BRIEF COMMENTS ON INVARIANCE IN PHONETIC PERCEPTION

A. M. Liberman

According to the instructions of my hosts, I have ten minutes to tell how I see the matter of invariance. So, getting right to the point, I should say that my concern is with invariance only in the conversion from sound to phonetic structure, then move immediately to the facts that such invariance ought, in my view, to take into account.

Because of the way we speak, the acoustic information for a phonetic segment commonly comprises a large number and wide variety of cues, most of them dynamic in form. These cues span a considerable stretch of sound, grossly overlap the cues for other segments, and are subject to a considerable amount of context-conditioned variation.

The phonetic perceiving system is sensitive—one might say exquisitely sensitive—to all the acoustic cues. None of them is truly necessary; all are normally used; and their relative importance bears little relation to their salience as it might be reckoned on a purely auditory basis.

Perception of phonetic structure is immediate in the sense that there is no conscious mediation by, or translation from, an auditory base. This is to say, most generally, that listeners are only aware of the coherent phonetic structure that the cues convey, not of the quite different auditory appearances the cues might be expected to have, given their overlap, context-conditioned variation, number, diversity, and dynamic nature. Thus, taking stop consonants and their dynamic formant-transition cues as a particular example, I note that listeners are not aware of the transitions as pitch glides (or chirps) and also as (support for) a stop consonant; listeners are only aware of the stop. Yet these same formant transitions are perceived as pitch glides (or chirps) when—on the nonspeech side of a duplex percept, for example—they do not figure in perception of a phonetic segment.

These facts have two implications relevant to our concern. One is that the invariance between sound and phonetic structure should be sought in a general relation between the two that is systematic but special, not in particu-

†Also Yale University and University of Connecticut.
Acknowledgment. This work was supported in part by NICHD Grant HD-01994.

[HASKINS LABORATORIES: Status Report on Speech Research SR-77/78 (1984)]
lar connections that are occasional and discrete. The relation we seek can be seen to be systematic to the extent it is governed by lawful dependencies among articulatory movements, vocal-tract shapes, and sounds, dependencies that hold for all phonetically relevant behavior, not just for specific and fixed sets of elements. The relation has got to be special because the vocal tract and its organs are special structures that behave, most obviously in coarticulation, in special ways. A second implication is that the special relation between sound and phonetic structure is acted on in perception by a system that is appropriately specialized for the purpose.

If the foregoing assumptions are correct, then the invariance in speech is not unique. Rather it resembles, at least grossly, the kinds of special invariances that are found in many perceptual domains. Accordingly, the system that is specialized for phonetic perception can be seen as one of a class of similarly specialized biological devices. All take advantage of a systematic but special invariance between the "proximal" stimuli and some property of the "distal" object. The result is immediate perception of that which it is most important to perceive--namely, the properties that make it possible to identify the invariant distal object.

Consider, as an example, visual perception of depth as determined by the proximal cue of binocular disparity. There is a general and systematic, yet special, relation between the distal property (relative distance of points in space) and the proximal stimulus (disparity). The relation is general and systematic in that it is governed by the laws of optical geometry and holds for all points (within its range) and for all objects, not just for some. The relation is special because it depends on the special circumstances that we have two eyes, that they are so positioned (and controlled) as to be able to see the same object, and that they are separated by a particular distance. Neurobiological investigation has revealed an anatomically and physiologically coherent system--a biological "module," if you will--that is specialized to process the proximal disparity and relate it to the distal depth. Given that specialization, perception of depth is automatic and immediate: there is no conscious mediation by, or translation from, the double images we would see if, in fact, we were perceiving the proximal disparity as well as the distal property it specifies.

Other perceptual phenomena have the same general characteristics. Auditory localization and the various constancies come immediately to mind, and, if we put aside questions about phenomenal "immediacy," so too do such processes as those that underlie echolocation in bats and song in birds. These are surely specializations if only because each such process, or module, is as different from every other as is the invariant relation it serves. The phonetic module differs from many of the others in at least two ways.

To make one of the differences clear I would turn again to binocular disparity and depth perception as representative of a large class. In this case the distal object is "out there," a physical thing in the narrow sense of physical, and the invariant relation between its properties and those of the proximal stimulus is determined, as already indicated, by optical geometry and the separation of the eyes. In speech, however, the distal object--a phonetic structure--is a physiological thing, a neural process in the talker's brain, and the invariant relation between its properties and those of the proximal sound is determined in large part by neuromuscular processes internal to the talker but available also to the listener. Thus, the specialized phonetic
module might be expected to incorporate a biologically based link between production and perception. Such a link is not part of the disparity module or of the other perceiving modules it exemplifies, though it may very well characterize the "song center" module of certain birds.

A second important difference in the nature of the invariance (and its module) has to do with the question: What turns the module on? In the case of binocular disparity, the answer is a quite specific characteristic of the proximal stimulus—namely, disparity. Notice, however, that disparity has no other utility for the perceiver but to provide information about the distal property, depth. There are, accordingly, no circumstances in which the perceiver could use the proximal disparity as a specification of, or signal for, some other property. This is to say that disparity and the depth it conveys do not compete with other aspects of visual perception such as hue, form, etc., but rather complement them. Not so in phonetic perception. There is, first of all, the fact that the speech frequencies overlap those of non-speech. More to the point, the formant transitions that we don't want to perceive as chirps when we are listening to speech are very similar to stimuli that we do want to perceive as chirps when we are listening to birds. Thus, almost any single aspect of the proximal stimuli can be used for perception of radically different distal objects: phonetic structures in a talker's head or acoustic events and objects in the outside world. What follows is that the module can hardly be turned on by some specific (acoustic) property of the proximal stimulus. Not surprisingly, then, we find in research on speech perception that the module is, in fact, not turned on that way, but rather by some more global property of the sound. Thus, just as in the perception of phonetic segments all cues are responded to but none is necessary, so too in identifying sound as speech.

How, then, is the module turned on? What invariant property of the sound causes the listener to perceive that the distal object is a phonetic structure and not some nonlinguistic object or event? I offer a suggestion. Suppose that auditory stimuli go everywhere in the nervous system that auditory stimuli can go, including, of course, the language center. Suppose, further, that the language center applies the principle: if the shoe fits, wear it. What is decided, then, by the language center is the answer to the question: could these sounds, taken quite abstractly, have been produced by linguistically significant articulatory maneuvers, also taken quite abstractly? If the answer is yes, then the module takes over the purely phonetic aspects of the percept, and the auditory appearances are inhibited. (Auditory aspects that are irrelevant to the phonetic, such as loudness or hoarseness, are perceived, of course, as attributes of the same distal object.) If the answer is no, then the phonetic module shuts down and the ordinary auditory appearances of the stimulus are perceived. Hence the common experience of those who work with synthetic speech that when the sound includes configurations that the articulatory organs cannot produce, as well as those it can, the percept breaks, correspondingly, into non-speech and speech. Phenomenally, the non-speech stands entirely apart from, and bears no apparent relation to, the speech, even though the acoustic bases for these wholly distinct percepts were perfectly continuous. The same arrangement for turning the module on (or off) might account for the fact that certain kinds of acoustic patterns—for example, sine waves in place of formants—can be perceived as speech or as non-speech depending on circumstances that in no way alter the acoustic structure of the stimulus. It also helps to explain how, as in the unnatural procedures of duplex perception, we can disable the mechanism that forces the choice be-
tween speech and nonspeech, and so create a situation in which exactly the same proximal formant transition is simultaneously perceived (in the same context and by the same brain) as critical support for a stop consonant and also as a nonspeech chirp. At all events, there is a kind of competition between phonetic perception and other ways of perceiving sound. A consequence is that the phonetic module produces a more or less distinct mode of perception in a way that modules like depth perception do not. This phonetic mode accommodates a class of distal objects that are distinguished, not only by their role in language, but also by the special nature of the invariant relation by which they are connected to sound.