641

Verticality unparalleled

Ignatius G. Mattingly and Alvin M. Liberman
Haskins Laboratories, New Haven, Conn. 06511

Having long found reason to believe that speech is special, we have, naturally enough, been surprised at the firmness with which others have asserted the contrary – that speech is just like everything else, or, what comes to the same thing, that everything else is special, too. Apparently, our claim has run counter to some deeply held conviction about the nature of mind. One of Fodor's achievements is that he makes this conviction explicit. On the orthodox view, as Fodor sees it, mental activities are "horizontally" organized; arguments for the specialness of speech and language fit better with the assumption that they are vertical. Of the many observations provoked by Fodor's lucid analysis of these opposing views, we can here offer only two. The first has to do with the relations among vertically organized input systems; the second, with the relations between input systems and output systems.

Fodor's input systems, being "domain specific" (p. 47), are in parallel, and their outputs complement each other. Thus, when two modules are sensitive to the same aspects of a signal,

representations from both modules should be cognitively registered. This assumption is surely plausible for those modules, such as for shape and color, that compute complementary representations of the same distal object. But the situation is different for speech. There, the linguistic module appears to take precedence over the module (or modules) that look after distal objects that are not linguistic. Given the same aspect of the signal, the linguistic and the nonlinguistic module are able to compute representations of different distal objects, but if a linguistic representation is computed, the nonlinguistic representation is not cognitively registered. Consider an example to which Fodor himself alludes (p. 49): the transition of the third formant during the release of a consonantal constriction in a consonant-vowel syllable. When artificially isolated from the rest of the signal, this transition is perceived nonlinguistically as a chirp or glissando (Mann & Liberman 1983; Repp, Milburn & Ashkenas 1983). But in its normal acoustic context, the same transition is not so heard. It simply contributes to the perception of a distal object that is distinctly linguistic: the place of articulation of the consonant.

Fodor's account of these facts would be that the isolated transition is ignored by the linguistic module but not by the nonlinguistic module, which registers it cognitively as a chirp. His account would also exclude the possibility that, for the transition in context, the linguistic module would register a chirp as well as a consonant. For the linguistic module, such a representation would be at most "intermediate" (pp. 55ff.) and hence inaccessible to central cognitive processes. (We ourselves doubt that the linguistic module computes any such representation *at all*; we prefer to believe that the earliest representation is an articulatory one.) But the simple parallel arrangement of the modules that Fodor assumes does cause trouble, for it means that "the computational systems that come into play in the perceptual analysis of speech . . . operate *only* upon acoustic signals that are taken to be utterances" (p. 49), but it does not preclude the possibility that other systems will operate on these same signals. It suggests that the transition in context will be registered not only phonetically, by the linguistic module, but also nonphonetically, by the nonlinguistic module. The listener would therefore hear both consonant and chirp. More generally, and more distressingly, the listener would hear all speech signals both as speech and as nonspeech.

What seems called for is a mechanism that would guarantee the precedence of speech but would not constitute a serious weakening of the modularity hypothesis. This precedence mechanism would ensure that, though both the linguistic and the nonlinguistic modules may be active (since speech and nonspeech may occur simultaneously in the world), a signal will be heard as speech if possible and otherwise as nonspeech, but not as both. It is rather compelling evidence for the existence of such a mechanism that it can be defeated under experimental conditions that evade ecological constraints. This is what occurs in the phenomenon known as "duplex perception" (Liberman, Isenberg & Rakerd 1981; Mann & Liberman 1983; Rand 1974). As we have noted, if a third-formant transition that unambiguously fixes the perception of a consonant-vowel syllable (for example, either as /da/ or as /ga/) is extracted and presented in isolation, it sounds like a nonspeech chirp. The remainder of the acoustic pattern, presented in isolation, is perceived as a consonant-vowel syllable, but in the absence of the transition, the place of the consonant is ambiguous. When the transition and the remainder are presented dichotically, a duplex percept results: The chirp is heard at the ear to which the transition is presented, and an unambiguous consonant (/da/ or /ga/, depending on the transition) is heard at the other ear; the ambiguous remainder is not heard (Repp et al. 1983). Thus, the transition is perceived, simultaneously, as a nonspeech chirp and as critical support for the consonant. Apparently, the precedence mechanism recognizes that the transition and the remainder belong together, but it is also aware that there are two signal sources, one at each ear, and that only one of them is speech. It therefore allows both the linguistic module and the nonlinguistic module to register central representations that depend on the formant transition.

How might this precedence mechanism work? An obvious possibility is that it scans the acoustic input and sorts speech signals from nonspeech signals, routing each to its appropriate module. But such a sorting mechanism would seriously compromise the modularity view, because, having to cut across linguistic and nonlinguistic domains, it would be blatantly horizontal. Fortunately for the vertical view, the horizontal compromise appears to be wrong on empirical grounds.

The point is that a sorting mechanism would require that there be surface properties of speech that it could exploit. These properties would be characteristic of speech signals in general, but not of nonspeech signals. Moreover, they would be distinct from those deeper properties that the linguistic module uses to determine phonetic structure. It is of considerable interest, then, that while natural speech signals do have certain surface properties (waveform periodicity, characteristic spectral structure, syllabic rhythm) that such a mechanism might be supposed to exploit (and that manmade devices for speech detection *do* exploit), none of these properties is essential for a signal to be perceived as speech. Natural speech remains speechlike, and even more or less intelligible, under many forms of distortion that destroy these properties (high- and low-pass filtering, infinite peak clipping, rate adjustment). And, more tellingly, quite bizarre methods of synthesis – for example, replacing the formants of a natural utterance by sine waves with the same trajectories (Remez, Rubin, Pisoni & Carrell 1981) – suffice to produce speechlike signals. Thus, speech appears to be speech, not because of any surface properties that mark it as such, but entirely by virtue of properties that are deeply linguistic. A signal is speech if, and only if, the language module can in some degree interpret the signal as the result of phonetically significant vocal-tract gestures. (In the same way, there are no surface properties that distinguish grammatical sentences from ungrammatical ones: a sentence is grammatical if, and only if, a grammatical derivation can be given for it.) We therefore reject this horizontal compromise, and consider two other possible precedence mechanisms, both thoroughly vertical.

The first is an inhibitory precedence mechanism that works across the outputs of the modules in this way: If the linguistic module fails to find phonetic structure, then the output of the nonlinguistic module is fully registered; if, on the other hand, the linguistic module does find phonetic structure, the link to the nonlinguistic module causes the "corresponding" parts of its output to be inhibited but leaves the phonetically irrelevant parts unaffected. Such a mechanism is certainly conceivable and, being a central mechanism, would not compromise modularity. It would, however, be most unparsimonious. For if the inhibitor mechanism were to know which aspect of the output of the nonlinguistic module corresponded to aspects of the signal that were treated as speech by the linguistic module, it would have to know everything that the two modules know: the relationships between phonetic structure and speech signals, as well as the relationship between nonlinguistic objects and nonspeech signals. Thus a central mechanism would, in effect, duplicate mechanisms of two of the modules.

Turning, therefore, to the second possible precedence mechanism, we propose that, while the outputs that the modules provide to central processes are in parallel, their inputs may be in series. That is, one module may filter or otherwise transform the input signal to another module. We suppose that the linguistic module not only tracks the changing configuration of the vocal tract, recovering phonetic structure, but also filters out whatever in the signal is due to this configuration, including, of course, formant transitions. What remains – nonlinguistic aspects of speech such as voice quality, loudness, and pitch, as well as unrelated acoustic signals – is passed on to the non-

linguistic module. This supposition is parsimonious in that it in no way complicates the computations we must attribute to the linguistic module; the information needed to perform the filtering is the same information that is needed to specify the phonetic structure of utterances (and ultimately the rest of their linguistic structure) to central processes.

A further point in favor of this serial precedence mechanism is that something similar appears to be required to explain the operation of other obvious candidates for modularity, such as auditory localization, echo suppression, and binocular vision. Consider just the first of these. The auditory localization module cannot simply be in parallel with other modules that operate on acoustic signals. Not only do we perceive sound sources (whether speech or nonspeech) as localized (with the help of the auditory localization module), but we also *fail* to perceive unsynchronized left- and right-ear images (with other modules). Obviously, the auditory localization module does not merely provide information about sound-source locations to central cognitive processes; it also provides subsequent modules in the series, including the linguistic module, with a set of signals arrayed according to the location of their sources in the auditory field. The information needed to create this array (the difference in time-of-arrival of the various signals at the two ears) is identical to the information needed for localization.

Unfortunately, hypothesizing a serial precedence mechanism does not lead us directly to a full understanding of duplex perception. Until we have carried out some more experiments, we can only suggest that this phenomenon may have something to do with the fact that the linguistic module must not only separate speech from nonspeech, but it must also separate the speech of one speaker from that of another. For the latter purpose, it cannot rely merely on the differences in location of sound sources in the auditory field, since two speakers may occupy the same location; it must necessarily exploit the phonetic coherence within the signal from each speaker and the lack of such coherence between signals from different speakers. It might, in fact, analyze the phonetic information in its input array into one or more coherent patterns without relying on location at all, for under normal ecological conditions, there is no likelihood of coherence across locations. Thus, when a signal that is not in itself speech (the transition) nevertheless coheres phonetically with speech signals from a different location (the remainder of the consonant-vowel syllable), the module is somehow beguiled into using the same information twice, and duplex perception results.

Our second general observation about Fodor's essay is prompted by the fact that language is both an input system and an output system. Fodor devotes most of his attention to input systems and makes only passing mention (p. 42) of such output systems as those that may be supposed to regulate locomotion and manual gestures. He thus has no occasion to reflect on the fact that language is both perceptual and motor. Of course, other modular systems are also in some sense both perceptual and motor, and superficially comparable, therefore, to language: simple reflexes, for example, or the system that automatically adjusts the posture of a diving gannet in accordance with optical information specifying the distance from the surface of the water (Lee & Reddish 1981). But such systems must obviously have separate components for detecting stimuli and initiating responses. It would make no great difference, indeed, if we chose to regard a reflex as an input system hardwired to an output system rather than as a single "input-output" system. What makes language (and perhaps some other animal communication systems also) of special interest is that, while the system has both input and output functions, we would not wish to suppose that there were two language modules, or even that there were separate input and output components within a single module. Assuming nature to have been a good communications engineer, we must rather suppose that there is but one module, within which corresponding input and output operations (parsing and sentence-planning; speech perception and speech production) rely on the same grammar, are computationally similar, and are executed by the same components. Computing logical form, given articulatory movements, and computing articulatory movements, given logical form, must somehow be the same process.

If this is the case, it places a strong constraint on our hypotheses about the nature of these internal operations. All plausible accounts of language input are by no means equally plausible, or even coherent, as an account of language output. The right kind of model would resemble an electrical circuit, for which the same system equation holds no matter where in the circuit we choose to measure "input" and "output" currents.

If the same module can serve both as part of an input system and part of an output system, the difference being merely a matter of transducers, then the distinction between perceptual faculties and motor faculties (the one fence Fodor hasn't knocked down) is perhaps no more fundamental than other "horizontal" distinctions. The fact that a particular module is perceptual, or motor, or both, is purely "syncategoramatic" (p. 15). If so, then the mind is more vertical than even Fodor thinks it is.

ACKNOWLEDGMENT

Support from NICHD Grant HD-01994 is gratefully acknowledged.

NOTE

I. G. Mattingly is also affiliated with the Department of Linguistics, University of Connecticut; Alvin M. Liberman with the Department of Psychology at the University of Connecticut and the Department of Linguistics at Yale University.