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.

10

The Motor Theory of Speech Perception[1]

CHRISTOPHER MOLE

There is a long-standing project in psychology the goal of which is to explain our ability to perceive speech. The project is motivated by evidence that seems to indicate that the cognitive processing to which speech sounds are subjected is somehow different from the normal processing employed in hearing. The Motor Theory of speech perception was proposed in the 1960s as an attempt to explain this specialness. It is currently enjoying a renewal of interest, partly on account of our developing understanding of mirror-neurons (the existence of which is suggestive but not conclusive) and partly on account of some recent work using Transcranial Magnetic Stimulation (Fadiga *et al.* 2002).

This essay has two parts. The first is concerned with the Motor Theory's explanandum and shows that it is rather hard to give a precise account of what the Motor Theory is a theory *of*. The second part of the essay identifies problems with the explanans: There are difficulties in finding a plausible account of what the content of the Motor Theory is supposed to be. The agenda of both parts is rather negative, and problems will be uncovered rather than solved. In the concluding section, I shall suggest where one might look if one wants to solve the Motor Theory's problems, but it is unclear whether the Motor Theory's problems *ought* to be solved, or whether the whole theory should be abandoned.

I.

Psychologists were first persuaded that speech perception is unlike the perception of other sounds by the failure of attempts to build reading machines

[1] This work was done while the author held the William Alexander Fleet Fellowship. Thanks are due to Ms Julia Fleet, who sadly died while this volume was in preparation.

for the blind. Nowadays our computers do a good job of rendering a written text into speech, but it was not always so easy. After the Second World War there were lots of recently blinded people, for whom a machine that could read aloud would have been a very good thing. There was also rather little computing power available. The task of building a machine that would turn text into *speech* seemed to be a practically impossible one, but the task of building a machine that would make *some* sort of distinct *noise* for each of the letters in a text seemed straightforward enough. The more ambitious task of building a machine that would make a distinct noise for each of the separate speech sounds in the text (that is, each of the *phonemes*) also looked like a real possibility.

Such a task may have been computationally tractable, but as a substitute for reading it was completely hopeless. The thing that made the project hopeless was that the listener's ears just couldn't keep up. If one's reading machine was making its sounds at a pace that was anything like the pace at which our mouths make sounds when we speak, then it was making sounds at a rate that was far too fast for the listener to resolve. If the sounds were given slowly enough for the listener to resolve them, then they came far too slowly to effectively communicate a text. Whichever sounds are allocated to individual letters or phonemes, the resulting auditory presentation of words takes much longer to comprehend and puts a much greater load on working memory than written text or speech. Training does little to help. As Alvin Liberman, one of the first and most prominent researchers on speech, puts it, 'Only the sounds of speech are efficient vehicles for phonetic structures; no other sounds, no matter how artfully contrived, will work better than about one tenth as well' (Liberman 1990).

The blind war veterans stood no chance at all of learning to recognize the rapidly presented sounds of the 'reading machines', but they had, like the rest of us, learned in infancy how to recognize the phonemes that make up speech, and these phonemes can be resolved at a very fast rate indeed. Why, then, could they not learn to resolve any other sounds at anything approaching the same pace? It seems that *somehow* speech must be special. Perhaps its specialness has something to do with its being learned so early on. Perhaps it is special because it is massively more familiar than other sounds, and massively more practiced, but perhaps it is special in other ways, too. Language so often is.[2]

[2] Not everyone in the psychology of speech perception has signed on to the 'speech is special' view, but dissent is certainly rare. The dissenting view can be found in the discussion of Miller's Auditory-Perceptual Pointer Model in Klatt (1989). Or Fowler's 'Direct Realist' theory (Fowler 1986).

To be told that speech is special is to be told *something* of interest, but it is not yet to be told anything sufficiently precise for us to be in a position to assess a scientific theory that purports to explain this specialness. Researchers (most of them based at Haskins Laboratories where the work on reading machines had been carried out) uncovered multiple effects that were thought to illuminate the *way* in which speech is special. Four, in particular, attracted the attention of psychologists more widely. These were: the categorical perception of speech, the lack of invariance, the 'duplex' perception of speech, and the McGurk effect. I'll explain each of these in turn.

Categorical Perception

To understand the categorical nature of speech perception, consider the discrimination function for normal non-speech sounds—the function that gives, for each magnitude of a perceivable property, the smallest discernible difference from that magnitude. By presenting normal listeners with pairs of sounds, and asking whether the two sounds sound the same or different, we can work out, for any variation in magnitude on any dimension, what the smallest discernible difference in that magnitude is. And with this data, we can plot a graph showing, for each of the pitches we have tested, how much deviation is needed from that pitch for the two sounds to sound different, or we can plot a graph showing how much change in *tone* is needed for two sounds to sound different, and so on for various other dimensions along which acoustic properties vary. These graphs showing the discrimination function for simple acoustic properties will tend to be gently curving lines. As we test a subject's ability to discern variations from louder and louder sounds, we don't come across any one particular volume from which deviations are much more easily detected than are deviations from slightly louder, or slightly quieter sounds. The ability to discriminate one pitch from another changes smoothly as a function of the initial pitch. The same is true for volume and timbre. This *isn't* true if the dimension of variation is a dimension that is relevant to differences in speech sound. If the sound we are testing is a speech sound—such as the syllable /pa/—and if the magnitude we are varying is one that can make a difference to which phonemes are heard—such as the voice onset time—then there *is* a privileged magnitude such that deviations from that magnitude are much easier to detect than deviations from the previous and from the subsequent magnitudes. The graph on which we plot the smallest discernible differences will not curve smoothly. There will be a sharp trough in it at the point where we suddenly become extremely sensitive to small variations. It will be helpful to spell out this example in some detail.

The syllables /pa/ and /ba/ both begin with bilabial plosive consonants. That is to say that the initial sound in both is produced by closing the lips, allowing a little air to build up behind them, and then releasing that air. They differ in the moment during this performance at which the vocal chords begin to vibrate. For a /ba/ the vocal chords vibrate almost as soon as the air is released. A /pa/ is produced when the vocal chords vibrate a little later. The time between the release of the air held at the lips and the beginning of the vocal chord vibration is called voice-onset time (VOT). VOT can be varied continuously and a spectrum of phonemes can be artificially produced, each of which differs from the preceding phoneme only by, say, an increase in VOT of 10 ms.

If we listen to each of the sounds in this continuously varying spectrum they are not *heard* as varying continuously. We don't hear a sequence of /ba/s gently sloping off into penumbral cases which gradually become recognizable as /pa/s. Instead, the first half of the spectrum is all heard as more or less the same /ba/ sound, and the second half of the spectrum is all heard as more or less the same /pa/ sound, and there is a narrow band in the middle of the spectrum, when VOT is around 26.8 ms, where subjects differ as to how they hear the sound. At this transition point subjects can recognize very slight variations in VOT, since these slight variations are enough to move the sound from one category to the next. From the point of view of the subject, the difference between a syllable with VOT of 26.8 ms and a syllable with VOT of 36.8 ms sounds like a really big difference, whereas a difference of the same objective magnitude, between, say, VOT of 30 ms and of 40 ms, sounds like a very slight difference. In the first pair, one of the syllables sounds like a /ba/ and the other sounds like a /pa/. In the second pair, the two syllables are more or less indistinguishable /ba/s.[3] This distinctive discrimination function is found for many of the dimensions of variation that, like VOT, make the difference between two consonants, although it isn't found for vowels.

This finding by itself is rather unremarkable. It isn't surprising to find a difference between the discrimination functions for complex sorts of variation, such as variation in VOT, and the discrimination functions for simple sorts of variation, such as variation in pitch. And it isn't surprising that we exploit the categorical perception of such variations in the boundaries that we use to indicate semantically relevant differences in speech. Something similar to the categorical perception of VOT holds for the perception of colors: we perceive two reds as more similar than a red and a yellow, even though the objective difference between the wavelengths may be the same. If you were designing a communication system with colored flags, you'd assign different meanings to red and yellow, instead of making the difference between two reds a semantically significant one. The categorical perception of speech sounds might be used in the speech code in something like this way, without the fact that the phonemic boundaries coincide with the boundaries of categorical perception showing any connection between categorical perception and the specialness of speech perception.

Categorical perception might be a feature of the perception of complex variations—a feature that the speech code makes use of, but that is not a special feature of speech perception as such. Two things suggest this. First, it is found that there is categorical perception of some non-speech sounds, as when, for example, musicians show categorical perception for semitone boundaries, or when normal people show unlearned categorical perception effects for certain 'buzz, noise, and relative timing continua' (Harnad 1987: 9). Secondly, and more impressively, categorical perception is found in some creatures that lack language (and lack anything that might be thought of as a proto-language). The categorical perception of speech sounds was found in chinchillas by Kuhl and Miller (1978), in Japanese quail by Kluender *et al.* (1987), and in the Mongolian gerbil by Sinnott and Mosteller (2001).

This is not to say that categorical perception doesn't *contribute* to the specialness of speech, only that the categorical nature of speech perception isn't by itself adequate for characterizing the way in which speech is special. There is, in fact, good evidence that categorical perception *is related* to the specialness of speech coming from the fact that the categorical perception of speech is affected by the role it plays in language. We know that it is affected in this way because we know that the pattern of categorical perception that people show for speech sounds is affected by the pattern of phonemic contrasts that can make for a difference in meaning in their native language. English speakers, for whom the difference between /l/ and /r/ can be a semantically relevant difference, perceive a continuous change from one to the other as a categorical change. Their discrimination function has the trough that is characteristic of categorical perception. In Japanese, the difference between /l/ and /r/ is not a difference that ever distinguishes two phonemes. (That is to say that the difference between whether an /l/ or an /r/ was said never makes for a difference in which words the speaker uttered, unlike in English, when

[3] This is a slight simplification of the experimental procedure used, but not, I think, a significant one. The usual procedure is not pair-wise comparison but comparison of triples. Subjects are presented with a pair of neighboring sounds—*A* and *B*—and then presented a third sound—*X*—which is just the same as either *A* or *B*. Their task is to discern which of the two initial sounds *X* repeats. With a VOT continuum where neighbouring items differ by 10 ms, subjects perform close to chance on the *ABX* task for almost all of the spectrum except for the point (when VOT is around 26.8 ms) at which *A* is heard as a /ba/ and *B* as a /pa/.

the existence of a mapping from simple features of the stimuli to perceived categories is not the norm, the absence of such a mapping cannot be evidence of specialness. It does not illuminate the specialness of speech.

It is tempting to think that one could create the appearance of an invariance problem for *any* categorization task if one started with a sufficiently low-level description of the input. One can recognize a large number of faces viewed at various angles and in various lights, and one can recognize them beneath a wide range of hats, spectacles, false noses, and so on. The invariances which one exploits in face recognition are at such a high level of description that if one were trying to work out how it was done given a moment-by-moment mathematical description of the retinal array, it might well appear impossible. There is surely no simple pattern of retinal stimulation that always and only occurs when I see a face as being my brother's. My ability to recognize him is not a matter of my being triggered by some simple property of the array he projects to my retina. A well-trained boy scout can recognize granny knots, reef knots, and sheet bends by touch alone, but the features that he uses when making these discriminations would be extremely hard to recover from a moment-by-moment presentation of the pressure that each knot exerts on his fingertips. There is certainly no profile of finger-tip pressures that is always or only associated with granny knots, and so the same sort of invariance shown by the acoustic signal for speech is shown by the haptic signal for knots. But it would be absurd to draw any conclusions about the specialness of boy scout knot-perception on the basis of this invariance. It is simply that the boy scout does not categorize on the basis of simple features that can be discerned in the moment-by-moment description of fingertip pressures. The speech spectrographs for which we find invariance in the speech signal are moment-by-moment descriptions of the low-level properties presented to the ear. It is no surprise that they fail to show any features that correspond with the phonemes we hear in speech, and it is no indication of specialness.

Duplex Perception

Duplex perception is a strange phenomenon and it occurs in a strange context, making it rather hard to interpret, but in their article 'A Specialization for Speech Perception', Liberman and Mattingly (1989) rest their whole case for the existence of cognitive resources that are devoted solely to speech processing on the phenomenon of duplex perception. Duplex perception occurs when headphones are used to play a different sound to each ear. More specifically, it occurs when the sound given to the first ear is a speech sound: a syllable like /da/ or /ga/, but a speech sound that has been doctored so that the initial burst of acoustic energy is absent. The result of this doctoring is that the sound, if heard in isolation, is ambiguous between /da/ and /ga/. The sound which is presented to the other ear is just that burst of acoustic energy needed to disambiguate the doctored sound—the burst of rising frequency sound that, if added to the first sound, would make it sound like a /ga/; or the burst of falling frequency sound that would make it sound like a /da/. This second sound, if heard in isolation, sounds like a little chirp. It does not sound like speech. Here is Liberman and Mattingly's (1989) account of what it's like to hear this combination. (I've replaced their jargon with mine):

> Listeners hear two sounds, one at each ear. At the ear receiving the [second sound], they hear a non-speech chirp, just as they do when the [second sound] is presented in isolation. At the ear receiving the [first sound] they hear /da/ or /ga/. But, surprisingly, these latter percepts are not ambiguous, as they were when the [first sound] is presented in isolation; rather, they are unambiguously determined to be /da/ or /ga/ by the [fact about whether the second sound is a chirp of rising frequency or falling frequency], just as when the undivided syllable is presented in the normal way. (Liberman and Mattingly 1989: 490)

Perhaps this result is, as Liberman and Mattingly say, a surprising one, but is it evidence of specialness? One would not expect this mingling of the sounds presented at either ear if simple non-speech sounds were presented, but, as we emphasized above, simple sounds are not the relevant control group. To see if the duplex effect shows speech to be special, we should compare speech sounds to non-speech sounds of comparable complexity. When Fowler and Rosenblum (1990) did this, comparing the duplex perception of syllables with the duplex perception of wooden and metal door slams, the speech sounds behave in more or less the same way as the non-speech sounds. Duplex perception seems not to indicate specialness.

The McGurk Effect

The story so far is this. We are trying to understand how it is that speech perception differs from normal perception in such a way that speech sounds can be resolved much faster than other sounds. We have looked at three phenomena that are said to illuminate this specialness. Two of these phenomena (the lack of invariance and duplex perception) we found to tell us nothing about the specialness of speech *per se*. They did nothing more than point towards some ways in which the perception of complex, composite sounds can be expected to differ from the perception of simple sounds. The other phenomenon we have looked at is the categorical perception of speech. We found there to be good evidence that this phenomenon is related to the specialness that we are trying to understand, but we also saw some good reasons to doubt that this relationship is an especially intimate one. We turn now to the fourth of the phenomena that has been thought to cast light on speech's specialness. This is the phenomenon known as the McGurk effect. In the McGurk effect, the syllable that a speaker is heard to have said is found to be influenced by lip movements that the speaker is *seen* to produce, as well as by the acoustic information given to the hearer's ear (McGurk and MacDonald 1976). The effect occurs in the following way: A video is taken of a speaker repeating the syllable /ga/ and an auditory recording is made of the speaker repeating the syllable /ba/. When the auditory recording is heard alone, listeners accurately recognize the syllable heard as a /ba/. If they are hearing these syllables *while watching the video of appropriately timed /ga/s being mouthed*, then the listener is subject to an illusion in which the sound *heard* is reported as being /da/.

The illusory syllable splits the difference between the syllable heard and the syllable seen. /ba/, which gets presented to the ears, differs from /ga/, which gets presented to the eyes, in its place of articulation. /b/s are bilabial (which is to say that they are articulated at the lips), while /g/s are velar (which is to say that they are articulated towards the back of the throat). What listeners hear in the McGurk effect is a /d/, which is an alveolar consonant, made towards the middle.

The effect may be thought of as a somewhat surprising instance of the context effects that we discussed under the heading of 'The Invariance Problem'. The facts about which phonemes a burst of sound is heard to contain are, as we saw, influenced by a great many features of the *context* of the sound. What the McGurk effect shows is that context effects are not limited to effects of a sound's *auditory* context. The effect is an effect of *visual* context on heard sound. We were unmoved by the invariance problem because context effects are the norm for the perception of complex stimuli. The McGurk effect is more impressive because cross-modal context effects are less obviously normal.

But cross-modal context effects are not *entirely* exceptional. If the McGurk effect shows that there is something special about speech, it is not because there is anything special about the fact that speech is a stimulus that is subject to influence from concurrently presented visual stimuli. Lots of stimuli other than speech are subject to that sort of influence. The influence is most frequently discussed in connection with examples from outside the auditory domain, such as the illusion of self-motion produced by motion in the periphery of the visual field (Lee and Lishman 1975). Cross-modal effects are found in the auditory domain, too. The McGurk case is not the only case in which vision and auditory modalities combine in illusory ways, and so it does not show that such illusory combinations are special to speech. Saldana and Rosenblum (1993) have shown that judgments of whether a cello sounds like it is being plucked or bowed are subject to McGurk-like interference from visual stimuli. Sound and vision can also interact to produce *visual* illusions, not just auditory ones. The number of flashes that a subject seems to *see* can be influenced by the number of concurrent tones that he *hears* (Lewald and Guski 2003).[4] It is not special to speech that sound and vision can interact to produce hybrid perceptions influenced by both modalities, without the subject's being aware of the influence.

This is not to say that the McGurk effect shows us nothing special about speech. The McGurk effect does reveal an aspect of speech that is in need of a special explanation because the McGurk effect is of a much greater *magnitude* than analogous cross-modal context effects for non-speech sounds. Although non-speech sounds *are* influenced by vision in much the same way that speech sounds are influenced in the McGurk effect, they do not seem to be influenced to the same extent. The particular *degree* of influence from vision on what seems to the subject to be the auditory perception of speech does seem to be an effect that needs to be explained by the postulation of something special about speech processing.

This is worth emphasizing because a *quantitative* difference between speech perception and the perception of other sounds may be explained by reference to a *quantitative* kind of specialness on the part of speech. Given that sounds in general are *somewhat* susceptible to McGurk-like effects, we do not need to postulate very much specialness to explain why speech is distinguished from other sounds by the degree of its susceptibility to such effects. Perhaps the unusually high susceptibility of speech sounds to the McGurk effect is explained by the fact that the contexts in which speech sounds are heard are, to a greater extent than are the contexts of other sounds, occasions where the source of the sound is visible and where the visual information is a potential source of useful disambiguating information. The existence of other auditory-visual cross-modal illusions shows that there are mechanisms in place by which visual stimuli can influence the perception of sound. The fact that speech sounds are unlike other sounds in the degree to which it is *useful* to make fine discriminations, and the fact that speech sounds are unlike other sounds in the frequency with which visual information from the sound source is *available* for helping with such discriminations, could together explain why the mechanisms of cross-modal influence (not

[4] A vivid demonstration, described in Kamitani and Shimojo (2001), can be found at: <http://www.cns.atr.jp/~kmtn/audiovisualRabbit/index.html>.

special in themselves) come to be especially influential on the perception of speech.

The magnitude of the McGurk effect does reveal something special about the psychology of speech perception, but the specialness accounting for the McGurk effect might just be that the normal mechanisms of audiovisual interaction are especially active for speech on account of the uncommon availability of occasions on which they can come into play, and the uncommon utility of their doing so. Our final verdict on the explanandum of the Motor Theory of Speech Perception is this: the fact of speech's fast resolution needs to be explained, but the phenomena that have been discussed as if they were revealing of the ways in which speech is special turn out to tell us rather little. The McGurk effect does show something special about speech perception, but fails to make anything clear about what sort of explanation this specialness needs. It fails to tell us whether the perception of speech is special because it differs from normal auditory perception by degree, or by some qualitative difference. I want to turn now to the theory that these various phenomena are supposed to support, and that purports to give an explanation of them, and of the specialness with which we began.

2.

Any interpretation of a scientific theory is probably mistaken if the theory is interpreted as saying something trivial, or something very obvious. It is equally likely to be mistaken if the theory is interpreted as saying something obviously false. The most discussed theory of speech's specialness is the Motor Theory of Speech Perception. The task of saying what the contents of that theory are proves to be much harder than one might expect. This is not because the theory has not been given a canonical statement, but because the theory seems from some points of view to be saying something trivial and from other points of view to be saying something that is obviously false.

The canonical statement of the theory was given in 1985 when Liberman and Mattingly wrote 'The Motor Theory of Speech Perception Revised'. They tell us that 'The first claim of the Motor Theory, as revised, is that the objects of speech perception are the intended phonemic gestures of the speaker'. 'First and fundamentally', we are told, 'there is the claim that phonetic perception is perception of gesture' (Liberman and Mattingly 1985: 21). How are we to understand this claim? There are at least two possibilities, suggested by a familiar

distinction from discussions of perceptual epistemology. In those discussions, we often encounter the distinction between two different perceptual relations distinguished in natural language by the difference between perceiving an entity and perceiving *that* something or other is the case. There are, corresponding to these two perceptual relations, at least two ways in which Liberman and Mattingly's claim about the object of speech perception could be interpreted. It could be a claim about the sort of thing that goes in the *y* place in true sentences of the form 'He heard *y*' (when the hearing in question is an instance of speech hearing). Or, alternatively, it could be a claim about the sort of thing that goes in place of the *P* in true sentences of the form 'He heard that *P*' (when the hearing in question is an instance of speech hearing). When Liberman and Mattingly talk of perceiving 'gestures', what they mean is that when we hear a /b/ the object of our perception is a bilabial plosive gesture; that when we hear a /n/ we hear an alveolar nasal gesture; and so on. What isn't clear is which of the two perceptual relations these gestures are supposed to be the objects of.

On a first reading, Liberman and Mattingly are claiming that when a listener hears speech, it is true that he hears intended phonemic gestures. If this is the correct reading of their claim, then their claim is surely true. Sentences of the form '*x* heard *s*' are true if and only if there is something identical to *s* that *x* heard. This context for '*s*' is an extensional one. So, for example, it is true that Miss Scarlett heard the gunshot just if it is true that there is something that Miss Scarlett heard, and true that that thing was the gunshot. It doesn't matter whether she recognized it as a gunshot, or even if she has any concept of gunshots. If Miss Scarlett thought that she was hearing a champagne bottle being opened, but the sound was in fact that of a gun firing, then it is nonetheless true that Miss Scarlett heard the gunshot. She heard it; she was just mistaken about *what* she heard. Understood in this way—as a claim about the object of the *x* heard *y* relation—the 'first claim of the Motor Theory' is uncontroversial. A speech act *is* a sequence of intended phonemic gestures, so the truth of sentences of the form 'He heard intended phonemic gestures' is guaranteed by the existence of truths of the form 'He heard the speech act'. Perhaps, like Miss Scarlett, we do not know what it is that we are hearing. The claim that we hear phonemic gestures is compatible with the claim that we hear such gestures unbeknownst to us.

This claim is true, and obviously so, but it won't do as an interpretation of what Liberman and Mattingly intend, because it can't do the work that the Motor Theory is supposed to do. To claim that the perception of speech is the perception of gesture in this sense is not to identify a feature that makes speech special. Even if speech perception were exactly the same as normal audition, then the object of perception in this sense would still be the gesture. If we want the Motor Theory to be making a non-obvious claim, we should understand it to be making a claim about the other of the two sorts of perceptual relation: It must make a claim about the object of the relation '*x* heard *that P*'. More is required for the truth of sentences with the form '*x* heard *that P*', than was required for the truth of '*x* heard *s*' because this context is an intensional one. Although Miss Scarlett can truly say, on learning about the circumstances of the death, 'I heard the gunshot', she cannot truly say that she heard that there was a gunshot. Not if she took it for the opening of a bottle. Hearing *that* there was a gunshot requires (among other things) that the event be heard *as being* a gunshot. If the Motor Theory claims that intended phonemic gestures are the objects of speech perception in the sense that speech perception involves perceiving *that* there were certain intended phonemic gestures, then the theory is committed to our hearing speech *as being* a set of intended phonemic gestures.

The claim that we hear speech as being phonemic gestures is rather counter-intuitive, and it is easy to produce an argument showing it to be false. Suppose we have a listener, who, being in the grip of some false theory about the phonemic gestures, believes that /b/ is not a bilabial plosive, but a postalveolar trill. Such a listener is easy to imagine. It is equally easy to imagine that such a listener is listening to a speech replete with instances of /b/, and that he is taking the experience at face value. He need not believe he is subject to any sort of illusion. If hearing the /b/s in the speech involved hearing them *as* bilabial plosives, then this thinker would be guilty of some sort of *irrationality*, just as one who believed that no gun had been fired would be guilty of irrationality if he persisted in his belief of having experienced a whole sequence of events as being gunshots. A false theory about phonetics is not so readily refuted: speech perception doesn't present us with the underlying gestures as a part of the content of experience. We could make the same point in more Wittgensteinian tones: One who is searching for a labiodental fricative may need a look-up chart to tell him when he has successfully found one. One who is searching for a red flower famously needs no such look-up chart. If phonemic gestures were given in the contents of experience, then 'labiodental fricative' would behave like 'red' in this respect. The contents of experience have to be non-inferentially *given*, and phonemic gestures aren't given in that way. It is not the case that when *x* perceives speech, *x* perceives that certain phonemic gestures were made.

It can seem that the Motor Theory is stuck with an irresolvable dilemma. Either it is making a claim about the relation of hearing, or it is making a claim about the relation of hearing *that*. If it is making the first claim, then it is saying something true, but something that cannot contribute to our understanding of the specialness of speech. If it is making the second claim, then it is saying something demonstrably false. This dilemma only arises because we take the Motor Theory to be making a claim about the object of a perceptual relation *in which the subject is a person*. Can the theory avoid these problems if it retreats to making a claim about a *subpersonal* relation? The first horn of the dilemma remains—'*x* heard *s*' is an extensional context for *s*, whatever we put in the *x* place, so the identity of speech with phonemic gesturing guarantees trivially that phonemic gestures are perceived when speech is. The move to a subpersonal perceiving subject can't help here. But perhaps it helps with the dilemma's other horn. The second horn of the dilemma does look like a place in which the tactic of moving to a subpersonal perceiving relation seems more promising. The problems at that horn were problems that arose because the theory seemed wrongly to convict a certain kind of thinker of irrationality. These are problems that the move to the subpersonal may help with, since the notions of rationality and irrationality are notions that lose their grip when we move to the subpersonal.

To avoid the problems set out above, the Motor Theory needs to be interpreted as making a claim about a subpersonal perceiving relation, and it needs this perceiving relation not to be an extensional one, or else the problems associated with the first horn of the above dilemma will arise again. How should we understand this notion of a subpersonal, non-extensional 'perceiving that' relation? When we were at the personal level, we had some intuitive grasp of the way in which the personal 'perceiving that' relation fails to be extensional, but at the subpersonal level, much more work is needed if we are to understand the source of the non-extensionality of the 'perceiving that' relation. At the personal level, hearing *that* there was a gunshot requires hearing the sound *as* a gunshot. If we are to make sense of the Motor Theory as claiming that speech is represented as phonetic gestures at the subpersonal level, then we shall need a subpersonal notion of representing *as*, corresponding to the personal-level notion of hearing *as*. It is a natural thought that the way to understand this notion is through some kind of connection with particular *concepts*. But, for reasons akin to those we've already seen, a conceptually demanding notion won't serve the motor theorist's purposes. One who lacks the concepts of phonemic gestures can nonetheless hear what's being said to him, and even if the thinker has those concepts, they do not seem to be engaged just because speech is being perceived.

The situation we are in is this: The Motor Theory makes the claim that gestures are the objects of speech perception. We are trying to understand what this could mean. We have seen that this can't be understood as being the claim that gestures are the objects of any *personal-level* perceptual relation, or of a *conceptually demanding* perceptual relation, or of an *extensional* perceptual relation. The suggestion might be made that a perceptual relation that avoids each of these problems can be built from the notion of carrying information. A subpersonal representation can *carry information about* some properties of a thing without the thinker needing concepts of that thing, and this notion of carrying information about some aspect of a thing allows us to individuate the contents of representations more finely than extension—and so it appears to enable us to find something non-vacuous to make of the idea that the representation of speech is the representation of vocalic gestures. But this appearance is misleading, and the problem of vacuity arises again. It arises because phonemes are *individuated by* the lip movements that produce them. A glance at the international phonetic alphabet will reveal that phonemes are classified by place of articulation (where in the mouth the sound is made) and by the sort of movement made (plosives, nasals, trills, taps, fricatives, and so on, are ways of moving the mouth parts). What it *is* for a word to contain a given consonant is for its pronunciation to involve mouth movements of a certain sort. Nothing can carry information about phonemes without carrying information about phonemic gestures. The phonemes that make up speech have to be encoded if one is to know what the speaker said, and so any representation that carries information about the words that a speaker said *ipso facto* carries information about the lip movements made.

The thing that is distinctive about the approach of Liberman and Mattingly and other Motor Theorists is that their notion of 'representing speech as phonemic gestures' is not tied to the notion of carrying information *about* such gestures, nor to the thinker's capacity to *think* about phonemic gestures, but instead to the thinker's capacity to *produce* such gestures in his own speech. This, finally, is the motor-related aspect from which the Motor Theory gets its name. It seems, on the face of it, that any move that links our ability to perceive speech to our ability to speak is an unappealing move, since it ought to be possible to hear speech without being able to speak oneself, and it is surely possible to hear speech sounds that one cannot produce oneself. The child born with an ill-formed mouth does not, of course, face deafness. Nor is the poor mimic unable to hear the speech of those with regional accents that it is beyond him to imitate. Liberman and Mattingly (1985) were moved by these sorts of considerations, and cite as an influence on their revision of the Motor Theory the finding by MacNeilage, Rootes, and Chase (1967) that 'people who have been pathologically incapable from birth of controlling their articulators are nonetheless able to perceive speech' (Liberman and Mattingly 1985: 24). On account of these findings, they moved from a claim about the vocal tract *itself* to a claim about an internal *model* of the vocal tract. The theory as revised does not claim that we actually use our mouths and throats in hearing speech, but that the perception of speech involves the use of 'an internal, innately specified vocal-tract synthesizer' (Liberman and Mattingly 1985: 26).

This move from a claim about the vocal tract to a claim about an internal model of the vocal tract brings with it a loss of clarity because it is not immediately obvious what it *takes* for a bit of neural apparatus to constitute an internal model of the vocal tract. Several suggestions could be made to help us understand the claim. One such suggestion would start with the observation that there are some contexts in which one system can be said to model another just if the model can be used to generate reliable predictions about the system modeled. This is the sense of 'model' in use when a load-bearing spring is said to model an inter-molecular force, the effects of LSD are said to model schizophrenia, and, perhaps, some computer programs are said to model the weather. If the Motor Theorist's claim that the apparatus of speech perception includes a model of the vocal tract is understood as a claim that involves this sense of modeling-as-prediction-generation, then problems arise along just the lines that we have already seen. There is a problem with saying that some part of our brain generates reliable predictions about vocal tract gestures if these predictions are personal-level states—normal perceivers of speech make no such predictions. And there is a problem if the 'predictions' in question are subpersonal representations encoding information about the vocal tract—any subpersonal state that encodes information about phonemes also encodes information about vocalic gestures, on account of phonemes being individuated by the vocalic gestures that produce them.

There is, however, another sense of 'model' on which the Motor Theorist's claims stand more chance of being both plausible and explanatory. We can say that one system models another if the first behaves in a way analogous to the behavior of the second, and if it does so *for analogous reasons*.[5] If one system is a model of another in this sense then it can be said to *represent* that system, and the particular states in the model that occupy the same functional role as a particular part of the system modeled can be said to represent those particular parts. This may give us a sense of 'represent' that we can use to understand the Motor Theorist's claim that we represent speech as phonemic gesture.

[5] This is really just a dynamic version of the common or garden concept of a model as we find it applied to model trains, and the like. It is nothing to do with the technical, logician's sense.

If one system can be said to model another in this sense (as opposed to the less demanding and already rejected sense of modeling-as-prediction-generating), then there must be a high degree of symmetry between the causal architecture of the model and that of the system modeled. If the system modeled includes two states, both of which originate from some single feature of the system modeled, then the corresponding states of the model must also share their origins. Similarly, if two states of the system modeled have *different* explanations, there should be a difference in the way the analogous states arise in the model. Moreover, where this causal isomorphism condition requires that a *single* state must feature in the explanation of some two states of affairs, that state must be a *genuinely unified* state. If the rain in Bristol and the flooding in Wales are both caused by the area of low pressure coming from the north, then my meteorological model is not good enough if its representing Bristol as rainy is a result of having access to information about rainfall, and its representing Wales as flooded is a result of having access to some quite separate body of information about river flow. The causal isomorphism requirement can't be met by gerrymandering a disjunctive state comprising both bodies of information.[6]

For the brain to contain a model of the vocal tract (and so, in this sense, for it to be able to represent speech as phonetic gestures), the system by which the brain gets from the sounds at the ear to the representation of words spoken must include a part in which the processing of representation is causally isomorphic with the treatment received by sounds as they pass from vocal chords to lips, and out. Could the brain's processing of speech proceed in a way that would satisfy this non-gerrymandered causal isomorphism requirement so that, in virtue of modeling the vocal tract, the brain could rightly be said to represent vocal tract gestures? I think not, but I do think that we have finally arrived at the correct way to understand the content of the Motor Theory of Speech Perception. The Motor Theory should be understood as a theory about the existence in the brain of a causally isomorphic model of the vocal tract. It may be that there is such a model, but, for a couple of reasons that we shall now turn to, it does not seem to be at all likely that there is.

There are two ways for the brain's processing of speech to model the vocal tract, and each is problematic. The model could work backwards—taking as input the acoustic profiles which the vocal tracts of our interlocutors put out and analyzing them to find the phonetic intentions that set the vocal tract going. This is the most obvious way for such a model to work, but the model could, alternatively, work *forwards*. The model could try out a whole range of various inputs, and use these to generate representations of various acoustic profiles which it then compares to the acoustic profile that has been encoded by the ear. When it finds a match between one of the generated acoustic profiles and the profile perceived, it can identify the input that produced the match. These two alternative ways of using a model correspond to the two strategies that, in the psychological literature, are given the unlovely names 'analysis by analysis' and 'analysis by synthesis'. An analogy will help to clarify the difference between the two approaches. Suppose that Mr Jones is playing notes, one at a time, on the piano, and that Mr Smith has the job of finding out which notes Mr Jones is playing. To help him in his task, Smith is seated in the same room as Jones, and at the keyboard of an exactly similar piano. There are two tactics Smith can use. The speediest tactic would be to lift the lid of his piano, press the sustain pedal so that the strings are not dampened, and watch to see which string resonates. This will be the string that corresponds to the note Jones is playing. The second tactic is for Smith to press each of the notes on his keyboard, one after the other, and listen to hear when the note he plays sounds the same as the note Jones plays. In each case, Smith uses his piano as a model of Jones's. The first tactic is analogous to analysis by analysis. The second tactic is analogous to analysis by synthesis. The method of analysis by analysis is the more efficient of the two.

To meet the causal isomorphism requirement, a speech processor which could successfully detect the /d/ at the beginning of 'di' and the /d/ at the beginning of 'du' would have to do so by the same means, for both /d/s result from the same pattern of gestures. But, as we saw in our discussion of the lack of invariance, the difference in the following vowel causes this pattern of gestures to produce different effects on the features of the acoustic profile. The degree to which there is a lack-of-invariance problem, as discussed above, shows that there can be no model of the vocal tract that satisfies the causal isomorphism requirement and conducts successful analysis by analysis. The causal isomorphism requirement calls for a single part of the model detecting all and only, for example, tongue-backing, while the lack-of-invariance problem tells us that there is no feature of the acoustic profile such that a device that operated as a detector of that feature would be responding to all and only tongue-backing.

Perhaps because they are aware of the tension between their claims about lack of invariance and the possibility of analysis by analysis, the advocates of the Motor Theory have typically accepted the prima-facie less plausible *analysis by synthesis* account, according to which the model of the vocal tract in the brain generates several representations of acoustic profiles, and then compares

[6] Spelling out when exactly a state is genuinely unified and when gerrymandered is, of course, not an easy matter. For our purposes the intuitive notion will have to suffice.

the profiles it has generated to the acoustic profile heard, so that, on finding a match, it is able to identify whatever input to the model of the vocal tract produced a representation that corresponds to the profile presented. Even if an initial bit of analysis by analysis is used to reduce the set of profiles that must be generated to a set of plausible candidates, the task of analysis by synthesis seems so vast that it could only be successfully completed in a realistic time frame if the candidate profiles are produced by massively parallel processing. There are a huge number of possible things that you could be doing with your mouth at any time, and the analysis by synthesis approach requires that a model of the vocal tract try each one of them out to see whether the acoustic consequences it generates match the sound heard. A single model of the vocal tract trying out each of these possibilities in series would have to be working at a colossal rate for speech to be perceived in real time. Analysis by synthesis is only plausible if parallel processing is employed, but parallel processors fail to meet the 'no gerrymandering' clause in the causal isomorphism requirement on modeling. To see that they must do so, suppose there are two models working in parallel, one of which tries out the lip movements corresponding to 'du' and the other of which tries out the lip movements corresponding to 'da'. On one occasion the sound presented is a 'du' and the profile produced by the first model gets matched to the profile of the sound heard. On another occasion the sound heard is a 'da' and the second model produces the match. In both cases a /d/ is recognized, and so to meet the causal isomorphism requirement there must be a single state featuring in the recognition of both sounds—but for there to be such a state is for there *not* to be separate paths operating in parallel. Analysis by synthesis is implausible unless the synthesizing models operate in parallel, but models operating in parallel fail to meet the causal isomorphism requirement, and the states of models operating in parallel therefore fail to count as representations of parts of the vocal tract.

We have tried various ways to interpret the claims of the Motor Theory of Speech Perception, but found none of them to be both plausible and meaningful as an account of how speech perception is done. We have also found that the evidence that has been thought to recommend the Motor Theory's approach is wanting. This might lead us to give the whole thing up. Nonetheless, I claimed above that I would end with a gesture in the direction of a place where these problems could be solved. That place is, I think, closer to the spirit of the original Motor Theory than it is to the more sophisticated theory that was developed in the light of the evidence and arguments that have been reviewed here. We rejected the idea that the apparatus of speech perception is the apparatus of speech production because the perception of speech that one cannot produce is so obviously possible. We saw that this consideration led the Motor Theorists to change their claims to claims about internal *models* of the vocal tract. They went from a claim about there being just one system to a claim about two systems, one of which was a model of the other. This seems to me to have been a source of unnecessary difficulties. There was no need for the original one-system suggestion to be abandoned so entirely. The Motor Theorist can perfectly well claim that there is a single common mechanism of speech production and perception, and that this mechanism represents phonemic gestures, without being committed to the problematic idea that the capacity to produce speech always accompanies the capacity to perceive it. A single common mechanism of production and perception need only have the function of directing speech production and perception when other things are equal; it need not be the case that *whenever* the system is able to perceive it is able to produce. There are plenty of ways in which the performance of a combined production/comprehension system could be impaired on the production side alone. The Motor Theory could then claim that there are resources of speech production, that these resources represent the phonemic gestures, and that these same resources are involved in speech perception. Overlaps in processing resources for comprehension and production are a familiar and obvious idea—the lexicon, presumably, serves both, as do some resources of grammatical analysis. We can understand the Motor Theory as proposing that the overlaps in representational resources continue out to the less abstract levels of representation needed to get the mouth to move in the right way, and needed to get us into a position to know which words are said to us. This is a sketch for the sort of proposal that might be made. If it is to be developed, then we shall need to be a lot clearer about the truth conditions of the various sorts of representation postulating claims that can be made in subpersonal cognitive psychology.

References

Best, C. and McRoberts, G. (2003). 'Infant Perception of Non-native Consonant Contrasts that Adults Assimilate in Different Ways'. *Language and Speech*, 46: 183–216.

Espy, K. A., Molfese, D. L., Molfese, V. J., and Modglin, A. (2004). 'Development of Auditory Event-Related Potentials in Young Children and Relations to Word-Level Reading Abilities at Age 8 Years'. *Annals of Dyslexia*, 54: 9–38.

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). 'Speech Listening Specifically Modulates the Excitability of Tongue Muscles: A TMS Study'. *European Journal of Neuroscience*, 15: 399–402.

Fowler, C. A. (1986). 'An Event Approach to the Study of Speech Perception from a Direct Realist Perspective'. *Journal of Phonetics*, 14: 3–28.

—— and Rosenblum, L. D. (1990). 'Duplex Perception: A Comparison of Monosyllables and Slamming Doors'. *Journal of Experimental Psychology: Human Perception and Performance*, 16: 742–54.

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

Ivry, R. B. and Justus, T. C. (2001). 'A Neural Instantiation of the Motor Theory of Speech Perception'. *Trends in Neuroscience*, 24: 513–15.

Kamitani, Y. and Shimojo, S. (2001). 'Sound-Induced Visual "Rabbit".' *Journal of Vision*, 1(3): 478.

Klatt, D. H. (1989). 'Review of Selected Models of Speech Perception', in W. Marslen-Wilson (ed.), *Lexical Representation and Process*. Cambridge, Mass.: MIT Press.

Kluender, K. R., Diehl, R. L., and Killen, P. R. (1987). 'Japanese Quail Can Learn Phonetic Categories'. *Science*, 237: 1195–7.

Kuhl, P. K. and Miller J. D. (1978). 'Speech Perception by the Chinchilla: Identification Functions for Synthetic VOT Stimuli'. *Journal of the Acoustical Society of America*, 63: 905–17.

Lane, H. (1965). 'The Motor Theory of Speech Perception: A Critical Review'. *Psychological Review*, 72: 275–309.

Lee, D. N. and Lishman, J. R. (1975). 'Visual Proprioceptive Control of Stance'. *Journal of Human Movement Studies*, 1: 87–95.

Lewald, J. and Guski, R. (2003). 'Cross-Modal Perceptual Integration of Spatially and Temporally Disparate Auditory and Visual Stimuli'. *Cognitive Brain Research*, 16: 468–78.

Liberman, A. (1990). 'Afterthoughts on Modularity and the Motor Theory', in I. G. Mattingly and M. Studdert-Kennedy (eds.), *Modularity and the Motor Theory of Speech Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

—— and Mattingly, I. G. (1985). 'The Motor Theory of Speech Perception Revised'. *Cognition*, 21: 1–36.

——— (1989). 'A Specialization for Speech Perception'. *Science*, new series, 243(4890): 489–94.

McGurk, H. and MacDonald, J. (1976). 'Hearing Lips and Seeing Voices'. *Nature*, 264: 746–8.

MacNeilage, P. F., Rootes, T. P., and Chase, R. A. (1967). 'Speech Production and Perception in a Patient with Severe Impairment of Somethetic Perception and Motor Control'. *Journal of Speech and Hearing Research*, 10: 449–67.

Mann, V. A. and Repp, B. H. (1980). 'Influence of Vocalic Context on Perception of the [sh]–[s] Distinction'. *Perception and Psychophysics*, 28: 213–28.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (1975). 'An Effect of Linguistic Experience: The Discrimination of /r/ and /l/ by Native Speakers of Japanese and English'. *Perception and Psychophysics*, 18: 331–40.

Saldana, H. M. and Rosenblum, L. D. (1993). 'Visual Influences on Auditory Pluck and Bow Judgments'. *Perception and Psychophysics*, 54(3): 406–16.

Sinnot, J. M. and Mosteller K. W. (2001). 'A Comparative Assessment of Speech Sound Discrimination in the Mongolian Gerbil'. *Journal of the Acoustical Society of America*, 110(4): 1729–32.