The interesting papers we heard all dealt in one way or another with a question that is surely central to an inquiry into the biology of language: Are linguistic processes in some sense special, different from the processes that underlie nonlinguistic activities and, perhaps, unique to man? To discuss that question, and the papers of the evening's session, I find it useful to distinguish two classes of specialized processes, auditory and phonetic.

Specialized auditory processes would serve, perhaps in the fashion of feature detectors, to extract those aspects of the acoustic signal that carry the important information. One is led to suppose that such devices might exist because it is true, and paradoxical, that some of the most important phonetic information is contained in parts of the speech sound that are not physically salient. Thus, a significant acoustic cue is in the formant transitions, though these are often of short duration and rapidly changing frequency. Perhaps there are devices devoted to detecting those transitions. If so, we should hold them up as examples of specializations in the auditory system. They would be important for the perception of language, but not properly part of its special processes.

If the acoustic signal were directly related to the phonetic message, then detection of the phonetically important cues would be sufficient for phonetic perception; no further processing would be necessary. But the relation between signal and message is peculiarly complex. [For summary accounts, see Fant (1962); Cooper (1972); Stevens and House (1972); Liberman (1974); Studdert-Kennedy (1974).] As a result, the specialized auditory detectors can only begin the job; the auditory display they produce must still be interpreted, because the phonetic message is there in such highly encoded form. If there are devices specialized to do that kind of interpreting, then I should consider them phonetic, not auditory. Since I will organize my comments on the papers of the evening in terms of that distinction, I should take a moment to illustrate what I mean.

Consider the formant transitions that are important cues for the perception of stop consonants in syllable-initial position, and call up in your mind's eye spectrographic representations (similar to those shown by Dr. Morse) of such transition cues as would be appropriate for [da] and [ba]. Now add a patch of


†Also University of Connecticut, Storrs, and Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-45/46 (1976)]
fricative noise—the hiss of [sa]—just before the [da]. If that patch is immediately in front of the [da], you will hear [sa], not [da]; the stop will have disappeared completely. But if the patch is moved away so as to leave about 50 msec of silence between the end of the hiss and the beginning of the formant transitions, then you will hear [sta]; that is, you will hear the stop once again. The generalization that captures those facts, and many others closely related to them, is that a necessary condition for the perception of syllable-initial stop consonants is a brief period of silence in front of the appropriate transition cues. But why should silence be necessary? Why should it be impossible to hear the stop when its acoustic cues follow closely on the fricative noise?

The simplest explanation, surely, is that we are here dealing with a characteristic of the generalized mammalian auditory system. That might seem reasonable if only because in putting the fricative noise in front of the transition cues we have conformed to the paradigm for auditory forward masking. But a search of the literature on such masking uncovers no reason to suppose that it could, in fact, provide the account we seek; forward masking does occur, but it is not nearly so strong as to produce the total disappearance of the stop consonant in [sa]. [See, for example, Elliott (1971) and Leshowitz and Cudahy (1973).]

Consider, now, a second interpretation. Suppose there are transition detectors of the kind I speculated about and suppose, further, that the fricative noise disables them, rendering them ineffective in extracting the transition cues for the stop consonant. In fact, there is very indirect evidence that such transition detectors may exist in man. Thus, work by Kay and Matthews (1972) suggests that there may be detectors sensitive to frequency modulations, at least within a certain range. More, and perhaps more indirect, evidence comes from studies on the so-called adaptation-shift phenomenon, first found in speech by Elmas and Corbit (1973) and since studied by a number of investigators. [For a review, see Cooper (1975) and Darwin (in press).] Among those studies is a recent one by Ganong (1975) that I will describe, if only briefly, because its outcome has several implications for our concern with specialized processes: it suggests, as do several other such studies, that transition detectors may exist, but it also indicates that such detectors are in no way disabled by the fricative noise of our example.

Ganong’s experiment went like this. Having first found the boundary between synthetic [da] and [ba], Ganong adapted his subjects with [da] and measured the resulting shift in the [da–ba] boundary. Then he put a patch of fricative noise in front of the [da] and adapted his subjects with the [sa] syllable that they all heard when the fricative-patch-plus [da] was sounded. The effect on the [da–ba] boundary was at least as great as when the adaptation was carried out with [da]. As a control against the possibility that [sa] had its effect because it worked on the same abstract phonetic-feature detector as [da] ([s] and [d] have the same place-of-production feature), Ganong adapted with a [sa] from which the formant transitions had been removed; in that condition the effect on the [da–ba] boundary was much smaller. Those results suggest that the adaptation shift in the [da–ba] boundary was caused by a change in the state of some device that responds to formant transitions; thus, they support the assumption that there are such things as transition detectors.
But Ganong's results also show, more generally, that the transition cues following the fricative noise were getting through in full strength, at least as auditory events. If those transition cues nevertheless failed to produce perception of a stop consonant, it was not because they were absent from the auditory display. [Other kinds of evidence for the same conclusion are reviewed in Liberman (in press).]

We are led, then, to a third explanation for the disappearance of the stop consonant: silence is necessary for the perception of stop consonants, not because it provides time to evade normal auditory forward masking, and not because it prevents the disabling of specialized transition detectors, but because it provides information. The information is that the speaker did indeed make the total closure of the vocal tract necessary to the production of a stop consonant. Thus, given enough silence to indicate a sufficient closure of the vocal tract, a specialized phonetic device could interpret the transition cues as reflecting a linguistic event that included the stop-consonant segment [d]. Hence the perception [sta] when a silent interval of about 50 msec is placed between the end of the hiss and the beginning of the transitions. Without that silent interval the only reasonable phonetic interpretation is that the vocal tract did not close completely. Hence [sa].

So much, then, for the possibility that there are at least two different kinds of devices specialized for speech. Let me now comment on the papers of the evening with reference to that distinction.

In the presentation by Dr. Andrews we saw interesting evidence that baboons change the configuration of their vocal tracts so as to produce something like formant transitions and, further, that such transitions may convey information from one baboon to another. If it is indeed the formant transitions that carry the information, and if the transitions are as brief and rapid as they sometimes are in human speech, then we should not be surprised to find feature detectors specialized to tract them. And in working with baboons we might, of course, expect to get at such devices more directly than we can in research on human beings.

Though baboons may produce and respond to rapid transitions, we have as yet found no reason to believe that they (or, indeed, any creatures other than man) produce or perceive phonetic strings. I should doubt, therefore, that we would find the specialized phonetic processor to which I referred. But what I doubt is surely not important. What is important, I should think, is that we can find out whether baboons do have something like transition detectors and also whether they behave toward speech as if they make a phonetic interpretation. Dr. Andrews has given us a good start in that direction.

The experiments that Philip Morse described are a model of how to learn about the biology of language. To select some interesting characteristic of human speech perception and then look for that characteristic in prelinguistic infants and nonhuman primates is surely one of the best ways to uncover whatever there may be of biological predisposition, specialized process, and species specificity. The experiments are certainly hard to do, but they are very much worth doing, and Dr. Morse does them very well indeed.

The results Dr. Morse told us about this evening were interpreted by him in terms of the possibility that there are devices like transition detectors. In
his view, such devices might explain categorical perception of the place dis-
tinction for stop consonants in infants and the somewhat in-between tendency
toward categorical perception he got in monkeys. I think it quite reasonable to
suppose that the output of such detectors would be categorical. I doubt, how-
ever, that the concept of feature detector could take us very far toward ex-
plaining the perception of stop consonants, except by a kind of metaphorical ex-
tension. Some of the reasons for my doubt will, perhaps, become clearer in con-
nection with the examples I mean to develop when I discuss Dr. Warren's paper in
a few moments, so I will say no more about those reasons now. In fairness to
Dr. Morse, however, I should emphasize that he was not trying to explain the
perception of stop consonants, nor even the perception of the place feature, but
only some data on discrimination and tendencies toward categorical perception in
infants and monkeys.

As for Dr. Morse's experiment, I should say that in using three formants
instead of two he gained the advantage of greater realism but at the cost of
some added difficulty in interpretation of the results. That difficulty arises
because when second- and third-formant transitions are both varied, it is harder
to scale physical similarity and therefore that much harder to assess tendencies
toward categorical perception. If one nevertheless prefers to use the three-
formant patterns because they are closer to what occurs in speech, he might re-
duce the difficulty I referred to by coupling the transition cues with a variety
of vowels, thus randomizing the acoustical similarities; if the discrimination
functions nevertheless come out the same way they did in Dr. Morse's experiment,
the conclusion would be quite compelling.

Still, the results so far obtained with infants are impressive. The in-
fants of Morse's study did show a strong tendency toward categorical perception
of the place distinction in the stops, and, as Morse pointed out, that result
accords with those obtained by other investigators. In the case of the monkeys,
however, it is a good deal less clear that perception of the stops is categor-
ical. There was, in the monkeys of Dr. Morse's experiment, some tendency in
that direction, though less apparently than with the infants. In that connec-
tion, we should keep in mind the results of the earlier study by Sinnott (1974),
to which both Morse and Warren referred. Using reaction time as the measure,
Sinnott found that her monkeys, like those of Morse, discriminated within phon-
etic categories; but they did not discriminate better across phonetic bound-
aries than within them. That is, Sinnott's monkeys did not show any appreciable
tendency toward categorical perception, though her human subjects did.

Since the experiment on discrimination of the voicing distinction by
chinchillas (Kuhl and Miller, 1975) was several times referred to by our speak-
ers, I should also comment on that. It is surely of interest that the chin-
chillas "classified" the speech stimuli so as to put the boundary in much the
same place that human listeners do. Given that the relevant acoustic cue is the
relative time of onset of two parts of the pattern, it is also of interest that
research with nonspeech sounds has found a categorical "notch" in the auditory
system at a relative displacement appropriate to the speech-sound boundary
(Miller, Pastore, Wier, Kelly, and Dooling, 1974). In the case of the voicing
distinction, it may be, therefore, that in the development of language, nature
took advantage of a categorical distinction characteristic of some mammalian
auditory systems, though special adjustments in the articulatory mechanisms
would presumably have been necessary to get them to produce accurately just that
small difference in timing required to put the sounds within the preset (and rather narrow) constraints of the ear.

I nevertheless have several reservations, even about this apparently simple case. Using an expanded range of the same stimuli that were used in the chinchilla experiment, Wilson and Waters (1975) found that variations in stimulus range caused rhesus macaque monkeys to shift their "boundary" from 28 msec, which happens to be about where the chinchilla boundary was, to 66 msec. (They also found some tendency toward categorical perception, wherever the boundary was.) That kind of change, which implies that the monkeys may have been splitting the range, does not occur in human subjects. [See, for example, Sawusch, Fisoni, and Cutting (1974).] The possibility that such a change might occur in chinchillas was not controlled for.

My other reservation arises from the fact that the human boundary is not fixed at either of the boundaries so far found with animals and with nonspeech sounds, but rather varies (together with the categorical notch) from 18 msec to as much as 45 msec as a function of the duration of the transitions and the frequency at which the first formant begins (Stevens and Klatt, 1974; Lisker, Liberman, Erickson, and Dechovitz, 1975). (The variation with duration of the transitions may reflect a normalization for rate of articulation.) I would be interested to know if the chinchilla's boundary moves in the same way. It would also be interesting to know if the chinchilla, or any other animal, appreciates that the voicing distinction is, indeed, the same in those cases in which the relevant acoustic cues are entirely different. What happens, for example, when the distinction is moved from initial position (e.g., [bi] vs. [pi]), which is the kind of distinction so far studied in animals) to intervocalic position following a stressed syllable (e.g., [rasbrd] vs. [raeprd]), where a sufficient cue is the time interval between the two syllables; or to final position (e.g., [raeb] vs. [raep], where a sufficient cue is the duration of the preceding vowel (plus consonant-vowel transition))? To "understand" that such distinctions have something in common despite gross difference in the acoustic cues would constitute an impressive demonstration of phonetic interpretation.

We come now to that part of this evening's program that touched more directly on the matter of specialized phonetic processes. The relevant paper was given by Richard Warren. He reminded us of his earlier experiments—very important experiments, in my view—in which he found that the auditory system does not measure up to one of the requirements of phonetic perception. The requirement is that the order of the phonetic segments be preserved; the word 'bad' is different from the word 'dab.' Now if we measure the rates at which speech is produced and perceived, we find that the durations we can allot to the phonetic segments are often very short. Indeed, those durations can be as little as 50 msec per segment or, for brief periods, even less. But Dr. Warren has found with nonspeech sounds that the ear cannot properly cope with segments of those temporal dimensions. At the very short durations that we can assign to phonetic segments, the ear can discriminate one order of segments from another—that is, it can hear distinctively different patterns—but, as Dr. Warren told us, it is unable to identify the separate components in the order of their occurrence. Now I will not here review or comment on Dr. Warren's solution to this very real problem. I will rather offer an alternative, which is that in perceiving the order of the phonetic segments we need not—and indeed do not—rely on the temporal order of acoustic segments. Indeed, I would argue that even if the ear were able to identify the order of very short-duration acoustic
segments, it could hardly make use of that ability in perceiving speech. That
would be so because the string of phonetic segments is drastically restructured
in the conversion to sound, with the result that segmentation of the sound does
not correspond directly to the segmentation of the message; accordingly, the
segments are not signaled simply by acoustic events in ordered sequence. But,
fortunately for the integrity of the message, information about segment order is
nevertheless conveyed, though by acoustic cues that could be interpreted, I
should think, only by a device that "knows" the secret of the code—that is, by
a phonetic device.

Let us consider, for example, the matter of segment order in the syllables
[ba] and [ab] and see how information about the phonetic structure is carried in
the sound. In producing those syllables, the gestures for the segments [b] and
[a] are not made discretely and in turn. Rather, as we well know, the gestures
are organized into units larger than a segment—something like a syllable, per-
haps—and then coarticulated. If the [ba] and [ab] syllables had been produced
at a moderately high rate of articulation, we should then see for [ba] an
acoustic signal lasting perhaps 70 or 80 msec and containing three formants that
rise from the beginning of the acoustic syllable to the end. For [ab] we should
see the mirror image—that is, three formants that fall. If we search out the
information about [b], we find that it exists not just at the beginning (for
[ba]) or at the end (for [ab]), but throughout the acoustic syllable. Informa-
tion about the vowel is also carried from one end of the sound to the other. It
is as if the coarticulation has effectively folded consonant and vowel into the
same piece of sound. As a result, there is no acoustic criterion by which one
can divide the speech signal into segments corresponding to the segments of
the phonetic message. A further consequence is that the cues for the segments must
necessarily exhibit a great deal of context-conditioned variation: the transi-
tion cues for the consonant, for example, are rising in the one case and falling
in the other. (It should be remarked that when we listen to those transitions
in isolation we hear rising and falling glissandos, just as our knowledge of
auditory psychophysics would lead us to expect.)

To explain how a listener might recover the identity of the segments—that
is, know that there is a consonant [b] and a vowel [a]—we might suppose that
there is a specialized phonetic device that can "hear through" the context-condi-
tioned variation in the acoustic cues and arrive at the canonical forms of
the segments. If so, then that same device could use the same context-condi-
tioned variation to discover the order of the segments: for if the rising pat-
ttern contains a [b], then it could only be a syllable-initial [b]; and if the
falling pattern contains a [b], it could only be a syllable-final [b]. Thus, I
would suppose that perceiving the order of the phonetic segments does not depend
on the ability of the ear to deal with discrete sounds of short duration, but
rather on the operation of a special phonetic device that is able to cope with
the fact that information about order is often encoded in the sound as vari-
ations in acoustic shape. Indeed, I would suppose that such encoding would seem
nicely designed to evade just those limitations of the ear that Dr. Warren's
research has revealed.

I should comment finally on the paper by Philip Lieberman. His work is
especially interesting from my point of view because it offers evidence for a
specialization associated with the production of speech that is, in an important
sense, analogous to the transition detectors of the auditory system. To see the
analogy, we should consider what might have occurred as grammar—hence language—
evolved. The view I want to present has been developed elsewhere (Liberman, 1974), so I will only outline it here.

If, as in an agrammatic system of acoustic communication, the messages were directly linked to sounds, the number of messages we could communicate would be limited to the number of holistically different sounds we can produce and perceive. And that is a relatively small number. But grammar drastically restructures the information in the message, making it appropriate, at the one end, for the great message-generating capabilities of the brain and, at the other, for the relatively limited abilities of the vocal tract and the ear to produce and perceive sounds. Viewed this way, the processes underlying grammar evolved as a kind of interface between two different kinds of structures, adapting the potentialities of the one to the limitations of the other. (My earlier comments on evading the auditory limitations described by Dr. Warren are an example of this kind of grammatical function at the very lowest level of the linguistic system—that is, at the conversion from phonetic message to sound.) But it is also possible that in this evolutionary process the structures being linked by the grammar might themselves have changed. On the perception side of the process an example would be the development of transition detectors in the auditory system to extract just that information which the phonetic (grammatical) system used in carrying out its peculiar function. And on the production side there are the changes in the vocal tract that Dr. Liberman has told us about. Those changes have apparently made the vocal tract less limited for phonetic communication, and so have reduced the mismatch between that organ and the message-generating intellect, a mismatch otherwise taken care of by the grammar. We might suppose that if we had to speak with the vocal tract of a nonhuman primate, the grammatical interface would have to be even more complex than it is.

I think I can justifiably end my comments on a hopeful note. Those of us who care about speech and the biology of language have reason to be encouraged. We now know enough about speech to be able to identify some of its most distinctive characteristics—those characteristics, that is, that most clearly imply the existence of specialized linguistic processes. As a result, we can fruitfully make comparisons with nonlinguistic processes in man and with any processes at all in prelinguistic infants and (presumably) nonlinguistic animals. Indeed, the comparisons are, for obvious reasons, easier to make at the level of speech than at the level of syntax, especially with infants and animals. Moreover, we have started to make those comparisons. But we have only just started. There are hundreds of experiments out there waiting to be done. Until we see what results they produce, we would be well advised, I think, to suspend judgment.

REFERENCES


