117

American Scientist

Reprinted from
May-June 1972
Volume 60
Number 3
pp. 327-337

Speech Cues and Sign Stimuli

Ignatius G. Mattingly

Ignatius G. Mattingly

Speech Cues and Sign Stimuli

An ethological view of speech perception and the origin of language

The perception of the linguistic information in speech, as investigations carried on over the past twenty years have made clear, depends not on a general resemblance between presently and previously heard sounds but on a quite complex system of acoustic cues, which has been called by Liberman et al. (1967) the "speech code." These authors suggest that a special perceptual mechanism is used to detect and decode the speech cues. I wish to draw attention here to some interesting formal parallels between these cues and a well-known class of animal signals, "sign stimuli," described by Lorenz, Tinbergen, and others. These formal parallels suggest some speculations about the original biological function of speech and the related problem of the origin of language.

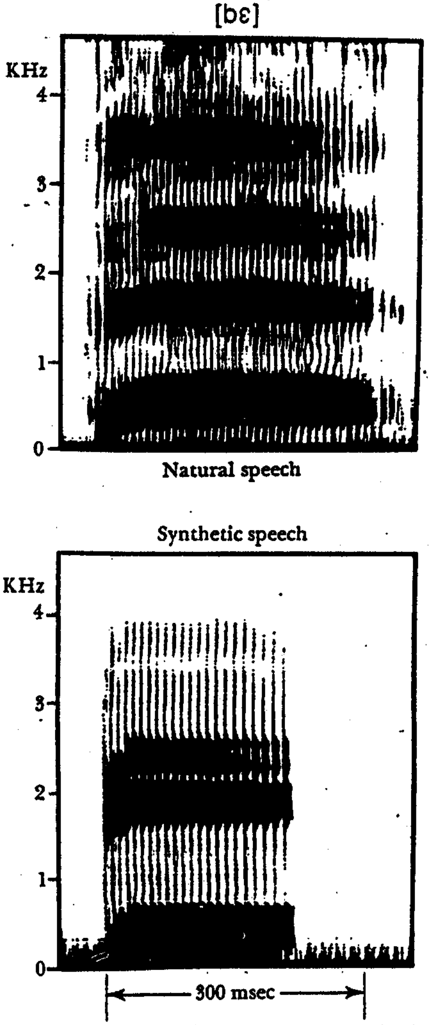

Figure 1. Spectrograms of the syllable [bɛ]: *top*, natural speech; *bottom*, synthetic speech.

A speech cue is a specific event in the acoustic stream of speech which is important for the perception of a phonetic distinction. Speech cues enable us to distinguish the back rounded vowel [u], as in *boot*, from the front unrounded vowel [i], as in *beet*, or the class of speech sounds with labial place of articulation [b, p, m] from the class of sounds with alveolar articulation [d, t, n] or velar articulation [g, k, ŋ].

The cues are most readily observed with the aid of the sound spectrograph. This device produces a spectral display in which the horizontal axis is time, the vertical axis, frequency, and the darkness of the record, intensity (see Fig. 1 and Fig. 2, top). The concentrations of energy that appear in the display as dark bars are "formants," and many of the acoustic cues depend upon the behavior of these formants (we postpone for the moment discussion of their physical basis). In Figure 1, the steady state part of both utterances is heard as [ɛ] as in *bet*. *F*1 (i.e. the first, lowermost formant) has a frequency of about 500 Hz and *F*2, the second formant, a frequency of about 1700 Hz. If *F*1 had been lower and *F*2 considerably higher, the vowel would be [i] as in *beet*. If both *F*1 and *F*2 had been very low, the vowel would have been [u] as in *boot*. The initial formant transitions are also speech cues. The rising transition of *F*1 cues a preceding stop consonant. The rising *F*2 transition is a cue that the place of articulation is labial.

It is possible by electronic means to convert a spectrogram back to speech, with little loss of intelligibility, by means of an optical scanning device called the Pattern Playback (Cooper et al. 1952). The Playback also permits one to convert hand-drawn spectrographic patterns into synthetic speech. Figure 2 shows a spectrogram of the natural utterance *to catch pink salmon* and, below it, the hand-drawn

Ignatius G. Mattingly, Associate Professor of Linguistics at the University of Connecticut and a member of the research staff of Haskins Laboratories, received his B.A. and Ph.D. from Yale and an M.A. from Harvard. For the academic year 1970–71, during which this article was written, he was a Fulbright visitor and a Fellow of King's College, Cambridge. He has published a number of papers about the reading process, speech perception, and speech synthesis, and has written a computer program that uses speech cues like those discussed in the present article to produce synthetic speech. He is deeply interested in the implications of ethological research for human perception.

The preparation of this paper has been supported by The National Institute of Child Health and Human Development, the Office of Naval Research, the Veterans Administration, the Fulbright-Hays Commission, and the Provost and Fellows of King's College, Cambridge. The figures from Tinbergen's The Study of Instinct *are reproduced with the kind permission of the author and the Clarendon Press, and the quotation from Eric Lenneberg's* Biological Foundations of Language *is used with permission of the author. Dr. Mattingly also wishes to acknowledge encouragement, discussion, and criticism given by G. Evelyn Hutchinson, Alvin M. Liberman, Charles Liberman, Benjamin Sachs, Jacqueline Sachs, Michael Studdert-Kennedy, Emily Kirstein, Philip Lieberman, Alison Jolly, Mark Haggard, Adrian Fourcin, and Dennis Fry.*

Address: Haskins Laboratories, 270 Crown St., New Haven, CT 06511.

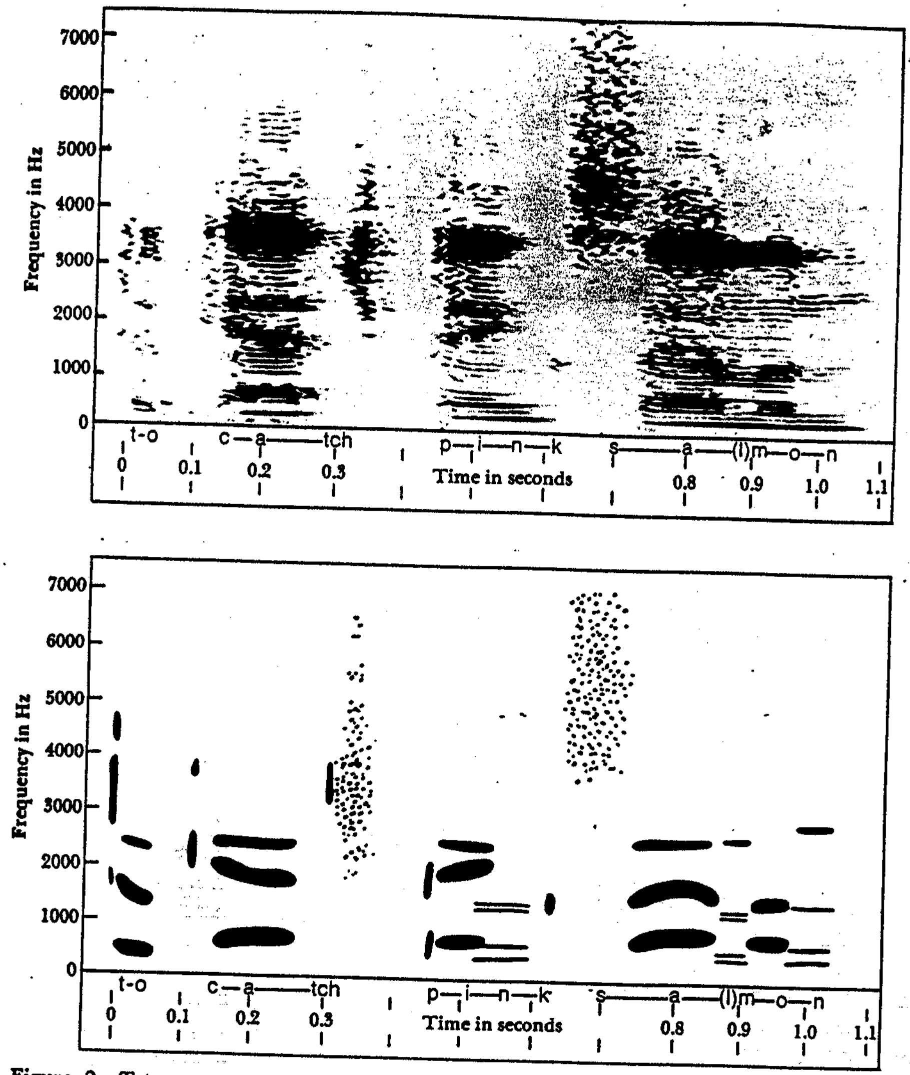

Figure 2. *Top,* spectrogram of the natural utterance *to catch pink salmon; bottom,* Playback spectrographic pattern for this utterance.

spectrographic pattern from which an intelligible artificial version of the original was synthesized. Many of the acoustic cues we have been discussing were discovered by perceptual experiments for which the stimuli were made on the Playback.

Some recent work indicates that human beings may possibly be born with knowledge of speech cues. While appropriate investigations have not yet been carried out for most of the cues, the facts are rather suggestive with respect to "voice-onset time," a major cue to the voiced-voiceless distinction. Voice-onset time refers to the timing of the onset of phonation relative to the release of the stop closure. This distinction can readily be observed by comparing the spectrograms of two synthetic syllables, [dɑ] and [tɑ], in Figure 3. During the 60 msec transition period following the release of the two stops, [dɑ] has voicing, represented by the vertical striations in the spectrogram, in all three formants, and an $F1$ transition rising from zero to the frequency of [ɑ], while [tɑ] has noisy rather than voiced $F2$ and $F3$ and no observable $F1$ transition. Not all languages distinguish between stops with immediate voice onset and stops with voice onset delayed after release, but for all those that do, the amount of delay in voice onset required for a stop to be heard as voiceless rather than voiced is about the same, 30–40 msec (Lisker and Abramson 1970; Abramson and Lisker 1970). This constraint on perception thus appears to be a true language universal, and thus likely to reflect a physiological limitation rather than a learned convention.

To explore the question more directly, Eimas et al. (1970), by monitoring changes in the sucking rate of one-month-old infants listening to synthetic speech stimuli, showed that the infants could distinguish significantly better between two stop-vowel stimuli which straddle the critical value of voice-onset time than between two stimuli which do not, even though the absolute difference in voice-onset time is the same. Thus the information required to interpret at least one speech cue appears either to be learned with incredible speed or to be genetically transmitted.

Sign stimuli and social releasers

Sign stimuli, with which I propose to compare speech cues, have been defined by Russell (1943), Tinbergen (1951), and other ethologists as simple, conspicuous, and specific characters of a display which under given conditions produces an "instinctive" response: the red belly of the male stickleback, which provokes a rival to attack, or the zigzag pattern of his dance, which arouses the female (Tinbergen 1951); the spots by which the ringed plover identifies her eggs (Koehler and Zagarus 1937); the red spot on the herring gull's bill which makes her chicks beg for food (Tinbergen 1951). These examples are visual, but sign stimuli are found in other modalities also: e.g. the monotone note of the white-throated sparrow's song by which he asserts his territorial claims (Falls 1969), or the chemical in the blood from a wounded minnow which causes other minnows to flee when they scent it in the water (Manning 1967). Responding properly to sign stimuli is normally of great value for the survival of the individual or the species. As Manning (1967:39) comments, "Sign stimuli will usually be involved where it is important never to miss making a response to the stimulus." It is this circumstance, perhaps, which accounts for the striking properties of sign stimulus perception which will mainly concern us here: the animal responds not to the display in general but specifically to the sign stimuli, and the strength of the response is in proportion to the number and conspicuousness of the sign stimuli. The perception of a sign stimulus and the response it produces have been attributed by Lorenz (1935) to a special neural "innate releasing mechanism."

The concepts of the sign stimulus and the innate releasing mechanism, as used in early ethological work, have come in for much justified criticism (e.g. Hailman 1969; Hinde 1970). It has been argued that sign stimuli cannot be shown to differ in principle from other stimuli; that some purported sign stimuli are not actually specific to particular responses but merely reflect the general capabilities of the animal's sense organs or associated perceptual equipment; that the word "innate" suggests too simple a dichotomy between nature and nurture; and that sign stimuli do not always lead to direct and immediate responses but influence behavior in other ways.

But when all these criticisms are taken into account, there remain some very striking phenomena. There are many cases in which a stimulus is selectively perceived by a particular species and not by others. The selectivity cannot be accounted for simply by an appeal to the general sensory capabilities of the species. The stimulus consistently elicits a direct response (or other specific behavior indicating that the stimulus has been perceived, as in the case of orientation). This response is adaptive. Moreover, in many instances (and in all the examples given above) the stimulus is a character of a display by a conspecific (or symbiotically related) individual; the entire pattern of behavior, consisting of the display and the response, is adaptive.

Displays of this latter sort have been called "social releasers" (Tinbergen 1951:171). Their component sign stimuli elicit appropriate responses from conspecific individuals in situations important for group safety or for the integrity and continuity of the species. Social releasers include alarm calls; "threat behavior" of many species, by which the adaptive ends of sexual fighting are achieved with few actual casualties; the displays which serve as reproductive isolating mechanisms, encouraging intraspecific and discouraging interspecific mating; and the signs by which parents and young identify each other, so that the latter are protected and fed. In all these adaptively important situations, displays composed of sign stimuli serve to authenticate the conspecificity of individuals.

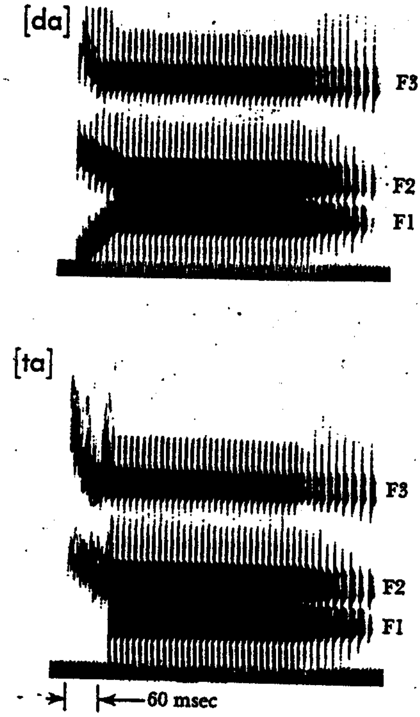

Figure 3. Spectrograms of synthetic [dɑ] and [tɑ]. In [tɑ], the *F*1 transition is not visible and the *F*2 and *F*3 transitions are noisy.

Parallels with speech

It has also been suggested that sign stimuli actually occur in human behavior. The facial characteristics and limb movements of babies evoke parental behavior (Tinbergen 1951). Babies, in turn, respond to adult facial characteristics, notably to eyes and to smiles, and women have a universal flirting gesture (Eibl-Eibesfeldt 1970). I think that speech cues may also belong to the class of human sign stimuli, despite some obvious differences, to be discussed shortly. But let us now consider the resemblances.

First of all, the speech cues, like the sign stimuli, do not require a natural context, or even a naturalistic one; the appropriate response can be elicited by drastically simplified models of the natural original. Tinbergen's sticklebacks would respond to an extremely crude model provided only that it had a red belly, but would disdain very naturalistic models which lacked this crucial feature (Fig. 4) (Tinbergen 1951:28). Lorenz (1954:-291, translated by Eibl-Eibesfeldt 1970:88), makes the general claim that "where an animal can be 'tricked' into responding to simple models, we have a response by an innate releasing mechanism." In the case of speech, most of the complexity of the spectrum can similarly be dispensed with. The natural and synthetic utterances in Figure 1 are linguistically equivalent, even though only the lower formants appear in the latter, and these in a very stylized configuration. The synthetic utterances in Figure 3 are clearly heard as [dɑ] and [tɑ]. The point is even more obvious in Figure 2, in which a hand-drawn Playback pattern can be directly compared with the natural utterance on which it was based.

The Playback pattern is not, however, simply an acoustic cartoon of the natural utterance. Though it shares with a cartoon the appearance of extreme simplicity and emphasis of salient features, it is rather a systematic attempt to represent, consistently but exclusively, the essential acoustic cues, all other details of the signal being discarded or neutralized. The principal loss in such synthetic speech is not intelligibility but only naturalness. This is rather surprising; one might reasonably expect that intelligibility would depend crucially on naturalness, that tampering with the observed spectrum of a natural utterance to any degree would alter its linguistic value, or cause it not to be perceived linguistically at all. I do not mean to imply that high-quality natural speech would not be more intelligible than synthetic speech, or that sticklebacks would not respond more strongly to a real stickleback with a red belly than to a dummy. In synthetic speech, a host of redundant minor cues, as yet unidentified, are no doubt sacrificed together with the linguistically irrelevant details of the signal.

In a similar manner, in the construction of the dummy, sign stimuli of minor importance have been ignored. But it appears that the dependence of artificial speech cues and sign stimuli on a naturalistic context is very small. Though the listener, and (for all we know) the stickleback, may be quite aware of the lack of naturalness, neither one appears to be disturbed by it. The relative naturalness of the speech cues and sign stimuli themselves is something else again, as will be seen shortly.

Both speech cues and sign stimuli exhibit what Tinbergen (1951:81), translating Seitz (1940), calls "the phenomenon of heterogeneous summation." That is, the same response can be elicited by separate and noninteracting sign stimuli: thus, either the

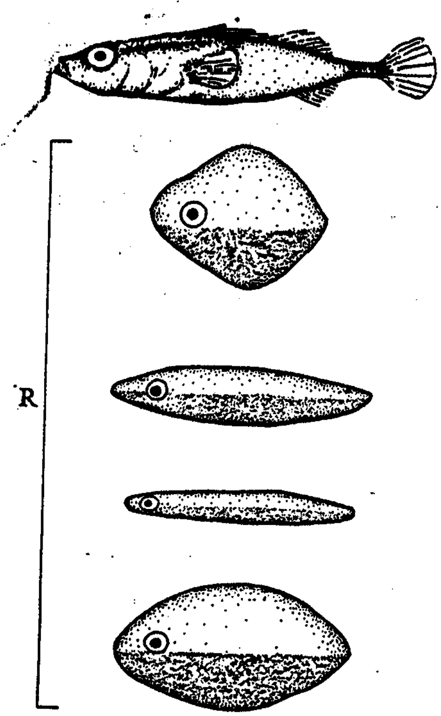

Figure 4. Stickleback models used by Tinbergen. The fairly realistic model, *N*, which lacked a red belly, provoked attack by male sticklebacks much less than the various crude models, *R*, which had red bellies (after Tinbergen 1951).

redness of the patch on the herring gull's bill or the contrast of the patch with the rest of the bill releases the chick's pecking response. Moreover, if two stimuli for the same response are present but one is defective, the second will compensate for the deficiency of the first.

A similar principle operates in speech perception. Multiple cues for the same phonetic feature are the rule. For example, place of articulation in stop consonants is cued not only by the *F*2 transition but also by the *F*3 transition and by a burst of noise at an appropriate frequency just after release of stop closure (Delattre et al. 1955; Halle et al. 1957; Harris et al. 1958). In the medial position, a voiced rather than a voiceless stop is cued by low-frequency periodic energy during closure, by lesser duration of closure, and by greater length of the preceding vowel (Lisker 1957). Furthermore, the perceptual weight of one cue appears to be independent of that of the others; all combine additively to carry a single phonetic distinction. If a cue is defective or absent, as is often the case in natural speech, the deficiency is compensated for by the presence of other cues.

Thus Hoffman (1958) compared perception of place of articulation for (1) synthetic stop-vowel syllables in which all three cues (burst, *F*2 transition, *F*3 transition) were present, (2) syllables lacking the burst cue, (3) syllables lacking the third formant and its transition, and (4) syllables lacking both the third formant and the burst, with only the *F*2 transition present. He found that the optimal version of a cue for a particular place of articulation is the same whether presented separately or in combination with other cues; that labeling is most consistent when all three cues are optimal for the same point of articulation; and that an optimal *F*3 transition would compensate for a nonoptimal burst cue, and conversely. A. M. Liberman (personal communication) points out that speech also carries multiple cues to the sex of the speaker: men's voices differ from women's both in pitch range and in formant frequency range. Thus, neither the perception of speech cues nor that of sign stimuli is a Gestalt (Hinde 1970).

An optimal speech cue is often not a realistic one; such a cue is the analog of a "supernormal" sign stimulus such as the pattern of black spots on a white background on the artificial egg (see Fig. 5), which the plover prefers to a natural egg with dark brown spots on a light brown background (Koehler and Zagarus 1937). "The natural situation," Tinbergen (1951:44) observes, "is not always optimal."

Similarly, if a human subject is asked to label randomly ordered syllables from the series schematized in Figure 6, in which the *F*2 transition is the variable and *F*1 and the steady state of *F*2 are held constant, he will hear the first few, those with rising transitions, as [bæ], the next few as [dæ], and the last few, those with falling transitions, as [gæ]. The stimuli with the less steeply sloping transitions are closer to what one observes in occurrences of [bæ] or [gæ] in natural speech, while the more extreme transitions are unlikely, perhaps even impossible to articulate. Yet in a labeling test, the more steeply rising the *F*2 transition, the more likely the subject is to hear [bæ], and the more steeply falling the transition, the more likely he is to hear [gæ]. Thus the subject will label more consistently not only when more cues are present but also when the cues that are present are more nearly optimal, i.e. supernormal.

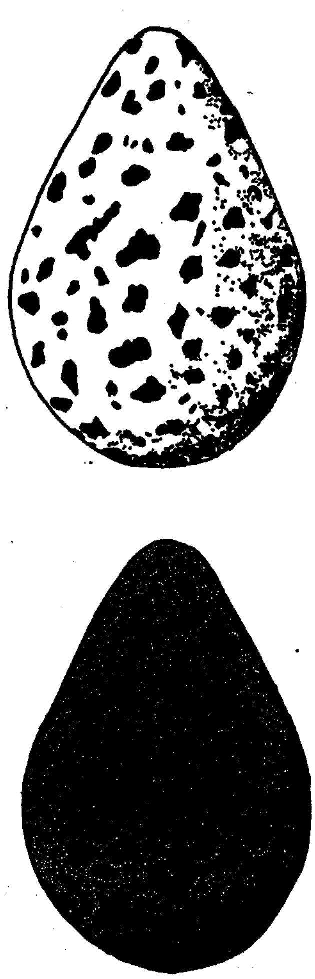

Figure 5. The supernormal plover egg (*top*) with black spots on a white background, preferred by the plover to the normal egg (*bottom*) with dark brown spots on a light brown background (after Koehler and Zagarus 1937, reproduced in Tinbergen 1951).

Again, vowels spoken in isolation will occupy more extreme positions on the *F*1–*F*2 plane than vowels in connected speech (Shearme and Holmes 1962), and they are easier to label than the "same" vowels excised from connected speech. As Manning (1967) says, the failure of a sign stimulus to evolve to the supernormal extreme can usually be explained by considering other functional requirements. Thus the low-contrast, brown-on-brown spotting of the plover's eggs also serves to camouflage them from predators; black on a white background would not be so effective. The vocal tract, likewise, is primarily a group of devices for breathing and eating. A vocal tract that produced supernormal formant transitions and extreme vowels at normal speech rates would probably be unable to perform the primary functions properly.

What is more interesting, as Manning

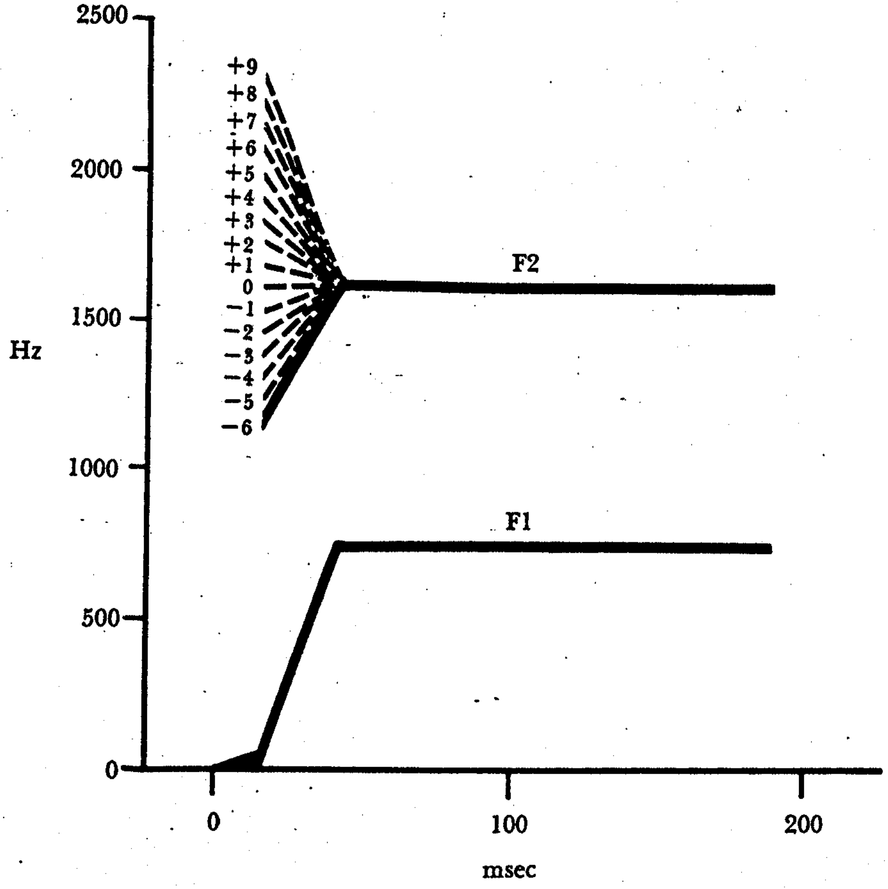

Figure 6. Schema for a series of stop-vowel syllables varying only in *F*2 transition. *F*2 steady state and *F*1 transition and steady state remain constant. Depending on the *F*2 transition, the stop will be heard as [b], [d], or [g] as the *F*2 transition changes from −6 to +9.

goes on to point out, is that the tendency to respond to the sign stimulus has not evolved so as to be perfectly adjusted to the naturally occurring form of the stimulus. Like heterogeneous summation, this must reflect a characteristic of the process by which sign stimuli are perceived, and speech perception must share this characteristic. When we listen to natural speech, presumably we respond best to that combination of cues which approaches the supernormal ideal most closely. Thorpe (1961:98), similarly, has observed that the best natural sign stimulus display is the one which "can come nearest to the supernormal for the largest number of constituent sign stimuli."

Finally, since the validity of the concept of a specialized neural mechanism to account for the selective perception of and response to sign stimuli is in dispute, the possibility that some such mechanism operates in speech perception is of special interest. The properties speech perception has in common with perception of sign stimuli point in this direction, for they are not characteristic of human auditory perception in general; so does the possibility of genetic transmission of knowledge of the cues.

There is also some other supporting evidence. If we ask a subject first to label randomly ordered synthetic stop-vowel syllables from the schematized series shown in Figure 6, and then to discriminate between pairs of syllables that are adjacent in this series, he will do very well near the boundaries implied by the cross-over points of his [bæ] and [dæ] labeling functions and of his [dæ] and [gæ] labeling functions and very poorly elsewhere.

The upper part of Figure 7 shows the labeling functions of a typical subject; the lower part shows his discrimination function for the syllables (solid line). He is discriminating categorically (Liberman 1957), a kind of discrimination that is quite unusual in psychoacoustic tasks. If we now give the subject a similar discrimination task, in which the stimuli are "chirps," i.e. *F*2 transitions in isolation, without *F*1 or the steady-state portion of *F*2 (see Fig. 8), his discrimination function, represented by the dashed line in the lower part of Figure 7, is quite different. He discriminates better than random for most of the series, but the peaks of the syllable discrimination function are absent. There is no indication of categorical perception, and the function is more typically psychoacoustic (Mattingly et al. 1971).

Additional evidence for a special mechanism comes from experiments in dichotic presentation of speech sounds. If different stop-vowel syllables are simultaneously presented to a subject's two ears, he will be able to report correctly the stimuli presented to the right ear more often than the stimuli presented to the left ear. The effect is attributed to the processing of speech in the left cerebral hemisphere (Kimura 1961; Studdert-Kennedy and Shankweiler 1970). No such right-ear advantage is found with nonspeech signals, such as musical tones (Kimura 1964). Experiments by Conrad (1964), Wickelgren (1966), and others suggest that the speech perception mechanism is somehow involved with, and may in fact include, "short-term memory."

To recapitulate, speech cues have a number of perceptual properties in common with sign stimuli. Their perception does not require a naturalistic context, they obey the law of heterogeneous summation, they are more effective as they approach a supernormal ideal, and there is reason to suppose that a special neural mechanism is involved. Some of these formal properties appear in other situations—heterogeneous summation is a property of human binocular vision, for instance—but it is their co-occurrence in both speech and sign stimuli that I find compelling. These properties are shared by the sign stimulus systems of many species, presumably for functional rather than for phylogenetic reasons. Thus, we are led to ask whether speech is in some way functionally similar to a sign stimulus system. But before considering this point, we ought to mention certain rather obvious differences between sign stimuli and the speech cues.

Differences from speech

First, the speech cues are transmitted at a rate much higher than the sign stimuli of any animal system. The displays in which sign stimuli occur, if not virtually static, are either relatively slow-moving or highly repetitive. But the acoustic events of speech that serve as cues occur extremely rapidly. The speech-perceiving mechanism not only keeps up with these events but is capable, as experiments with speeded speech have demonstrated, of speeds more than three times greater than normal speaking rates (Orr et al. 1965). A further gain in transmission speed is obtained by "parallel processing": the speaker produces and the listener extracts cues for different phonetic distinctions more or less simultaneously from the same acoustic activity (Liberman et al. 1967). Thus in a consonant-vowel syllable, the slope of the transition will carry information about the place of articulation of a consonant, its manner class (stop, fricative, semivowel), and the quality of the vowel, while the excitation of these same transitions will cue the voicing distinction. The information rate of speech can be as high as 150 bits per second, and the question of the adaptive value of such a high rate arises.

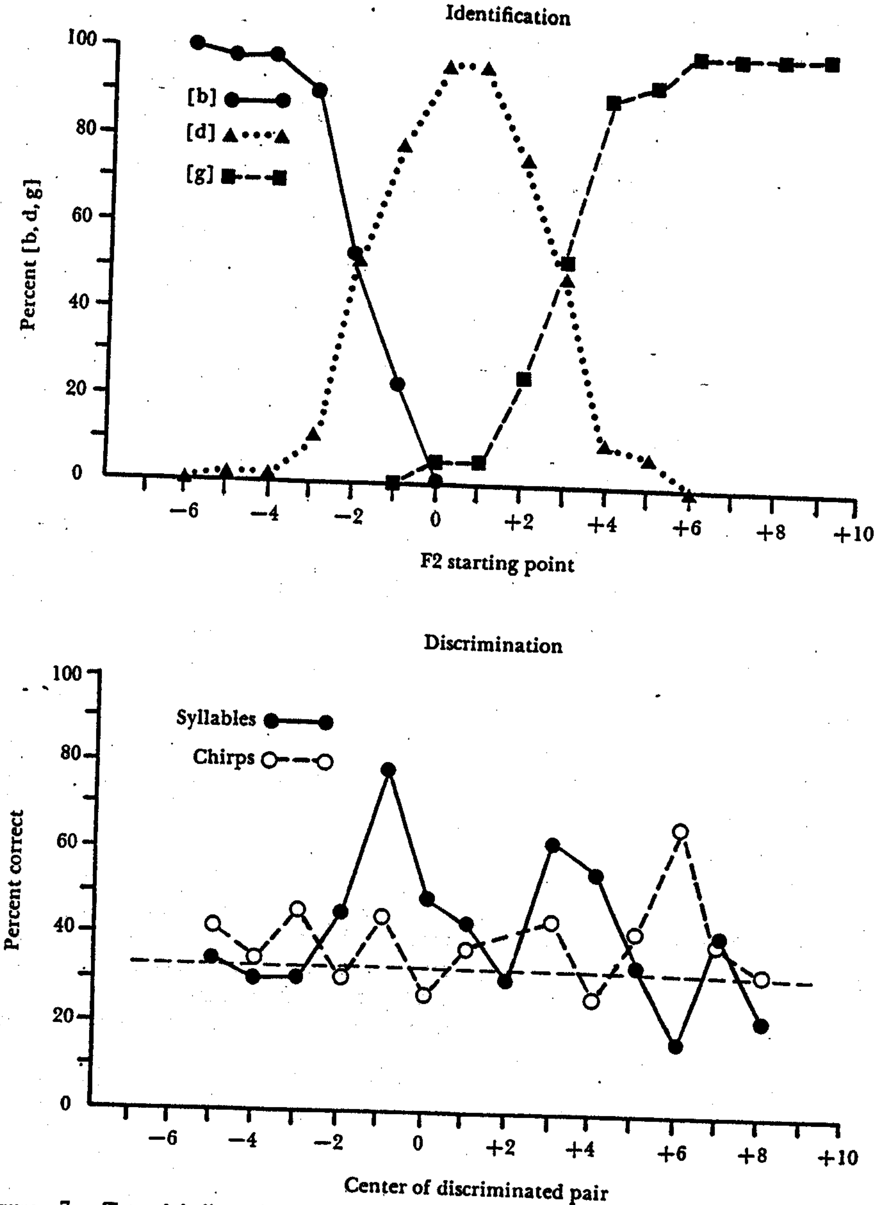

Figure 7. *Top,* labeling (identification) functions for one subject for the series of stimuli in Figure 6. *Bottom,* discrimination functions (solid line) for this subject show characteristic peaks corresponding to cross-over points of labeling functions. Dashed line is the discrimination function for the stimuli of Figure 8.

Another difference between speech cues and sign stimuli is implicit in our use thus far of such conventional phonetic terminology as "place of articulation." Natural speech is produced by exciting the resonant frequencies of the pharyngeal, the oral, and sometimes the nasal cavities of the vocal tract, either by vibration of the vocal cords or by noise from various sources. The changing frequencies of the formant bars of the spectrogram represent the changing resonances of the vocal tract as the various articulators move. Thus, *F*1 depends upon the openness of the major constriction of the vocal tract ("manner"). *F*2 depends upon the location of this constriction ("place of articulation"). The occurrence and timing of vibration of the vocal cords underlie the "voicing" distinction. But apart from these facts about speech production, there is reason to believe that perception of phonetic distinctions is essentially articulatory rather than acoustic. One indication of the articulatory reference of the cues is that a series of stimuli may be perceived as belonging to the same phonetic category, even if they are not neighbors on an acoustic continuum, though they must not fail to be close together on some articulatory continuum.

Thus the spectrographic patterns for [di] and [du] in Figure 9 are heard as beginning with [d], despite the difference in the slopes of the second formant transition. This seems less odd when we look at a series of [d]'s before an articulatory and acoustic vowel continuum beginning with a high front unrounded vowel [i] and ending with low back-rounded [u], schematically represented in the upper part of Figure 10. The second formant appears to have an imaginary origin or "locus" at 1800 Hz (Delattre et al. 1955). It is as if the listener, given the later part of the *F*2 transition, could extrapolate backward through an imaginary earlier part (missing because of the stop closure) to a frequency characterizing the place of articulation of [d]. But the ability to make such an extrapolation implies tacit knowledge of the way the vocal tract works, a knowledge so compelling that the acoustic differences between the [d] in [di] and the [d] in [du] go unperceived. The articulatorily mediated relationship makes it impossible to hear the difference between the two.

Still, the series of syllables beginning with [d] in Figure 10 is an acoustic as well as an articulatory continuum, and it could be argued that the "locus" has only an acoustic reality. But in the case of [k, g] the acoustic continuum is incomplete, because the concept of the locus fails to apply consistently; the locus for [k, g] with low back vowels appears to be much lower and less clearly specifiable than for high front vowels (lower portion of Fig. 10). Yet the perception is constant because the articulation is similar (Liberman 1957). Conversely, the series of stimuli schematized in Figure 6, which does form an acoustic continuum, divides into [b, d, g] because the articulatory reference changes abruptly at two points on the continuum.

Furthermore, a view which refers speech perception to articulation is parsimonious in two distinct ways. It makes use of knowledge that the hearer needs to know in any case if he himself is to be a speaker, and it relates, as correlates of a single articulatory gesture, an apparently unrelated cluster of acoustic events. The cues for, say, the alveolar sounds [t, d]—a high-frequency burst, an *F*2 transition which has a locus at about 1800 Hz, and an *F*3 transition with a locus at 3200 Hz—seem like a highly arbitrary selection if they are regarded as purely acoustic events. Moreover, the events do not occur synchronously; and, as we have just noted, they are interspersed with cues for other phonetic distinctions. But if these same events are interpreted as acoustic correlates of the simple articulatory gesture which produces [t, d], both the selection of events themselves and their relative timing appears quite straightforward. Because of such considerations, it seems reasonable to regard speech as an acoustic encoding of articulatory gestures—or rather of the motor commands underlying those gestures (Lisker et al. 1962; Liberman et al. 1963; Studdert-Kennedy et al. 1970). We may call the sequence of motor commands which determines the speaker's output the "phonetic representation." The listener, because of his intuitive knowledge of the speech code, can recover this representation.

The view that we have just summarized has been called "the motor theory of speech perception" (Lisker et al. 1962; Liberman et al. 1967). So far as we know, no parallel for this kind of indirect, encoded perception has been suggested for any animal other than man. If we accept the motor theory we are obliged to explain what adaptive function it has, how the speech code could have evolved, and why the code is based on articulatory gestures rather than, for example, the size and shape of the cavities of the vocal tract.

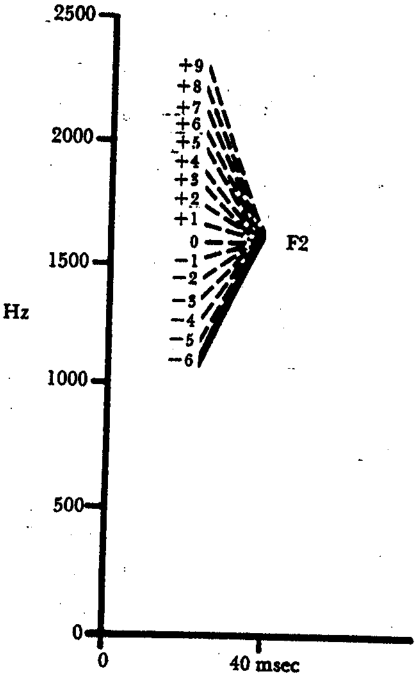

Figure 8. Schema for a series of "chirps" (isolated *F*2 transitions). The discrimination function for chirps (see Fig. 7) is very different from that for stop-vowel syllables (see Fig. 6) for which *F*2 transition is the only variable.

The most notable difference, however, between speech cues and sign stimuli is that, while sign stimuli typically produce a stereotyped behavioral response, speech cues do not. The reason the response to speech is not stereotyped is of course that, unlike sign stimulus displays, a phonetic representation has no fixed significance apart from the linguistic system in which it functions; in itself it is a meaningless pattern, related only quite indirectly to the semantic values of the speakers and hearers. Speech does not stand by itself; it functions as part of language. The meaning of an utterance and the nature of the ultimate behavioral response depend not just on the characters of the stimulus, the environmental context, and the internal state of the perceiver but also upon something not found in conjunction with any set of sign stimuli—a grammar. By virtue of a system of grammatical rules, shared by speaker and hearer, the speaker can evoke not just a few stereotyped responses but a wide variety, many of which are delayed or covert, and, in principle, an infinite range of semantic values can be expressed. The problem is to explain why and how such a powerful system should have evolved.

Grammar and evolution

It is with this problem that most attempts to find precedents for human language in animal behavior have begun. The cries of animals grossly resembling man, as well as animal communications systems which transmit a substantial amount of information, even though the physical nature of the signals may be very different from human speech, have been scrutinized by many investigators for linguistic properties. These efforts have consistently failed. The properties treated as linguistic by some investigators have been so abstract—for example, the Hockett-Altmann "design features" (Altmann 1967; Hockett and Altmann 1968)—that those characteristics which distinguish language from purposive behavior in general are lost to view (Chomsky 1968:60), and really fundamental features are placed on a level with trivial ones.

Thus Hockett's design feature 3, "rapid fading," a property shared by all acoustic phenomena, is apparently just as important as design feature 13, "duality of patterning," which, as we shall see, is truly significant. It is perhaps noteworthy that, according to Hockett and Altmann, the stickleback's communication system, which is of great interest from the viewpoint adopted here, lacks most of the linguistic design features.

Other investigators have tried indiscriminately to force the phenomena of animal behavior into standard linguistic categories. In Lenneberg's (1967:228) words, they have attempted

> to count the number of words in the language of gibbons, to look for phonemes in the vocalizations of monkeys or songs of birds, or to collect the morphemes in the communication systems of bees and ants. In many other instances no such explicit endeavors are stated, but the underlying faith appears to be the same since much time

and effort is spent in teaching parrots, dolphins, or chimpanzee infants to speak English.

Such efforts, I think, are doomed to failure, and those who have insisted most strongly on the biological basis of language—Chomsky and Lenneberg—share this view. Chomsky (1968:62) suggests that human language "is an example of true 'emergence'—the appearance of a qualitatively different phenomenon at a specific stage of complexity of organization." Lenneberg (1967) believes that language has for the most part evolved covertly. In his view, we cannot expect that the steps in the evolution of a characteristic *A* from some quite different characteristic *B* will necessarily be manifest. The nature of the process of genetic modification is such that the intervening steps must in many cases remain obscure. This, he suggests, is the case with human language. While Lenneberg's general position on the nature of evolution may well be essentially correct, to take refuge in this position in the case of a particular evolutionary problem, such as the origin of human language, is essentially to abandon the problem.

Even if precedents for grammar existed in animal communication, it would be very difficult to learn about them. Most of what we know of the grammatical aspects of human language we know not from observations of human behavior but by virtue of our special status as members of the human species. The work of the linguist depends on the availability to him of the intuitions of speakers of a language that certain utterances are, or are not, grammatical. A member of another species, however intelligent, would find it difficult to deduce the most elementary grammatical concepts by observing and manipulating behavior: he would have, somehow, to consult the grammatical intuitions of a human speaker. We are similarly at a loss when speculating about the possible grammars of animal communication systems. Despite the lack of precedents for grammar, I think that Chomsky and Lenneberg are perhaps unduly pessimistic, and that the parallels between the speech cues and the sign stimuli suggest some interesting speculations about the origins of language.

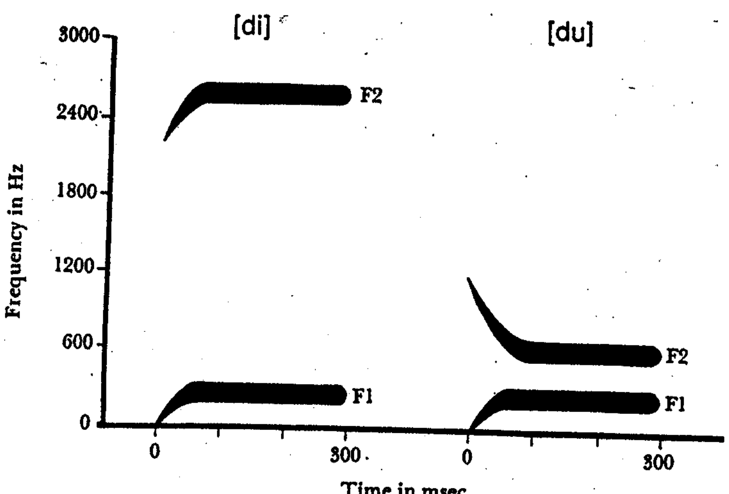

Figure 9. Two-formant spectrographic patterns for [di] and [du]. The *F*1 transition is the same in both patterns, but the two *F*2 transitions are quite different, though both are cues for [d].

One of the traditional explanations of language is that it developed from cries of anger, pain, and pleasure (see, e.g., Rousseau 1755). The difficulty with this explanation is that it does not attempt to account for the transition from cries to names, or for the emergence of grammar. But let us put these problems to one side for the moment, and postulate, just as the traditional explanation does, a stage in man's evolution when speech existed independently of language. Such speech, we suppose, had no syntax or semantics. But it was more than just expressive, because it had phonetic structure, though early man's phonetic repertoire was doubtless more limited than that of modern man (Lieberman and Crelin 1971). His utterances were phonetic representations encoded by acoustic cues. If we ask what function such prelinguistic but structured speech could have had, the parallels we have discussed between speech cues and sign stimuli suggest a possible answer.

Since speech is intraspecific, we suggest that it may have been, at this stage of evolution, a social releaser. If this speculation is correct, prelinguistic speech may have served early man as a vehicle for threat behavior, as a reproductive isolating mechanism, and as a means for mutual recognition of human parents and offspring. By means of phonetic representations underlying his utterances, man elicited appropriate behavioral responses from his fellows in each of these crucial situations. It is probably pointless to speculate as to what particular phonetic representations evoked what responses, but it perhaps reflects the primitive function we have attributed to speech that, while the segmental aspects of speech have been adapted for linguistic purposes, the prosodic features remain as a primary means of physically harmless fighting, of courting, and of demonstrating and responding to parental affection.

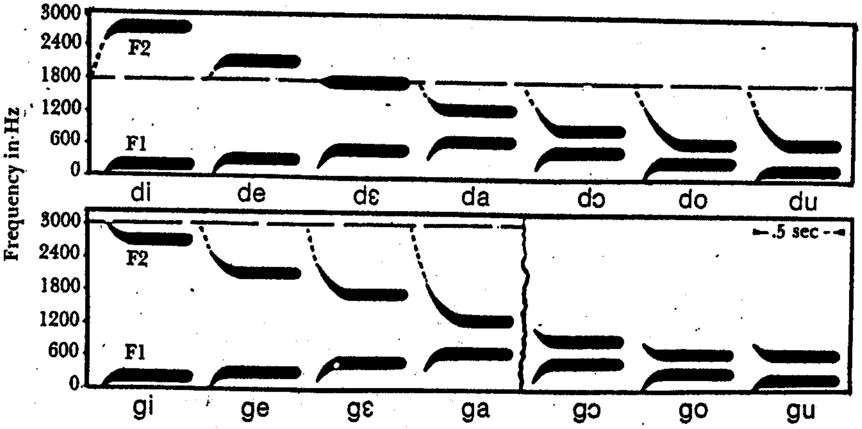

Figure 10. Two-formant spectrographic patterns for [d] and [g], with vowels ranging from [i] to [u]. While the [d] locus is consistent at 1800 Hz, the [g] locus at 3000 Hz breaks down before back-rounded vowels.

If speech was once a social releaser system, we should expect it to show adaptation in the direction of "communications security." While being as conspicuous as possible on appropriate occasions to conspecific individuals, social releasers should be otherwise as inconspicuous as possible, in particular to prey and to predators. In the case of visual releasers, various camouflaging arrangements are found: outside the courtship period, the stickleback changes the color of his belly to a less noticeable shade, and birds hide their brilliant plumage (Tinbergen 1951). In the case of acoustic releasers, the animal can become silent when this is expedient; the simplicity of this solution is the great advantage of acoustic systems. As for speech, two of the differences we have noted between sign stimuli and speech cues are probably to be interpreted as further adaptations in the direction of security. The rapid rate at which the speech cues can be transmitted and their parallel processing mean that, when necessary, transmissions can be extremely brief, making it so much the more difficult for an enemy to locate the source of the signal.

Moreover, if we regard speech as a social releaser system, a natural explanation is available for an old problem. The fact that no other animal except man can speak, not even the primates to whom he is most closely related, has long been a cause for wonder and speculation. But, of course, a social releaser is required, almost by definition, to be species-specific: it must be so if it is to perform its authentication function effectively. It is thus no more surprising that speech should be unique to man than that zigzag dances should be unique to sticklebacks.

Origins of speech and language

Let us now consider how the concept of prelinguistic speech as consisting of a system of phonetic social releasers bears on the problem of the origin of language. Most speculations on this topic suppose that man's unusual intelligence must have been the principal factor in the development of language. The weaker version of this view (which would have been that of many post-Bloomfieldian linguists) assumes that man's intelligence differs from that of animals in degree: he alone is intelligent enough to divide the world into its semantic categories and to recognize their predicative relationships. The structure of his language, insofar as it is not purely a matter of convention, reflects the structure of human experience. The stronger version of this view (which I think it is fair to attribute to Chomsky and his colleagues) assumes that man's intelligence differs in kind from that of other animals, and that the structure of language, properly understood, reflects specific properties of the human intellect. Speech, according to either version, serves simply as the vehicle for the abstract structure of language. The anatomy of the vocal tract imposes certain practical constraints on linguistic behavior, but has only a trivial relationship to linguistic structure.

The difficulty with this view is not only that it makes no attempt to account for the choice of speech as the vehicle of language but also that many animals display some degree of intelligence, and a few display intelligent behavior comparable in some ways to man's. One would expect to find some limited linguistic behavior among animals of limited intelligence, or something approximating human linguistic behavior among animals whose intelligence seems to resemble man's. But, as we have seen, precedents of any kind are lacking, and it is argued that language is an instance of evolutionary "emergence."

I wish to suggest a somewhat less drastic alternative to emergence. This is that language be regarded as the result of the fortunate coexistence in man of two independent mechanisms: an intellect capable of making a semantic representation of the world of experience, and the phonetic social-releaser system, a reliable and rapid carrier of information. From these mechanisms a method evolved for representing semantic values in communicable form.

Before this could happen, a means had to be found for the speaker-hearer to recode semantic representations into phonetic representations, and phonetic representations into semantic representations. Clearly this recoding is a complex process, if only because the intellect, being capable of representing a wide range of human experience, probably has a very large number of categorical features available for semantic representations in long-term memory, while the phonetically significant configurations of the vocal tract can be described in terms of a very small number of categorical features—fifteen or twenty at most (Chomsky and Halle 1968). It would thus be impossible to accomplish the recoding simply by mapping semantic features onto phonetic features. It was necessary for another mechanism to evolve: linguistic capacity, the ability to learn the grammar of a language. The grammar is a description of the complex but rule-governed relationships, in part universal, in part language-specific, which obtain between semantic representations and phonetic representations. By virtue of his grammatical competence, a person can speak and understand utterances in the language according to the rules of the grammar.

One component of the grammar is the lexicon, a list of morphemes with which semantic, syntactic, and phonological information is associated. The stock of morphemes in a language is large but finite, while the number of conceivable semantic representations is infinite. But an infinite number of grammatical strings of morphemes can be generated by the syntactic component of the grammar, and from these, the semantic component can generate a correspondingly infinite number of semantic representations. The phonological component parallels the semantic component: for each string of grammatical morphemes, a phonetic representation can be generated. The speaker's task is thus to find a phonetic representation which corresponds grammatically to a given semantic representation, while the hearer's task is to find a semantic representation corresponding to a given phonetic representation. In both his roles, the speaker-hearer, in order to recode, must determine heuristically the probable input to a grammatical component, given its output and the rules which generate output from input. Very little is known about how he performs these tasks.

(The account of the organization of grammar given here, necessarily oversimplified, is based on Chomsky [1965, 1966]. I have ignored for simplicity's sake the obvious fact that there are not one but many languages, each with its own grammar. To Rousseau (1755) and von Humboldt (1836), to explain the diversity of human languages was regarded as a problem second in importance only to that of explaining the origin of language. Recently, Nottebohm (1970) has offered the intriguing suggestion, based on an analogy with bird song, that language diversity enables some members of a species to develop traits appropriate to their particular environment without an irreversible commitment to subspeciation.)

For our purposes, however, the important point is that a grammar has an obvious symmetry. There is a core, the syntactical and lexical components, and two other components, the semantic and the phonological, which generate the semantic and phonetic representations, respectively. The nature of the semantic component and the representation it generates appear to be appropriate for storage in long-term memory. The nature of the phonological component and the representation *it* generates are appropriate for on-line transmission by the vocal tract. To relate these two representations is the main motivation of the grammar, and its form is determined both by the properties of the intellect and those of the phonetic social-releaser system. It is thus surely not correct to view speech as though it were merely selected by happenstance as a convenient vehicle for language.

Once the grammar had begun to develop, we should not be surprised to find that it exercised a reciprocal influence on the development both of the phonetic system and of the intellect. In the case of the former, it has been argued very persuasively (Lieberman and Crelin 1971; Lieberman et al. 1972) that the vocal tract of modern man has evolved from something rather like that of a chimpanzee to its present form, with a shorter jaw, a wider and deeper pharynx, and vocal cords for which the tension is more finely controlled, and that these modifications not only have no other discernible adaptive value than to increase the reliability and the richness of structure of human speech but are actually disadvantageous for the vocal tract's primary functions of chewing, breathing, and swallowing.

This view reminds us of a difference between speech cues and sign stimuli not yet resolved: how did speech come to be encoded in articulatory gestures? At least a conjecture is possible if we consider that the behavior of nonhuman primates, such as chimpanzees, that most closely resembles in function the social releasers is their repertory of facial grimaces. If such a grimace—for example the protruding and rounding of the lips—is accompanied by vocal-cord vibration, an acoustic event characteristic of lip-rounding—a drop in the frequency of all formants—will be heard. Thus, even before the evolution of the articulators in the direction of speech production began, there was a potential acoustic cue corresponding to an overt articulatory gesture. Since most of us lip-read when face to face with another speaker, especially in a noisy environment, it appears that we still relate the overt labial articulation to its acoustic coding.

Primitive man, on the other hand, may have used the acoustic cue when darkness or distance made face-to-face contact impossible. Similarly, he could relate visible jaw movement to its acoustic correlate—the rise in *F*1 as the jaw is lowered. Once having developed this limited ability to interpret acoustic cues in articulatory terms, it would not seem unreasonable that this ability could be generalized to include covert as well as overt articulators. Thus, speech might be regarded as a way of making invisible faces. In fact, if Lieberman and his colleagues are correct in their conclusion that the articulators evolved in ways that are adaptive for speech production, we could add that this evolution may also have been determined so as to meet the requirements of a preexisting, primitive form of the neural mechanism for speech perception. The articulators of modern man evolved as they did because their position and movement produce acoustic events which can serve as cues because they are well matched to man's capacity for decoding speech. In this way, the speech code became an increasingly useful device for grammatical purposes.

The evidence for the development and specialization of the human intellect as a result of its grammatical affinities is of course far less concrete. But the very least that can be said is that the capability of symbolizing things and ideas by words permits a degree of conceptual abstraction without which the kind of thinking human beings regularly do would be impossible. If the function of a grammar is to serve as an interface between the phonetic and semantic domains, it is hardly surprising that precedents for linguistic behavior have not been found. The speech production and perception system is a highly specific mechanism; so also is the human intellect. Their co-occurrence in man was a remarkable piece of luck; other animals, which on behavioral or physiological grounds appear to be of high intelligence, had no opportunity to develop language because they lacked a suitable prexisting communications system. Moreover, even if high intelligence and an appropriate communications system had coincided in some other species and combined to form a "language," its grammar would be utterly different in form from any human grammar, because the intellectual and communicative mechanisms from which it evolved would be quite different in detail from the corresponding human mechanisms. In the circumstances, the most we can hope for is to understand more about the separate evolution of the intellect and that of the speech code, and to interpret human grammars in terms of their dual origin.

To summarize, I have called attention to certain parallels between the speech cues and sign stimuli. These parallels suggest the speculation that prelinguistic speech may have functioned as a social-releaser system, which would explain the fact that speech is species-specific. It is suggested, furthermore, that human language is not simply the product of the human intellect but is rather to be viewed as the joint product of the intellect and of this prelinguistic communications system. Grammar evolved to interrelate these two orginally independent systems. Its dual origin explains the lack of precedents for language in animal behavior and its apparent "emergence."

References

Abramson, A. S., and L. Lisker. 1970. Discriminability along the voicing continuum: Cross-language tests. In *Proc. 6th International Cong. Phonetic Sciences*. Prague: Academia.

Altmann, S. A. 1967. The structure of primate social communication. In S. A. Altman, Ed., *Social Communication among Primates*. Chicago: Univ. of Chicago Press.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: M.I.T. Press.

Chomsky, N. 1968. *Language and Mind*. New York: Harcourt Brace.

Chomsky, N. 1966. *Topics in the Theory of Generative Grammar*. The Hague: Mouton.

Chomsky, N., and M. Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Conrad, R. 1964. Acoustic confusions in immediate memory. *Brit. J. Physiol.* 55:75–83.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. 1952. Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Amer.* 24:597–606.

Delattre, P. C., A. M. Liberman, and F. S. Cooper. 1955. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Amer.* 27:769–73.

Eibl-Eibesfeldt, I. 1970. *Ethology*. New York: Holt, Rinehart and Winston.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. 1970. Speech perception in infants. *Science* 171:303–06.

Falls, J. B. 1969. Functions of territorial song in the white-throated sparrow. In R. A. Hinde, Ed., *Bird Vocalizations in Relation to Current Problems in Biology and Psychology*. Cambridge: Cambridge University Press.

Hailman, J. P. 1969. How an instinct is learned. *Scientific American* 221:(6)98–106.

Halle, M., G. W. Hughes, and J.-P. A. Radley. 1957. Acoustic properties of stop consonants. *J. Acoust. Soc. Amer.* 29:107–16.

Harris, K. S., H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. 1958. Effect of third-formant transitions on the perception of the voiced stop consonants. *J. Acoust. Soc. Amer.* 30:122–26.

Hinde, R. A. 1970. *Animal Behavior*. (2nd Ed.) New York: McGraw-Hill.

Hockett, C. F., and S. A. Altmann. 1968. A note on design features. In T. A. Sebeok, Ed., *Animal Communication*. Bloomington, Ind.: Indiana Univ. Press.

Hoffman, H. S. 1958. Study of some cues in the perception of the voiced stop consonants. *J. Acoust. Soc. Amer.* 30:1035–41.

Kimura, D. 1961. Cerebral dominance and the perception of verbal stimuli. *Canad. J. Psychol.* 15:166–71.

Kimura, D. 1964. Left-right differences in the perception of melodies. *Quart. J. Exp. Psychol.* 16:355–58.

Koehler, O., and A. Zagarus. 1937. Beiträge zum Brutverhalten des Halsbandregenpfeifers (*Charadrius h. hiaticula* L.) *Beitr. Frotpflanzungs biol. Vögel* 13:1–9. Cited by Tinbergen 1951.

Lenneberg, E. 1967. *Biological Foundations of Language*. New York: John Wiley.

Liberman, A. M. 1957. Some results of research on speech perception. *J. Acoust. Soc. Amer.* 29:117–23.

Liberman, A. M., F. S. Cooper, and K. S. Harris. 1963. A motor theory of speech perception. In *Proceedings of the Speech Communications Seminar*. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. 1967. Perception of the speech code. *Psycol. Rev.* 74:431–61.

Liberman, A. M., P. C. Delattre, L. J. Gerstman, and F. S. Cooper. 1956. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J. Exp. Psychol.* 52:127–37.

Lieberman, P., and Edmund S. Crelin. 1971. On the speech of Neanderthal man. *Linguistic Inquiry*. 2:203–22.

Lieberman, P., E. S. Crelin, and D. H. Klatt. 1972. Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist* 74: in press.

Lisker, L. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Lang.* 33:42–49.

Lisker, L., and A. S. Abramson. 1970. The voicing dimension: Some experiments in comparative phonetics. In *Proc. 6th International Cong. Phonetic Sciences*. Prague: Academia.

Lisker, L., F. S. Cooper, and A. M. Liberman. 1962. The uses of experiment in language description. *Word* 18:82–106.

Lorenz, K. 1935. Der Kumpan in der Umwelt des Vogels. *J. f. Ornith.* 83:137–213, 289–413. Tr. in C. H. Schiller, ed., 1957, *Instinctive Behaviour*, London: Methuen.

Lorenz, K. 1954. Das angeborene Erkennen. *Natur und Museum* 84:285–95.

Manning, A. 1967. *An Introduction to Animal Behavior*. Reading, Mass.: Addison-Wesley.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. 1971. Discrimination in speech and nonspeech modes. *Cognitive Psychol.* 2:131–57.

Nottebohm, F. 1970. Ontogeny of birdsong. *Science* 167: 950–66.

Orr, D. B., H. L. Friedman, and J. C. C. Williams. 1965. Trainability of listening comprehension of speeded discourse. *J. Educ. Psychol.* 56:148–56.

Rousseau, J.-J. Essay on the origin of languages. Originally written c. 1755. Tr. by J. H. Moran in J. H. Moran and A. Gode, eds., *On the Origin of Language*, New York: Ungar, 1966.

Russell, E. S. 1943. Perceptual and sensory signs in instinctive behavior. *Proc. Linnaean Soc. London* 154:195–216.

Seitz, A. 1940. Die Paarbildung bei einigen Cichliden I. *Zs. Tierpsychol.* 4:40–84. Cited by Tinbergen 1951.

Shearme, J. N., and H. N. Holmes. 1962. An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1–formant 2 plane. In *Proc. Fourth Int. Cong. Phonetic Sciences*. The Hague: Mouton.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. 1970. Motor theory of speech perception: a reply to Lane's critical review. *Psychol. Rev.* 77:234–49.

Studdert-Kennedy, M., and D. Shankweiler. 1970. Hemispheric specialization for speech perception. *J. Acoust. Soc. Amer.* 48:579–94.

Thorpe, W. H. 1961. Introduction to: Experimental studies in animal behaviour. In W. H. Thorpe and O. L Zangwill, eds., *Current Problems in Animal Behaviour*. Cambridge: Cambridge University Press.

Tinbergen, N. 1951. *The Study of Instinct*. Oxford: Clarendon Press.

von Humboldt, Wilhelm. *Linguistic Variability and Intellectual Development*. Originally published 1836. Tr. by G. C. Buck and F. A. Raven. Coral Gables, Fla.: Univ. of Miami Press, 1971.

Wickelgren, W. A. 1966. Distinctive features and errors in short-term memory for English consonants. *J. Acoust. Soc. Amer.* 39:388–98.