on our phenomenological experience of sound. Let us look now at the empirical arguments.

4. How do we Come to Hear Words in the Sounds People Utter?

In hearing speech sounds, we are presented with a continuous sound stream with no gaps indicating the boundaries between words that we find in written language. If there were gaps between words in human speech, it would sound unnatural and hard to follow. Yet the fact that we confront a largely

in part by her physical and social environment'. But this is not just an externalist thesis about meanings: 'command of a word's meaning is a mental capacity … the mind [is] the locus of our manipulations of meanings … Meanings are in the mind but as [Putnam's] argument establishes, they cannot be in the head; therefore, we ought to conclude, the mind is not in the head' (276). Recognizing another's meanings is recognizing a bit of their mind, and because meanings as parts of the mind are literally out there, as part of the external environment we encounter, so too are these parts of their mind.

[7] McDowell tries to lessen our qualms about this rather magical sprinkling of meanings on the exterior surfaces of things in the world by describing such a world as 'enchanted'.

uninterrupted acoustic signal is not easy to reconcile with what we take ourselves to be hearing when listening to others speak. What we 'hear' is the articulation of discrete words and syllables that do not, strictly speaking, occur in the acoustic speech signal. In fact, much of what we supposedly 'hear' in the speech signal makes no public appearance at all. Word boundaries, non-overlapping syllables, restored phonemes; none of these items is present in the speech signal, and yet all are perceived as being there. Somehow the mind imposes such items on the sound stream presented to us. So, what are we listening to when we hear another speak, and how does auditory perception give us knowledge of it? I will claim that what we are listening to is the voice of the person talking. What we 'hear' that person as saying depends on processes that go beyond the information given.

Speech perception depends on a 'set of processes by which the listener extracts words from the continuous, rapidly changing, acoustic signal of speech'. In recognizing the sentence uttered from the continuous signal and 'the multidimensional properties of [the] acoustic stimulus, we have to analyze the frequency spectrum, identify phonetic features, segment phonological units', as well as initiate word recognition and deploy syntactic information. And we do all of this at lightning speed. On average, 'we perceive and produce about three words per second or one phone every tenth of a second' (Trout 2001, 2003). The difficulty of the task cannot be overestimated, 'because the acoustic realizations of a given word can vary greatly depending on speech rate, speaker's voice features, context, etc.' (Dehaene-Lambertz *et al.* 2005: 21). And yet, 'Despite their apparent variability, words, and the phonemes that constitute them, are … most often effortlessly identified' (ibid.).

Just how are perceptual constancy and categorical perception effects achieved when attending to 'a continuous, rapidly changing acoustic speech signal'? Is it by means of 'general auditory mechanisms or special speech decoding processes'? Are properties of phoneme perception essentially dependent on physiological properties of the auditory system, psychoacoustic mechanisms, or are they the upshot of domain-specific speech processors? The overwhelming evidence finds in favor of specialized speech processing mechanisms rather than just general auditory mechanisms.

In an experiment by Dehaene-Lambertz *et al.* (2005), subjects are presented with computer-generated sounds akin to speech sounds, which, after a while, subjects suddenly come to hear as syllables:

> Many people exposed to sine wave analogues of speech first report hearing them as electronic glissando and, later, when they switch into a 'speech mode', hearing

them as syllables. This perceptual switch modifies their discrimination abilities, enhancing perception of differences that cross phonemic boundaries while diminishing perception of differences within phonemic categories. (Dehaene-Lambertz *et al.* 2005: 21)

Different cortical regions are activated depending on whether the perceiver is in speech or non-speech mode. Event-related potential (ERP) and functional magnetic resonance imaging (fMRI) studies show that switching to the speech mode significantly enhanced activation of certain brain areas (left superior gyrus and sulcus) and were 'activated significantly more by a phonemic change than by an acoustic change' (with the same acoustic stimuli). These results and many more like them serve to 'demonstrate that phoneme perception in adults relies on a specific and highly efficient left-hemisphere network which can be activated in a top-down fashion when processing ambiguous speech/non-speech stimuli' (Dehaene-Lambertz *et al.* 2005: 21).[8] Such dedicated and, most likely, innate, speech-processing mechanisms make a significant contribution to the perception of speech. They are responsible for phoneme constancy, for our perceiving word and syllable boundaries, and much else. Clearly, such mechanisms go beyond the information given in their inputs. Not only is some information from the acoustic signal discarded in the process of chunking, ordering, and reducing the amount of auditory information we are exposed to, but crucial information can be added, as is shown in the phoneme restoration effect (Warren 1970). Phonemic representation is computed faster and more efficiently than corresponding acoustic representation of the *same* stimulus. The phonemic network, once activated by our speech processors, can have an inhibitory effect on the concurrent auditory representations to prevent interference from non-linguistically pertinent differences (Liberman *et al.* 1981).

So, for those who understand the language, the experience of speech sounds is richer than the experience of those who do not. But, in certain environmental ways, it is poorer, too. For although the auditory processing of an acoustic signal as sound and as speech may proceed in parallel, the result of experiencing sounds as speech may inhibit the auditory processing of non-speech sounds. This may be why when we listen to speech in a familiar language we do not listen to the sounds; but, as I shall argue below, we do listen to the source of the sounds: the voice of the person who is producing them.

Not only do speech-processing mechanisms have an effect on speech perception, but visual information from faces can also affect the auditory

[8] There are many other empirical arguments in favor of a specialized speech processor. See Dehaene-Lambertz *et al.* (2005) for review and references.

perception of phonemes. The powerful McGurk effect occurs when subjects listening to the sound /ba/ while seeing on film a face making the lip movements for /ga/ hear the sound /da/. What they 'perceive' is a blend of the audio and visual information (MacDonald and McGurk 1978). In addition, there is neuroimaging evidence of the interaction of cortical areas for voice and face recognition when listening to a familiar speaker (von Kriegstein *et al.* 2005). What normally sighted listeners take themselves to hear may always be an amalgam of information from different sources.

At the level of phonemes, the evidence that we simply pick up linguistic information from the environment is scanty. First, we do not detect discrete units like phonemes in the acoustic speech signal, so there is considerable rupture between features of the acoustic signal and what we perceive as being uttered. And yet, without the direct perception of phonemes, there cannot be direct perception of words made up from those phonemes, let alone the perception of word boundaries. The phenomenological experience we have when we hear speech cannot easily be reconciled with the empirical findings. The same perceived phonemes correspond to quite different acoustic properties, and the same acoustic properties correspond to differently perceived phonemes. And where a phoneme is deleted in the middle of a word and replaced by a cough, one will report hearing an utterance of the whole word, including the missing phoneme, with a cough in the background (Warren 1970). What all this shows is that perceived speech sounds do not correspond to actual surface features of the speaker's acoustic signal. Now it may be objected that this is because we are trying to locate the surface of speech in the wrong place; in what Wittgensteinians like McDowell call 'sub-bedrock' terms.[9] But what the empirical findings really show is the extreme difficulty of locating what we experience when we perceive speech sounds at the phonemic level in anything that could properly be regarded as a surface in any sense. And yet an alignment between what is perceived and what occurs out there on the exterior of speech is precisely what the direct realist account requires. Surely, it is more plausible that the supposed surface of speech is in fact a percept of hearers due to the information they bring to bear in the course of processing the auditory information they are

[9] When describing phenomena like speech and other human actions, Wittgenstein reminds us to recognize the ground that lies before us as the ground, and warns us not to dig below bedrock. The point is that below bedrock, justifications give out, and there is nothing to support the attribution of normative notions like meaning and intention at the level above: the level Wittgenstein is calling 'bedrock'. See Wittgenstein (1983: VI, 31) and McDowell's (1998b: 249–54) gloss on this.

given. We hear speech *as* the articulation of distinct phonemes making up words that contribute to a whole sentence. But to hear a sequence of sounds *as* the utterance of meaningful words and sentences is not the same as saying we hear the words *in* the sounds or hear meanings in words present on the surface of speech. The conception of a surface as McDowell describes it is more plausibly construed in terms of the phenomenological experience of hearers than as anything lying on the exterior surface of the speaker's behavior.

The moral is that 'linguistic information is projected by means of articulations but is not embodied in them'. Linguistic information is 'read into' rather than 'read off' these sounds. It is part of our 'specifically human way with sounds' to do so (Harris and Lindsay 2000: 203). The speaker may take himself to be going public on what he is thinking and see himself as putting his meanings right out there in his words, but however things strike him phenomenologically, he cannot succeed in putting more into the sounds he emits than they can actually bear. And in many cases, he simply cannot make the crucial linguistic properties appear publicly at all. All that is out there are sounds and marks. And it is language users like us, with the cognitive systems we have, that can make something of these linguistically caused items. From the perspective of the speaker experiencing himself as producing a rich string of meaningful words and phrases, he is like a person tapping out, or whistling, a tune for others to recognize. All he puts out there are some impoverished noises, but he hears the sounds he produces with the rich inner accompaniments that make what he is tapping to or whistling seem so obvious. To the listener trying to recognize the tune, it may sound like mere tapping or noise, much as speech sounds sound like noise to those listening to a foreign language. Those who know the language and identify words in the sounds uttered will hear the sounds with the meanings they have for them as a matter of their inner experience.

A correct view of language and our knowledge of language needs to account for our capacity to hear complex meaning in speech sounds and to produce sounds imbued with meanings in indefinitely many cases; it will have to explain our immediate readiness to produce and comprehend utterances of sentences we have never used or heard before. And it will have to explain how by these means we succeed in making our minds available to one another. However, the way knowledge of the language helps us to perceive more in the speech sounds of a familiar language is not by giving us an ability to directly perceive 'unproblematically available facts' that lie on the surface of speech. It is by bringing our linguistic knowledge to bear on auditory inputs in order to

recover *more* than is given in the sound waves themselves. Much of this will be done by automatic and unconscious processes, though, as I acknowledged above, this is not where word meaning is to be found.

In effect, we have to compare two conceptions of language and knowledge of language:

(A) Speakers' knowledge latches onto properties of an external language.
(B) Speakers' knowledge determines the properties of their internally represented language.

The (A) conception of language, popular with philosophers of language and the folk, supposes that our competence is based on our acquiring knowledge of the observable facts of a public language. These facts are often supposed to be matters of convention we are taught and gradually adopt. The (B) conception of language, widely held by generative linguists, is that there is a largely innate basis for language in the brains of human language users, where language is now understood as the internal mechanism that enables us to speak and understand. According to this conception, many linguistic properties are due to the organization of our language faculties. Thus, the correct grammatical generalizations about our language—the ones we actually conform to—are neither consciously arrived at nor conventional regularities. They are the upshot of the workings of an internalized grammar.[10] Speakers have an innate capacity for language because of their native endowment with a universal grammar and their initial exposure to linguistic data: data that do not fully determine the language or (I-language) acquired, as the poverty of stimulus arguments tell us. The linguistic structures we deal with are internally generated in the mind of the speaker and assigned to sounds and marks which otherwise carry no linguistic information.

By contrast, McDowell thinks that what we perceive in speech, by virtue of having learned a language, is something lying open to view—the *surface* of linguistic practice. These are linguistic phenomena already there that we come to perceive as a result of acquiring knowledge of the language: a range of facts that were not previously (directly) perceptible come into view as we 'find our way into' the language.

[10] To appreciate how radically different the linguist's notion of a grammar and language are from the traditional folk conception embraced by many philosophers, consider these remarks by Chomsky: '... what should we take as a language ...? The natural choice is g, the generative procedure; thus a person who knows language L has a specified method for interpreting arbitrary expressions, such as [''Who do you think that John saw'' and ''What do you wonder who saw'', ''Who do you wonder what saw'', ''He likes John'', ''His mother likes John'', ''John likes him'', ''John's mother likes him'', J (a sentence of Japanese)]. Let us call g_E the I-language that some particular speaker of English (Jones) has acquired' (Chomsky 1987: 181).

5. Syntax and the Surface of Speech Behavior

McDowell's picture of the meaningful surface of speech faces even greater difficulties when we look at the syntactic structure of sentences. The semantic interpretations we can give to word strings depend on what syntactic analysis they are given. What we hear an uttered sentence as meaning depends systematically on its linguistic form. An ambiguous word string can be heard first one way and then another, depending on how we perceive it as structured:

(1) He talked to the woman from the sailing boat.

Connections between linguistic form and meaning are what compositional theories of meaning set out to describe:

> If (but only if) speakers of the language can understand certain sentences they have not previously encountered, as a result of acquaintance with their parts, the semanticist must state how the meaning of these sentences is a function of the meanings of those parts. (Evans 1975: 344)

And they can do so only if they identify the semantically relevant structural constituents of a sentence. This depends on the internal syntactic organization of the sentence. Syntactic configurations constrain the interpretations that can be given to word strings. In the following examples, there are certain interpretations they cannot have and others they must have:

(2) I know Mary expected to feed herself.
(3) I wonder who Mary expected to feed herself.

In (2), 'Mary' and 'herself' can only be construed as referring to the same person, while in (3) 'Mary' and 'herself' cannot be so construed despite the same sequence of words appearing in both (2) and (3). Speakers know these facts but they do not know how they know them. In particular they do not know that 'who' is construed as referentially dependent on a phonetically empty category in the syntax, PRO, that serves as the arbitrary subject of 'to feed herself'. What language users hear these sentences as meaning—and what they cannot hear them as meaning—are systematically correlated with facts about the syntactic configuration of these strings. But where should we locate these syntactic facts and why do the interpretations speakers and hearers give to particular word strings conform to linguistic generalizations defined over such facts? The linguistic generalizations speakers conform to cannot be captured in terms of surface properties of these strings—assuming for the sake of argument we could unproblematically recognize words in the surface sound string. They

consist, rather, in facts about hierarchical relations among constituents of sentences, only some of which appear in the surface string. It is well known in linguistic theory that we cannot describe the syntactic structure of sentences by reference to linear arrangements of word strings, and that we must posit levels of syntactic structure remote from surface form.[11] The question for us is how does syntactic information impact what we are able to hear in listening to speech in a familiar language?

McDowell recognizes the importance of systematicity and states the requirement of *system* in a theory of meaning as follows: 'We want to see the content we attribute to foreign sayings as determined by the contribution of distinguishable parts or aspects of foreign utterances, each of which may occur, making the same contribution, in a multiplicity of utterances' (1998a: 145).

This will be achieved by constructing a truth theory for the language, whose axioms deal with the primitive expressions of the language and feature as premises in the derivation of T-theorems that deal with sentences in which those expressions occur. Does this requirement only apply to theories constructed to interpret foreign languages? After all: 'Comprehension of speech in a familiar language is a matter of unreflective perception, not bringing a theory to bear' (McDowell 1998a: 179).

However, theory has a role in the home case, too, as it provides a means of describing the range of our capacity to perceive speech in a familiar language: 'The ability to comprehend heard speech is an information-processing capacity, and the theory would describe it by articulating in detail the relation, which defines the capacity, between input information and output information' (179). The range of facts about sentences, as we see deduced in the theory's output theorems, amounts to a description of the extent of the speaker's capacity.

> For theorems to be so deducible, utterances must be identifiable in terms of structures and constituents assigned to them by a systematic syntax; and it must be possible to match up those structures (if necessarily obliquely, through transformations) with configurations observable in physical utterance-events. (McDowell 1998a: 145)

More surprisingly, we hear:

> The hard physical facts, then, that constrain the construction of a truth-characterization for a language actually spoken are (i) the structural properties of physical

[11] For more on this point and on the significance of it for the folk view of language, see Smith (2006a).

utterance-events that permit the language to be given a syntactic description; and (ii) the complex relations between behaviour and the environment that permit (some of) the behaviour to be described and understood in intentional terms. (1998a: 146)

It is at the level of theorems dealing with the content of whole sentences that the truth-theory 'makes contact with the hard physical facts'. If the theorems are to be deducible in the systematic ways described, they 'must characterize utterances in terms of structures and constituents; so that the relation of match … must hold between the structures assigned to sentences by the syntax with which the theory operates … and configurations observable in the physical utterance-events'. So, the requirement of systematicity 'makes itself felt … in connection with the match between theoretical syntax and actual utterance-events' (McDowell 1998a: 146). But the talk of a match between syntactic structure and the hard physical facts about actual utterance-events is even less empirically plausible than the identification of phonemes and syllables in the acoustic speech signal.

The syntactic structures that feature in the linguistic generalizations speakers conform to are not consciously recognized or manipulated by speakers in producing and hearing meaningful utterances; but in order to conform to such generalizations, speakers must be able somehow to register the relevant facts about the underlying syntactic structure of a sentence. Thus, it is much more plausible to suppose that 'the language-input system specifies, for any utterance in its domain, its linguistic and maybe its logical form', and that 'the language processor delivers, for each input utterance, a representation which specifies its lexical constituents *inter alia*' (Fodor 1983: 90–1). The resultant understanding of the uttered sounds depends on speech processing in which a syntactic analysis is provided for the lexical items recovered from the input. Not all of the properties of sentences' linguistic or logical form are phenomenologically accessible, but what is accessible—our hearing a sentence as structured—depends crucially on what syntactic structure our fast, automatic, and unconscious speech processors assigns to the input string. Once again, the linguistic properties we rely upon in understanding the speech of others are not properties we find on the surface of speech. The syntactic constituents, their categories, and syntactic dependence—along with phonetically null, empty categories like traces and PRO—are simply not found in the sound string. The case for locating the meaning-determining properties of syntax in the sounds we encounter is without empirical foundation. Instead, what we see at work in McDowell's picture, as in the case of phonemes, is the myth of the externally given nature of language.

At this point, McDowell and others with an exteriorized conception of language could suppose there was a dichotomy between the phonological and syntactic properties of language that belong to the language faculty, and the publicly accessible properties of word and sentence meaning that are consciously accessible. While the former cannot be located on the surface of speech, perhaps the latter reside in the sounds speakers make. In effect, this would be to deny that there was a single locus of linguistic significance, and to suppose instead that the meaning properties of words were public and social, while the phonological and syntactic properties were part of our internal cognitive psychology. This hybrid picture would preserve the idea of word meanings occurring at the personal level, while most of the other linguistic properties were represented sub-personally in the language faculty.

How plausible is the hybrid picture? Prima facie, it faces severe difficulties in describing how the *meanings* of words and sentences come to be directly related to sounds, since the words and grammar they depend on cannot be located auditorily. A supporter of this picture would have to show how the properties that reside at these different levels and locations either interact or could be aligned so as to respect linguistic generalizations. I doubt whether this could be done, but I think that such a position faces greater difficulties still, and that the very notion of a surface to speech becomes problematic when we reflect on the nature of our auditory experience of sounds.

6. Sounds and the Phenomenological Experience of Speech

It is crucial to McDowell's picture that the unproblematic meaning facts we supposedly perceive in others' speech reside on the surface, or outward aspects, of linguistic behavior. According to this view, we are able to know what people say only because we perceive meaning in the sounds they utter. The idea of locating the surface of speech where linguistic meaning is to be found makes sense only if we can locate the speech sounds in which meaning is meant to reside. But can we? I shall now argue that there is no surface to speech because our auditory perception of speech fails to *locate* the sounds speakers produce.

Sounds, in general, are hard to place in the spatial world and auditory perception gives us no clues as to where they might occur. When we reflect on the metaphysics of sounds, there appear to be only three candidate

locations. Sounds are (i) at their sources, (ii) with us when we hear them, or (iii) somewhere in between. None of these options is satisfactory.[12] Sounds cannot be where their sources are, since the source of a loud explosion may be hundreds of miles away when we hear it. On the other hand, to treat sounds as occurring where I am when I hear them is to suppose that different hearers cannot literally hear the same sound. A bell is struck and it chimes in many minds at once. Surely they hear the same sound? On the second view, this could not be true. The final view usually treats sounds as identical with the sound waves that travel from the source to my ears, which would require sounds to travel, to get nearer and nearer to us as hearers. But we simply do not hear sounds themselves as moving. Of course, we can hear the source of the sound as getting nearer or farther from us, but sounds themselves are not heard as traveling towards us.

So what should we say? The origin of the view of sounds and sound perception I want to endorse is found in the work of Brian O'Shaughnessy (Chapter 6; 2000) and is developed further by Matthew Nudds in this volume (Chapter 4). O'Shaughnessy begins with a version of the second thesis, saying that 'the sound that I hear is *where I am* when I hear it'. Whether or not this is the correct account of sounds, the reasons he gives for this view offer an important insight into the non-special nature of our perception of sounds:

> [H]earing the sound to be coming from point p is not a case of hearing it to be *at p*. This is because the sound that I hear is *where I am* when I hear it. Yet this latter fact is liable to elude us because, while we have the auditory experience of hearing that a sound *comes from p*, we do not have any experience that it is here where it now sounds. (Rather, we work that one out.) And this is so for a very interesting reason: namely, that we absolutely never immediately perceive sounds to be *at* any place. (O'Shaughnessy 2000: 446; emphasis in original)

Although we do not hear sounds as located, we do hear their sources as located. In developing this view, Nudds (Chapter 4) points out that sounds need not exhaust the immediate objects of our auditory attention. We commonly listen to their sources, not to the sounds themselves. We seldom pay attention to the sensory qualities of sounds, but we focus instead on what is producing them. We mostly perceive sounds in terms of the objects that produced them. We hear the sound *of* a violin, the sound *of* a dog barking, the sound *of* the logs on the fire, the sound *of* the gas being lit. We hear and are interested in these distal sources of the sounds, and we hear them because the experience of a sound represents its source and the properties of its source. Through the processing

[12] For stout defenses of the first and second positions, see Casati and Dokic (Chapter 5) and O'Callaghan (Chapter 2).

of sound waves we are able to tell quite a lot about the size, movement, and density of the objects producing the sounds. It is these properties of the things producing the sounds—and not the sounds themselves—that we are interested in. Nudds (Chapter 4) stresses that 'we cannot explain why the auditory system groups the frequency components that it detects in the way it does, other than in terms of a process that functions to extract information about the objects [sources] that produced those frequency components' (p. 74), and thus, 'we can only explain why we experience the sounds we do in terms of a process that functions to tell us about the sources of sounds' (p. 75). 'Auditory perception tells us about the sources of sounds' (p. 72).

In hearing sounds, we listen to what (or who) is producing those sounds. And in the case of speech, we listen primarily to a voice. Voices are the sources of speech sounds, and voices are special. A voice belongs to a person, an embodied subject who intentionally produces the sounds we hear. We can recognize a lot about the producer of those sounds from properties of the voice, and we succeed in recognizing people's voices after only a brief exposure to their speech. On the radio, we identify a person speaking from his or her voice. Recognizing a voice is in normal circumstances recognizing who is speaking. In normal conditions, a voice provides a unique sensory print of a person.[13] Very specific information about an individual is conveyed by the voice. The identity of an individual is recognized by voice quality—recognition of a voice is usually recognition of a person. Even emotional states are largely recognized by non-semantic properties of speech, as demonstrated by several experiments that show vocal expression of emotions as being reliably recognized in content-masked speech signals (Fukuda and Kostov 1999; Scherer *et al.* 1972). Voice typically conveys information about the size, age, and gender of the speaker. When we do pay attention to the sound of someone's voice, it is because it can tell us something about the source of the sounds: the person himself and the state of mind he or she is in. We may attend to the tone of his voice or to its loudness, and we may hear its tremor or its catch. The auditory system detects these slight variations in voice quality—even in the sounds of unfamiliar voices—and registers them as signs of the nervousness or irritation in the speaker. We listen to such sounds when they tell us something specific about a person's mind. Perception of a voice as the source of speech sounds connects us immediately and intimately to the mind of another. There is a unique and direct meeting of minds, and all of this happens without semantic understanding. We sometimes hear the sound of voices talking in another room without hearing what is being said. But,

[13] I owe the idea of putting things this way to Anne-Lise Giraud.

again, the import of the experience is not just that there are sounds I am experiencing. I am hearing the voices of *people* talking. Through hearing a voice, we hear ourselves being addressed by a person, and, if the experience is veridical, we are hearing the mind of an individual who is addressing us. None of this information is conveyed via the content of what the person is saying. We do typically hear certain speech sounds *as* meaningful. But do we literally hear the meaning *in* the sounds? How can we say this on the basis of perceptual experience? The experience of speech sounds locates their sources (or apparent sources), but not the sounds.[14] We treat the source—a voice and therefore a subject—as the originator of the meanings we take the uttered words to have, but we do not perceive the meanings to be anywhere.

We do not have to claim that speech sounds do not have a location, though perhaps they do not. Rather, we only need the phenomenological observation that our experience of speech sounds fails to locate them. After all, where do you hear the sounds of someone's speech to be occurring? Look at a speaker's mouth moving and note what you hear the speaker saying. *Where* in this sequence do you observe the speech sounds to be occurring? It is impossible on the basis of our phenomenological experience of speech to give any location to the speech sounds in which meanings are meant to reside, we cannot give any external location to the surface of speech. The location is neither on the speaker's lips, where I am looking when I hear someone speak, nor in the air between us. (I hear the sounds as coming from that person—in other words, my experience locates the source of the sounds.) Without a place for speech sounds to be, there is no exterior surface on which to locate the meanings of words. Auditory speech perception simply gives us experience of sounds that presents voices and properties of those voices. One hears this as a result of auditory speech processing involving the segmenting and grouping of sound waves, with perhaps some knowledge of the properties of the human voice attended to. We do not experience the meaning of words as lying 'open to view, in publicly available facts about linguistic behavior in its circumstances', or as occurring anywhere. The phenomenological experience is of listening to a voice, the voice of a person. And what we take the words we recognize being uttered to mean is what we take the person who is voicing their thoughts to mean.

[14] We need to say 'apparent sources', because of the ventriloquism effect made use of in cinema, where an unlocalized source of a voice sound comes to be identified with the location of a visual cue of a mouth making speech sounds. Attentional capture of the auditory system by visual cues happens only so long as the lip synchronization with the sounds is close enough, otherwise the illusion breaks down. I am grateful to Charles Spence for this point.

The conviction we have that we are in touch with someone when talking to them, that there is no barrier between minds, is largely due to features of face and voice we recognize independently of understanding the content of their speech. The recognition of a voice, is, normally, the recognition of a person.[15] There are dedicated neural areas in the superior temporal sulci (STS) that respond selectively to voices more than to other sounds in the environment (Belin *et al.* 2000).[16] In particular, the anterior area is dedicated to voice processing and not the linguistic analysis of speech sounds (von Kriegstein *et al.* 2003). One reason why the recognition of a speaker happens so quickly is that it involves such early processing areas in the brain, such as the fusiform gyrus for face recognition and regions of STS for voice. Activation of these cortical regions helps us to quickly identify and form the capacity to recognize a person, and there is evidence from neuroimaging of the interaction of face and voice areas in the recognition of a speaker (von Kriegstein *et al.* 2005). Cross-modal integration occurs where we focus our attention visually on a speaker we want to listen to. We have to direct our visual attention in order to enhance our hearing of a particular person speaking at the other end of a table. In a crowded room or restaurant where many people are speaking at once, we need to direct our visual attention to orient our auditory perception to the *person* speaking as the source of the sounds we want to hear.

All of this shows that even without locating the content of speech sounds in auditory perception, there can still be a direct meeting of minds, due to our awareness of an individual subject or person as the *source* of the speech sounds we are hearing. It is natural to take the subject to be the originator of the meanings we attach to the words we retrieve from her acoustic sound stream in the course of lexical processing.

I hear you as saying something. But what I hear you as saying is the result of the meanings the words you utter have for me. I can only hear your words with the meanings I attach to them. Who else's meanings would I use, other than my own? Thus, if a word (a set of phonemes) has certain meaning for you that it does not have for me, I can only hear it with my meaning, not with yours. Similarly, if a word is ambiguous for me but not for you, you simply cannot *hear* it with one of the meanings I give it, even if you can come to know it has that other meaning for me. Thus, if a shop assistant in Glasgow says, 'Would

[15] Although voice is a property of a person, a person can have more than one voice. We talk of someone's 'singing voice' and are surprised by it even though we know their 'speaking voice'.

[16] STS is the area where we find mirror neurons that resonate to observed actions of others, perhaps suggesting that we may be able to find neural evidence for the motor theory of speech perception, according to which we are helped to perceive sounds by our motor system's matching of the articulatory movements that produce them.

you like a wee poke?' I may hear the question as meaning, 'Would you like a paper bag?' while you may not.

7. How Do I Come to Know What Others are Saying?

Through the early learning of word meanings, children come—in contexts involving another language user and under conditions of joint attention—to attach a meaning to a word or sound they hear. It may appear as though the child is being given the meaning of that word, but from the child's point of view, it is learning to endow that sequence of sounds with a meaning. And it is the meanings speakers have endowed their words with that count as their default understanding of these words whenever they encounter them. This is what they hear the words as meaning. Of course, the default case can be overridden and one can be wrong to take this to be what someone else means by their use of these words. But it will fix, initially, what we *hear* them as saying. And this will be a matter not of detecting meanings in their overt speech, but in our contributing the meanings we usually attach to the words we perceive to our understanding of their speech. When these are also the meanings they attach to those words, we will count as knowing what they mean, as being correct in what we take them to be saying.[17]

In listening to your voice, I am directly in contact with you as a person; but in hearing you say certain things, I supply meanings for the words I recognize you to be uttering. I simply always experience these words, at first, as said or heard with the meanings they have for me—the meanings I have endowed them with. The immediacy of the experience I have in hearing what you say is due to the inseparability for me of these words and these meanings. If my immediate understanding of you does not work, and the default condition—where you and I have attached the same meanings to these words—fails, I need to distance myself from my immediate understanding and engage in interpretation.

Notice how this picture differs from the one McDowell seeks to resist. He told us that 'the significance of utterances in a language must, in general, lie open to view, in publicly available facts about linguistic behavior in its circumstances' (McDowell 1998a: 314). Otherwise, understanding would

[17] For an account of how we first attach meaning to words and use these in a first-person-based epistemology of understanding, see Smith (2006b, 2006c).

consist in 'hypotheses about inner states of the speaker lying behind the behavior' (McDowell ibid.: 314, 331). But on my picture, where meanings do not 'lie open to view' on the surface of linguistic behavior, we are not as listeners hypothesizing about others' inner states. We are just hearing the words retrieved from the speaker's speech signal with the meanings they must have for us. Initially, we have no choice but to hear them this way. Our task is not to infer what goes on with others, but just to hear them as we are naturally and immediately inclined to do. By default, what we take someone to be saying will be what they are saying. According to this view, the direct connection with the mind of another will occur via perception of the *source*, and not the *content* of the speech. It is the sound of a person, not what the person says, that establishes a meeting of minds.

I can be mistaken about what you are saying, but if you are addressing me, I will not be mistaken about its being *you*—my interlocutor—who seems to be saying these things. But can we not make mistakes about who or what is addressing us in speech? Not if the experience is veridical, I say. But what is required for our experience of speech to be veridical? Nudds points out that 'the experience of sounds commits us to the existence of something other than sounds'. The experience of a violin being played is veridical if there is a violin being played and it is the source of the sounds heard.

Sound waves carry information about the things that produce them, and, thus, we can perceive those things through the auditory experiences that represent those objects and their properties. The auditory speech system functions to produce experiences of hearing a voice, and having a voice is a property of a person. In the auditory perception of speech, we hear the speech sounds as coming from a person who is speaking to us. So, when it is veridical, our experience of speech sounds commits us to the existence of voices which belong to persons. The correctness conditions for auditory experiences of speech sounds have these existential commitments, but they do not carry commitments about meanings residing on the surface of speech or lying open to view. Since the experience of sounds and their sources does not commit us to a surface for speech. Our experience of sounds does not provide location for those sounds, only for their sources. Usually, we know who is producing these sounds, or think we do. And yet, speech synthesizers produce powerful illusions as if we were encountering a person with a personality. We hear meaning in what these faux voices say, in the usual way, but there is a strong pull to misperceive the source of the sounds as a person. People frequently describe the 'voice' of a speech synthesizer in automatic telephone or satellite navigation systems as sounding insistent or strident or cold. These are abnormal cases. Normally, we hear *someone* saying

such and such, and that person is perceived as being the source, or apparent source, of the sounds.

Sounds from speakers in our immediate community are heard *as* meaningful but the linguistic meanings and forms on which perceived meaning depends are not there in the sounds we hear. The internal organization of language users provides all the linguistic significance there is.[18] The real object of speech perception is the *voice* of the producer. We hear the minds of others in the sounds they make but not what they say. The sounds do not carry meaning; they trigger the awareness of meaning in us. Producing meaningful speech sounds is like tapping out a tune for others to catch on to, and those who have learned the same tunes may hear the sounds in the same way. All the richness we hear in meaningful speech is not in the sounds but in us.

References

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). 'Voice Selective Areas in Human Auditory Cortex'. *Nature*, 403: 309–12.

Chomsky, N. (1987). 'Reply'. *Mind and Language*, 2: 178–97.

Dehaene-Lambertz, G., Pallier, C., Serniiclaes, W., Sprenger-Charolles, L., Jobert, A., and Dehaene, S. (2005). 'Neural Correlates of Switching from Auditory to Speech Perception'. *NeuroImage*, 24: 21–33.

Evans, G. (1975). 'Identity and Predication'. *Journal of Philosophy*, 72: 343–63.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, Mass.: MIT Press.

Fukuda, S. and Kostov, V. (1999). 'Extracting Emotion from Voice'. *Systems, Man, and Cybernetics, IEEE SMC '99 Conference Proceedings*, 4: 299–304.

Harris, J. and Lindsay, G. (2000). 'Vowel Patterns in Mind and Sound', in N. Burton-Roberts, P. Carr, G. J. Docherty (eds.), *Phonological Knowledge: Conceptual and Empirical Issues*. Oxford: Oxford University Press, 185–206.

Liberman, A. M., Isenberg, D., and Rakerd, B. (1981). 'Duplex Perception of Cues for Stop Consonants: Evidence for a Phonetic Mode'. *Perception and Psychophysics*, 30(2): 133–43.

MacDonald, J. and McGurk, H. (1978). 'Visual Influences on Speech Perception Processes'. *Perception and Psychophysics*, 24: 253–7.

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

——(1998b). *Mind, Value and Reality*. Cambridge, Mass.: Harvard University Press.

Marslen-Wilson, W. and Tyler, L. (1981). 'Central Processes in Speech Understanding'. *Philosophical Transactions of the Royal Society*, 8: 317–22.

[18] Notice, however, that although some of the objects of our linguistic knowledge are mental and internal, they depend for their content on others and on aspects of the environment.

O'Shaughnessy, B. (2000). *Consciousness and the World*. Oxford: Clarendon Press.

Quine, W. V. (1960). *Word and Object*. Cambridge, Mass.: MIT Press.

Scherer, K. R., Koivumaki, J., and Rosenthal, R. (1972). 'Minimal Cues in the Vocal Communication of Affect: Judging Emotions from Content-Masked Speech'. *Journal of Psycholinguistic Research*, 1: 269–85.

Smith, B. C. (2006a). 'What I Know When I Know a Language', in E. Lepore and B. C. Smith (eds.), *The Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press.

——(2006b). 'Davidson, Interpretation and First-Person Constraints on Meaning'. *International Journal of Philosophical Studies*, 14(3): 385–406.

——(2006c). 'Publicity, Externalism and Inner States', in T. Marvan (ed.), *What Determines Content? The Internalism/Externalism Dispute*. Cambridge, Mass.: Cambridge Scholars Press.

Trout, J. D. (2001). 'The Biological Basis for Speech: What to Infer from Talking to the Animals'. *Psychological Review*, 108: 523–49.

——(2003). 'Biological Specialization for Speech: What Can the Animals Tell Us?' *Current Directions in Psychological Science*, 12(5): 155–9.

von Kriegstein, K., Eger, E., and Kleinschmidt, A. (2003). 'Modulation of Neural Responses to Speech by Direction Attention to Voice or Verbal Content'. *Cognitive Brain Research*, 17: 48–55.

——Sterzer, P., and Giraud, A. (2005). 'Interaction of Face and Voice Areas During Speaker Recognition'. *Journal of Cognitive Neuroscience*, 17(3): 367–76.

Warren, R. (1970). 'Perceptual Restoration of Missing Speech Sounds'. *Science*, 167: 392–3.

Wittgenstein, L. (1983). *Remarks on the Foundations of Mathematics*, rev. edn. Cambridge, Mass.: MIT Press.