The Human Aspect of Speech*

Ignatius G. Mattingly+
Haskins Laboratories, New Haven, Conn.

Underlying the recent work of many of us at this conference is the question I would like to consider today. It may at first seem strange and unnecessary, but I hope to persuade you of its relevance and importance. My question is, what aspects of speech, if any, are peculiarly and distinctively human?

At one time, it would have been enough to answer that speech is the vehicle for language, and only human beings have language. Both of those propositions are now in dispute. Some people seriously question whether there is really any justification for reserving the term "language" for human communication: Premack (1972) and Lehrman (1973) suggest that chimpanzees have language. Students of sign and gesture point out that early man may have communicated linguistically without benefit of speech (Hewes, 1973), and that many deaf persons certainly do so (Bellugi and Fischer, 1972; Stokoe, this conference); speech is perhaps only one of several vehicles for language. Moreover, there is reason to believe that speech evolved independently of language: structural parallels have been noted between speech and various animal communication systems that no one would call languages (Mattingly, 1972). Thus we should not attach undue significance to the fact that speech is specific to man; its peculiarities might prove on closer observation to be of a not very profound kind, like the details that distinguish the courtship display of one species of gull from that of another (Tinbergen, 1951). But in the face of these considerations, I would maintain that there is a truly human aspect of speech, something that marks it unmistakably as a product of man's cognitive powers.

Let me begin my pursuit of the human aspect of speech with a very general account of linguistic capacity. We suppose, with the generative grammarians, that the speaker-hearer has to deal with two significant versions of an utterance: a phonetic representation and a semantic representation (Chomsky, 1965). The phonetic representation of the utterance is in a form suitable for transmission by the vocal apparatus; the semantic representation is in a form suitable for storage in long-term memory (Liberman, Mattingly, and Turvey, 1972). It is convenient to conceive of both of these representations as n-dimensional arrays.


+Also University of Connecticut, Storrs.

of features. The features of the phonetic representation are few in number and refer to the properties of the vocal tract (Chomsky and Halle, 1968), while those of the semantic representation are presumably far more numerous and refer to the whole of human experience. In the course of speaking or understanding an utterance, the speaker-hearer forms both of these representations. He can do so because he knows, tacitly, how the phonetic representation relates to the acoustic speech signal, how the semantic representation relates to the contents of long-term memory, and how the two representations relate to one another. This way of describing linguistic capacity suggests that phonetics, grammar, and semantics exhibit significant parallels, and this is just the impression I am striving to create. Let me try to bring out the parallelism by looking at each of these forms of cognitive activity in turn.

Speech differs in interesting ways from other natural communication systems. To be sure, some animal communication systems share many "design features" (Hockett and Altmann, 1968) with speech, and there are striking parallels between the perception of speech and the perception of "sign stimuli" (Mattingly, 1972). But for none of these systems is it difficult to imagine how the perceptual powers of the users can cope with the amount of information known to be contained in the signal. The messages are typically very simple indeed. They are central to the survival of individuals and species, but their information content is low. With speech, however, the problem is to explain how the system can convey as much information as it does, without overwhelming the ear. Most of us will recall Liberman's (in Kavanagh, 1968) account of the obstacles encountered in developing various alphabets composed of discrete sounds to be used in a reading machine for the blind. As in speech, linguistic information was being communicated by an acoustic signal. Yet none of the sound alphabets could be understood at information rates comparable to that of natural speech. At a considerably lower rate, the individual alphabet sounds merged in a buzz (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967).

Again, consider the difference between speech and various possible gesture systems that might conceivably convey linguistic information. The comparison is the more appropriate because speech itself is a system of articulatory gestures. Perhaps the simplest possible gesture system would be one consisting of gross bodily movement—shrugging one's shoulders, tossing one's head, turning one's back. Such gestures, we know, can convey attitudinal meaning either alone or in conjunction with speech (Eibl-Eibesfeldt, 1970). But obviously the repertoire is not large and the semantic possibilities are limited. However, if we allow independent movements of the arms, the potential repertoire of gestures increases. If the two hands work together, and can point to or touch the various parts of the upper body, a still larger repertoire is available, and the perceiver can concentrate his attention on a fairly small part of his visual field. This is the method of sign language. Finally, consider facial gestures. The various physiognomic features can move independently and quite rapidly, and they are collocated in a fairly small visual area. Nonhuman primates communicate with facial gestures and human beings use them expressively, exploiting the extreme mobility of the face to express an extraordinary range of attitudes. One could even conceive of a form of sign language in which linguistic information would be transmitted by rapid and quasi-independent motions of brow, eye, nose, mouth, and chin. Such a physiognomic communications system might in principle carry more information than a manual system. However, it would be difficult or impossible to track the concurrent movements of the various facial features. Perhaps the motion of two hands is as much as the eye can comfortably follow.
But speech apparently offers a way of overcoming this kind of limitation. This is in a way rather surprising, for speech would not seem to be a highly efficient communication system. In certain respects it resembles our imaginary physiognomic system: a group of collocated articulators—larynx, tongue, velum, jaw, lips—are moving quasi-independently, and the perceiver has the difficult task of following these different movements. But there is a further problem. The perceiver of physiognomic gestures has a direct and continuous display, whereas the display available to the perceiver of speech is partial and indirect: it consists of the sounds that happen to be made as air passes through the shifting cavities and passages formed by the moving articulators. Yet from the indirect record contained in the acoustic signal, the listener can extract without difficulty the information carried by the articulatory gestures.

Speech can circumvent the limitations on both ear and eye because, unlike human gesture systems and animal communication systems, it is "encoded" (Liberman et al., 1967). To clarify what this means, consider the mapping of the phonetic representation on the speech signal. Most of the information is carried by a few prominent acoustic features: the fundamental frequency, the first two or three formants, the plosive bursts, and the patches of fricative noise. The rest of the signal can be discarded with little or no loss in intelligibility. Moreover, the acoustic signal does not consist of temporal segments corresponding to successive phones. Rather, the acoustic features typically carry information about two or more phones in parallel. These economies mean that the load on the input channel is much lighter than it would be if the perceiver had to attend to separate acoustic units, or to a visible array of gestures. Thus, phonetic information can be transmitted at a much higher rate.

The price for this gain at the input is the complexity with which the cues that are the basic data for speech perception are represented by the acoustic features (Mattingly and Liberman, 1969). Thus we find that the cues for two or more successive phones may be carried simultaneously by the same acoustic event: a second-formant transition cues the place of articulation of a stop, and the place of the adjacent vowel. On the other hand, different cues for one phone may be far apart: when two vowels are separated by a consonant, the place of the second vowel may be cued not only by the quasi-steady-state of its second formant, and the adjacent transition, but also by the second-formant transition between the first vowel and the consonant. Finally, quite different cues will signal the same phone in different environments: an alveolar consonant is cued by a rising second-formant before a front vowel and a falling second-formant before a back vowel. Indeed, the dispersal of information in time and frequency offers ample justification for Hockett's comparison of the speech signal to smashed Easter eggs (Hockett, 1955:210).

When we investigate the sources of this complexity, we find that a restructuring of information occurs in at least three different ways. The most obvious restructuring is an acoustic one. Variation in the spectrum of the speech signal is determined in part by the changing shapes of the vocal tract cavities (Fant, 1960). The relation of cavity shape to spectrum is hardly straightforward and sometimes ambiguous, but it is the movements of the articulators that are significant, and these are only indirectly reflected in the changes in cavity shapes over time. However, the matter is even more complicated. If an articulator consistently moved in such a way that, as a consequence of each motion, it attained a target position associated with some phonetic value, it would be possible to relate to each phone a target articulatory configuration, and hence a target
shape and a target spectrum. But while such targets can be hypothesized, they are actually attained only in simple cases. More commonly, targets are merely approached, and the different articulators participating in the production of one phone do not generally come closest to their targets at the same time (Lindblom, 1963). Furthermore, the motion of an articulator is ordinarily complex, determined by preceding and following phones as well as by the current one (Öhmann, 1966). With electromyographic techniques, the different gestures underlying complex articulatory motion can frequently be distinguished at the neuromotor level (K. S. Harris, personal communication) as commands from different muscles. Yet restructuring can take place even at this stage of production; the muscle commands themselves are sensitive to phonetic context (MacNeilage and DeClerk, 1969).

Thus the task of speech perception is even more complicated than we originally suggested. What the listener has to recover from the acoustic data is not the mere physical motion of the vocal organs but the articulatory plan that is realized in this motion. How can we do this? At least part of the answer is that he is able to bring to his task information that severely constrains his perceptual hypotheses. My colleagues at Haskins Laboratories have argued that the listener has tacit knowledge of certain properties of the vocal tract, and they have proposed a "motor theory of speech perception" (Liberman et al., 1967). Rather than reviewing their arguments, let us take it that the theory is essentially correct, and consider the kind of knowledge the theory imputes to the listener. Certainly, it is not the kind of knowledge that could be deduced from communication theory, or even from an analysis of the acoustic speech signal alone. The vocal tract is a highly eccentric collection of disparate structures with distinct primary functions. Though it has undergone some remodeling to make it a more serviceable signaling device (Lieberman, 1968), it is essentially a bizarre arrangement that can be rationalized only in evolutionary terms. Nor is it the kind of knowledge that the listener might be supposed to derive from his own experience as a speaker. Experience in speaking is neither necessary, as is known from clinical cases in which damage to, or congenital deformation of the vocal tract does not interfere with speech perception (Lenneberg, 1967), nor sufficient, since it would not be adequately generalized knowledge. We need to assume that the listener's tacit knowledge is of a more abstract character if we are to account for his ability to recover phonetic information from the output of vocal tracts of different shapes and sizes. We might imagine his knowledge as the equivalent of a dynamic vocal tract model, an ideal speech synthesizer. With a few adjustments, the model is good for any speaker. It is a highly selective model, enabling the processes of perception to extract information from the signal received by the ear, even though this signal is a complexly encoded record of articulation.

If such a model seems over-elaborate, consider that it is required also to constrain the articulatory plan of an utterance if the speaker is not to make inconsistent or impossible demands on his articulators, and to monitor the utterance as it is being produced. Production and perception are regulated by the same tacit knowledge. The model is also needed to account for the fact that the

---

1 I do not mean by these comparisons to suggest a "process" model. Neither the proposed model nor any synthesizer actually recapitulates the processes of production; rather, they demonstrate the relationship of selected phonetic variables to acoustic output.
infant must deduce the phonetic rules of his language from speech produced by adult vocal tracts very different from his own (Lieberman, Crelin, and Klatt, 1972), and must learn to manipulate his own vocal tract, accommodating to its individual variations and to its changes in shape and size as it matures.

Speech perception, then, is a very powerful process because it applies to the analysis of a certain kind of very complex data a profound, specialized knowledge about such data. This cannot be said of sign language, or of the gestural communication systems imputed to early man.

Let us turn now, adopting the conceptual framework of generative grammar, to the relationship between the phonetic and semantic representations. If we compare these two representations for some utterance, we find that in the phonetic representation, the elementary propositions of the semantic representation ("deep structure") are internally reordered and combined with one another, that some lexical morphemes have been deleted or anaphorically replaced, and that other morphemes have been introduced, any one of them perhaps representing two or more semantic or syntactic elements. At morpheme boundaries, as well as within morphemes, there is extensive phonological revision: sounds have been inserted, deleted, changed in one or more distinctive features, or transposed with one another. In short, the phonetic representation is an encoding of the semantic representation (Mattingly and Liberman, 1969). The effect of the encoding is to make the syntax of the utterance as compact and the articulation as efficient as possible. The speaker-hearer's ability to produce appropriate phonetic representations, as well as his ability to reconstruct the semantic representation from the phonetic representation, is dependent upon his competence: his internalized knowledge of the grammatical rules of his language. This grammar is so highly specified that he is able to judge the grammaticality of any utterance and to recover its deep structure. The grammar is generative; that is, the grammatical analysis of an utterance is its derivation from deep structure according to rules. Infinitely many such utterances can be derived.

A generative grammar plays a role in the production and the understanding of sentences similar to the role played by the vocal tract model in speech production and perception. It embodies the knowledge that enables the encoding to be correctly imposed and removed. The parallel may not seem immediately obvious because our way of knowing grammatical facts is very different from our way of knowing phonetic facts. Phonetic activity is to some extent observable, but grammatical activity is not. On the other hand, we have considerable intuitional insight into grammar and little or none into phonetics. Grammar is customarily presented as a system of formal rules rather than as a neurophysiological model.

This formalism, however, brings out certain interesting restrictions: the division of the grammar into components with different types of rules, the nonoccurrence of certain transformational patterns, the type of context that must be stated in a phonological rule, and so forth (Chomsky, 1957). These restrictions are not functional; they reflect indirectly the idiosyncrasies of the underlying neurophysiological apparatus, though we cannot observe it directly as a flesh and blood reality, as we can the vocal tract. Moreover, such formalism is what we would expect of our articulatory model if it is to be the abstract thing we have suggested, independent of particular vocal tracts yet capturing their essential common features.
Our claim, then, is that the listener's grammatical analysis and his phonetic analysis are really quite similar forms of cognitive activity. In each case, the listener brings to a complex array of data tacit knowledge sufficiently well-specified for him to determine the underlying structure of the array and to remove the encoding.

Consider finally the question, how does the speaker get the semantic representations that he encodes linguistically? This is of course a special case of a problem central to cognitive psychology: how is information stored in and recovered from memory? We know that the capacity of long-term memory must be enormous, but we do not know how we thread our way through the labyrinth so readily. Yet some inferences useful to our present purposes are prompted by the familiar phenomenon of paraphrase (Lieberman et al., 1972). If someone is asked to recall a sentence he has previously heard, he responds, typically, with a sentence that probably differs semantically (and also grammatically and phonetically) from the original, though probably also remaining semantically consistent with it. In fact, we would find it strange if on a certain day A were to say to B, "I'm coming tomorrow," and B were to report this the following day saying, "A said I'm coming tomorrow," without indicating by appropriate intonation that he was quoting A directly. We would expect rather a paraphrased response that takes into consideration changes in the time, the speaker, and the world ("A said that he's coming today."). The phenomenon of paraphrase suggests that the semantic representations of the sentences one hears are not ordinarily preserved in memory as separate items. Indeed, this would be a most inefficient way of storing information, given the redundancy of human discourse. Rather, the semantic content of the sentence is somehow incorporated into a more general record of experience that is the basis of both paraphrases and newly created sentences. In fact, the two cannot logically be distinguished.

Bartlett (1932) has suggested that experience is represented in memory by "schemata." A schema is not a chronological record of previous perceptions, each separately preserved, but rather an integration of these perceptions into an "organized setting" relating to a particular sort of experience. A new perception is drastically influenced by the schemata, and also modifies the schemata themselves. The same, presumably, can be said of the semantic representation of a newly heard sentence. It does not seem likely that there can be any simple mapping of parts of the semantic representation onto parts of the schema. The schema is the product of a great many semantic representations and, of course, of percepts of other kinds. On the other hand, a single semantic representation may conceivably modify many different aspects of the schema, and the nature of the modification may depend on the state of the schema itself. If we had available some convenient visual display of a schema in memory—an enormous spectrogram, as it were—we would doubtless find it very difficult to identify unambiguously the correlates of a particular sentence. Memorial processes are neither directly observable nor intuitively accessible, but we suspect that the integration of a semantic representation is a kind of encoding. It is similar in principle to the other encodings we have discussed, though different both in the content of what

---

2Lieberman (1973) thinks that the encodedness of speech is peculiarly human, just as I do, but that the cognitive ability underlying language differs only quantitatively from the "logical" ability underlying the conditioned responses of lower animals.
is encoded and in its scale, for the schema is an encoding not of one but of many sentences.

But if old sentences are lost, where do new sentences come from? The general affinity between perception and recall, and Bartlett's shrewd comment about remembering, are suggestive: "...the organism would say, if it were able to express itself: 'this and this and this must have occurred, in order that my present state should be what it is'" (Bartlett, 1932:202).

Putting this in the terms we have been using, to recover semantic information is to produce a semantic representation that if encoded would prove to be consistent with the current state of the schemata. Moreover, for independent linguistic reasons, we want the semantic representations underlying the sentences the speaker produces to be formally similar to the semantic representations of the sentences he hears: there is only one kind of deep structure. What is needed, therefore, both for storage and for recall, are rules that relate semantic representations to schemata: a sort of generative grammar of memory. These rules must reflect the nature of human experience that can be remembered. They must also reflect any purely nonfunctional properties of memory, attributable to its evolutionary history. And since we would not expect the processes of recall to vary markedly depending on the schemata of the individual, the rules must be abstract and general enough to transcend such individual differences. The storage and recovery of semantic information in memory is thus a further instance of the kind of cognitive operation that we have already observed in speech and language.

To recapitulate, I have tried to trace a cognitive pattern common to the processes of language, speech, and memory. In each of these processes information is thoroughly reorganized for functional reasons. The relationship of the original information to its reorganized form is complex: we have termed it encoded. Recovery of the information is accomplished with the help of a grammar which specifies the relationship between the unencoded and the encoded form of the information, and whose formal properties consequently reflect the nature of the encoding device. We cannot say that this cognitive pattern is unknown in lower animals: for example, the spider's knowledge of his web seems not dissimilar. But surely only in man is the pattern so highly developed and so diversely manifested. Of these manifestations, speech is of special interest. It is not as complex a system as language or memory, and it does not claim our attention so immediately when we reflect on the character of our knowledge. But it exemplifies nonetheless a thoroughly and peculiarly human kind of knowing.

REFERENCES