FRANKLIN S. COOPER

How Is Language Conveyed by Speech?

In a conference on the relationships between speech and learning to read, it is surely appropriate to start with reviews of what we now know about speech and writing as separate modes of communication. Hence the question now before us: How is language conveyed by speech? The next two presentations will ask similar questions about writing systems, both alphabetic and nonalphabetic. The similarities and differences implied by these questions need to be considered not only at performance levels, where speaking and listening are in obvious contrast with writing and reading, but also at the competence levels of spoken and written language. Here the differences are less obvious, yet they may be important for reading and its successful attainment by the young child.

In attempting a brief account of speech as the vehicle for spoken language, it may be useful first to give the general point of view from which speech and language are here being considered. It is essentially a process approach, motivated by the desire to use experimental findings about speech better to understand the nature of language. So viewed, language is a communicative process of a special—and especially remarkable—kind. Clearly, the total process of communicating information from one person to another involves at least the three main operations of production, transmission, and reception. Collectively, these processes have some remarkable properties: open-endedness, efficiency, speed, and richness of expression. Other characteristics that are descriptive of language processes per se, at least when transmission is by speech, include the existence of semantically "empty" elements and a hierarchical organization built upon them; furthermore, as we shall see, the progression from level to level involves restructuring operations of such complexity that they truly qualify as encodings rather than encipherings. The encoded nature of the speech signal is a topic to which we shall give particular attention since it may well be central to the relationship between speech and learning to read.

The Encoded Nature of Speech

It is not intuitively obvious that speech really is an encoded signal or, indeed, that it has special properties. Perhaps speech seems so simple because it is so common: everyone uses it and has done so since early childhood. In fact, the universality of spoken language and its casual acquisition by the young child—even the dullard—are among its most remarkable, and least understood, properties. They set it sharply apart from written language: reading and writing are far from universal, they are acquired only later by formal instruction, and even special instruction often proves ineffective with an otherwise normal child. Especially revealing are the problems of children who lack one of the sensory capacities—vision or hearing—for dealing with language. One finds that blindness is no bar to the effective use of spoken language, whereas deafness severely impedes the mastery of written language, though vision is still intact. Here is further and dramatic evidence that spoken language has a special status not shared by written language. Perhaps, like walking, it comes naturally, whereas skiing does not but can be learned. The nature of the underlying differences between spoken and written language, as well as of the similarities, must surely be relevant to our concern with learning to read. Let us note then that spoken language and written language differ, in addition to the obvious ways, in their relationship to the human being—in the degree to which they may be innate, or at least compatible with his mental machinery.

Is this compatibility evident in other ways, perhaps in special properties of the speech signal itself? Acoustically, speech is complex and would not qualify by engineering criteria as a clean, definitive signal. Nevertheless, we find that human beings can understand it at rates (measured in bits per second) that are five to ten times as great as for the best engineered sounds. We know that this is so from fifty years of experience in trying to build machines that will read for the blind by converting letter shapes to distinctive sound shapes [Coffey 1963; Cooper 1950; Studdert-Kennedy and Cooper 1966]; we know it also—and we know that practice is not the explanation—from the even longer history of telegraphy. Likewise, for speech production, we might have guessed from everyday office experience that speech uses special tricks to go so fast. Thus, even slow dictation will leave an expert typist far behind; the secretary, too, must resort to tricks such as shorthand if she is to keep pace.

Comparisons of listening and speaking with reading and writing are more difficult, though surely relevant to our present concern with what is learned when one learns to read. We know that, just as listening can outstrip speaking, so reading can go faster than writing. The limit on listening to speech appears to be about 400 words per minute [Orr, Friedman et al. 1965], though it is not yet clear whether this is a human limit on reception (or comprehension) or a machine limit beyond which the process used for time compression has seriously distorted the speech signal. Limits on reading speed are even harder to determine and to interpret, in part because reading lends itself to scanning as listening does not. Then, too, reading has its star performers who can go several times as fast as most of us. But, aside from these exceptional cases, the good reader and the average listener have limiting rates that are roughly comparable. Is the reader, too, using a trick? Perhaps the same trick in reading as in listening?

For speech, we are beginning to understand how the trick is done. The answers are not complete, nor have they come easily. But language has proved to be vulnerable to experimental attack at the level of speech, and the insights gained there are useful guides in probing higher and less accessible processes. Much of the intensive research on speech that was sparked by the emergence of sound spectrograms just after World War II was, in a sense, seduced by the apparent simplicities of acoustic analysis and phonemic representation. The goal seemed obvious: it was to find acoustic invariants in speech that matched the phonemes in the message. Although much was learned about the acoustic events of speech, and which of them were essential cues for speech perception, the supposed invariants remained elusive, just as did such promised marvels as the phonetic typewriter. The reason is obvious, now that it is understood: the speech signal was assumed to be an acoustic cipher, whereas it is, in fact, a code.

The distinction is important here as it is in cryptography from which the terms are borrowed: "cipher" implies a one-to-one correspondence between the minimal units of the original and final messages; thus, in Poe's story, "The Goldbug," the individual symbols of the mysterious message stood for the separate letters of the instructions for finding the treasure. In like manner, speech was supposed—erroneously—to comprise a succession of acoustic invariants that stand for the phonemes of the spoken message. The term "code" implies a different and more complex relationship between original and final message. The one-to-one relationship between minimal units has disappeared, since it is the essence of encoding that the original message is restructured (and usually shortened) in ways that are prescribed by an encoding algorithm or mechanism. In commercial codes, for example, the "words" of the final message may all be six-letter groups, regardless of what they stand for. Corresponding units of the original message might be a long corporate name, a commonly used phrase, or a single word or symbol. The restructuring, in this case, is done by substitution, using a code book. There are other methods of encoding—more nearly like speech—which restruc-

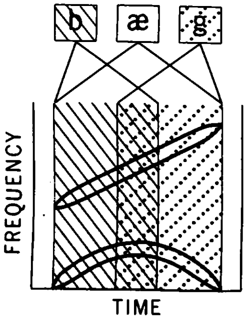

Figure 1. Parallel transmission of phonetic segments after encoding (by the rules of speech) to the level of sound (Liberman, 1970, p. 309).

ture the message in a more or less continuous manner, hence, with less variability in the size of unit on which the encoder operates. It may then be possible to find rough correspondences between input and output elements, although the latter will be quite variable and dependent on context. Further, a shortening of the message may be achieved by collapsing it so that there is temporal overlap of the original units; this constitutes parallel transmission in the sense that there is, at every instant of time, information in the output about several units of the input. A property of such codes is that the output is no longer segmentable, that is, it cannot be divided into pieces that match units of the input. In this sense also the one-to-one relationship has been lost in the encoding process.

The restructuring of spoken language has been described at length by Liberman, Cooper et al. [1967]. An illustration of the encoded nature of the speech can be seen in Figure 1, from a recent article [Liberman 1970]. It shows a schematic spectrogram that will, if turned back into sound by a speech synthesizer, say *bag* quite clearly. This is a simpler display of frequency, time, and intensity than one would find in a spectrogram of the word as spoken by a human being, but it captures the essential pattern. The figure shows that the influence of the initial and final consonants extend so far into the vowel that they overlap even with each other, and that the vowel influence extends throughout the syllable. The meaning of "influence" becomes clear when one examines comparable patterns for syllables with other consonants or another vowel: thus, the pattern for *gag* has a U-shaped second formant, higher at its center than the midpoint of the second-formant shown for *bag*; likewise, changing the vowel, as in *bog,* lowers the frequency of the second formant not only at the middle of the syllable but at the beginning and end as well.

Clearly, the speech represented by these spectographic patterns is not an acoustic cipher, that is, the physical signal is not a succession of sounds that stand for phonemes. There is no place to cut the syllable *bag* that will isolate separate portions for [b] and [æ] and [g]. The syllable is carrying information about all of them at the same time (parallel transmission), and each is affected by its neighbors (context dependence). In short, the phonetic string has been restructured, or encoded, into a new element at the acoustic level of the speech signal.

But is speech the only part of language that is encoded? Liberman's article, from which the illustration was drawn, asserts that comparable processes operate throughout language; that the encoding of speech and the transformations of syntactic and phonological structures are broadly similar and equally a part of the grammar. Thus, Figure 2 from the same article shows diagramatically the kind of restructuring and temporal compression that occurs in the syntactic conversion between deep and surface structure. Conventional orthography is used to represent the three deep-structure sentences and the single composite sentence at the surface. Again, there are overlapping domains, and compactness has been bought at the price of substantial changes in structure.

Encoding and Decoding

We see then, in all of spoken language, a very substantial degree of encoding. Why should this be so? Does it serve a purpose, or is it merely an unavoidable consequence of man's biological nature, or both? We have seen, in speech, that there is a temporal telescoping of the phonetic string into syllables and that this speeds communication; also, at the level of syntax, that there is a comparable collapsing of the deep structures into surface structures, with further gains in speed. Moreover,

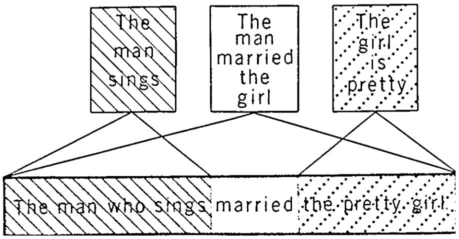

Figure 2. Parallel transmission of deep structure segments after encoding (by the rules of syntax) to the level of surface structure (Liberman, 1970, p. 310).

there are cognitive advantages that may be even more important, and that may explain why the encoding seems to have been done in stages, resulting in a hierarchical structure for language. George Miller [1956] has given us an account of how the magic of encoding lets us deal with substantial quantities of information in spite of limited memory capacity.

These are impressive advantages, but the price seems very high. We would suppose, from the foregoing, that the task of the person who listens to speech is staggeringly difficult: he must somehow deal with a signal that is an encoding of an encoding of an encoding. . . . Indeed, the difficulties *are* very real, as many people have discovered in trying to build speech recognizers or automatic parsing programs. But the human being does it so easily that we can only suppose he has access to full knowledge (even if implicit) of the coding relationships. These relationships, or a model of the process by which the encoding is done, could fully rationalize for him the involved relation of speech signal to underlying message and so provide the working basis for his personal speech decoder [Liberman 1970].

Our primary interest is, of course, in how speech is perceived, since this is where we would expect to find relationships with reading and its acquisition. It is not obvious that a person's implicit knowledge of how his own speech is produced might help to explain how another's speech can be perceived. Actually, we think that it does, although, even without such a premise, one would need to know how the encoding is done since that is what the decoder must undo. So before we turn to a discussion of how speech is perceived, let us first consider how it is produced.

The Making of Spoken Language

Our aim is to trace in a general way the events that befall a message from its inception as an idea to its expression as speech. Much will be tentative, or even wrong, at the start, but can be more definite in the final stages of speech production. There, where our interest is keenest, the experimental evidence is well handled by the kinds of models often used by communications engineers. This, together with the view that speech is an integral part of language, suggests that we might find it useful to extrapolate a communications model to all stages of language production.

The conventional block diagram in Figure 3 can serve as a way of indicating that a message (carried on the connecting lines) undergoes sequential transformations as it travels through a succession of processors. The figure shows a simple, linear arrangement of the principal processors

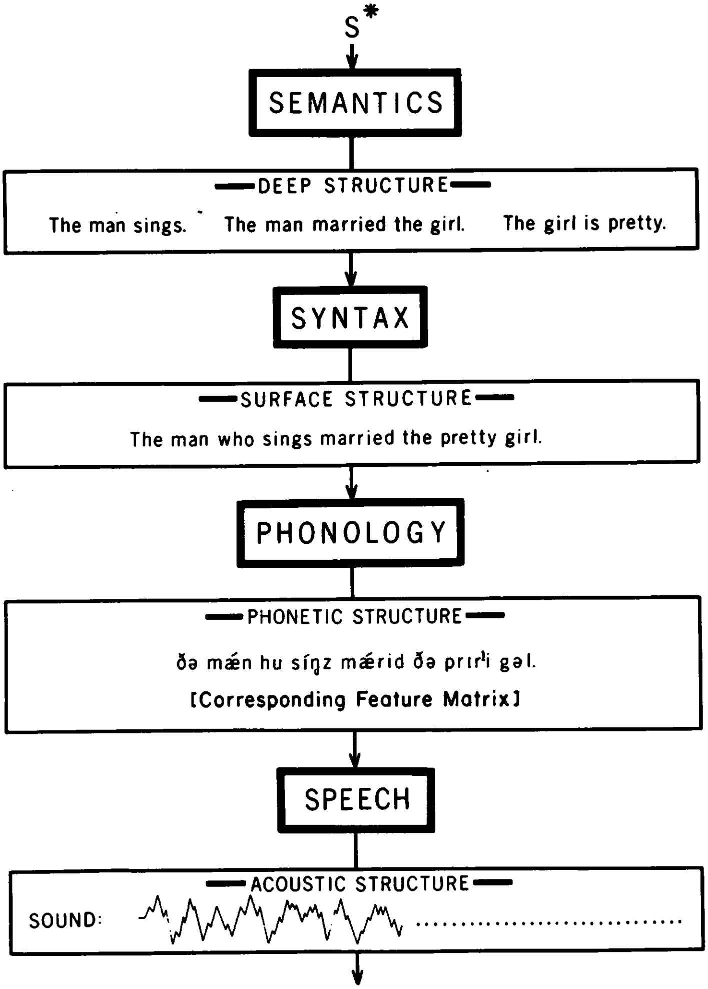

Figure 3. A process model for the production of spoken language. The intended message flows down through a series of processors (the blocks with heavy outlines). Descriptions are given (in the blocks with light outlines) of the changing form of the message as it moves from processor to processor. (Adapted from Liberman, 1970, p. 305.)

(the blocks with heavy outlines) that are needed to produce spoken language and gives descriptions (in the blocks with light outlines) of the changing form of the message as it moves from processor to processor on its way to the outside world. The diagram is adapted from Liberman [1970] and is based (in its central portions) on the general view of language structure proposed by Chomsky and his colleagues [Chomsky 1957, 1965; Chomsky and Miller, 1963]. We can guess that a simple, linear process of this kind will serve only as a first approximation; in particular, it lacks the feedback and feedforward paths that we would expect to find in a real-life process.

We know quite well how to represent the final (acoustic) form of a message—assumed, for convenience, to be a sentence—but not how to describe its initial form. S*, then, symbolizes both the nascent sentence and our ignorance about its prelinguistic form. The operation of the semantic processor is likewise uncertain, but its output should provide the deep structure—corresponding to the three simple sentences shown for illustration—on which syntactic operations will later be performed. Presumably, then, the semantic processor will somehow select and rearrange both lexical and relational information that is implicit in S*, perhaps in the form of semantic feature matrices.

The intermediate and end results of the next two operations, labeled Syntax and Phonology, have been much discussed by generative grammarians. For present purposes, it is enough to note that the first of them, syntactic processing, is usually viewed as a two-stage operation, yielding first a phrase structure representation in which related items have been grouped and labeled, and second a surface structure representation which has been shaped by various transformations into an encoded string of the kind indicated in the figure (again, by its plain English counterpart). Some consequences of the restructuring of the message by the syntactic processor are that (1) a linear sequence has been constructed from the unordered cluster of units in the deep structure and (2) there has been the telescoping of the structure, hence encoding, that we saw in Figure 2 and discussed in the previous section.

Further restructuring of the message occurs in the phonological processor. It converts (encodes) the more or less abstract units of its input into a time-ordered array of feature states, that is, a matrix showing the state of each feature for each phonetic event in its turn. An alternate representation would be a phonetic string that is capable of emerging at last into the external world as a written phonetic transcription.

This is about where contemporary grammar stops, on the basis that the conversion into speech from either the internal or external phonetic representation—although it requires human intervention—is straightforward and essentially trivial. But we have seen, with *bag* of Figure 1 as an example, that the acoustic form of a message is a heavily encoded version of its phonetic form. This implies processing that is far from trivial—just how far is suggested by Figure 4, which shows the major conversions required to transform an internal phonetic representation into the external acoustic waveforms of speech. We see that the speech processor, represented by a single block in Figure 3, comprises several subprocessors, each with its own function: first, the abstract feature matrices of the phonetic structure must be given physiological substance as neural signals (commands) if they are to guide and control the production of speech; these neural commands then bring about a pattern of muscle contractions; these, in turn, cause the articulators to move and the vocal tract to assume a succession of shapes; finally, the vocal tract shape (and the acoustic excitation due to air flow through the glottis or other constrictions) determines the spoken sound.

Where, in this sequence of operations, does the encoding occur? If we trace the message upstream—processor by processor, starting from the acoustic outflow—we find that the relationships between speech waveform and vocal tract shape are essentially one-to-one at every moment and can be computed, though the computations are complex [Fant 1960; Flanagan 1965]. However, at the next higher step—the conversion of muscle contractions into vocal tract shapes—there is substantial encoding: each new set of contractions starts from whatever configuration and state of motion already exist as the result of preceding contractions, and it typically occurs before the last set is ended, with the result that the shape and motion of the tract at any instant represent the merged effects of past and present events. This alone could account for the kind of encoding we saw in Figure 1, but whether it accounts for all of it, or only a part, remains to be seen.

We would not expect much encoding in the next higher conversion—from neural command to muscle contraction—at least in terms of the identities of the muscles and the temporal order of their activation. However, the contractions may be variable in amount due to preplanning at the next higher level or to local adjustment, via gamma-efferent feedback, to produce only so much contraction as is needed to achieve a target length.

At the next higher conversion—from features to neural commands—we encounter two disparate problems: one involves functional, physiological relationships very much like the ones we have just been considering, except that their location in the nervous system puts them well beyond

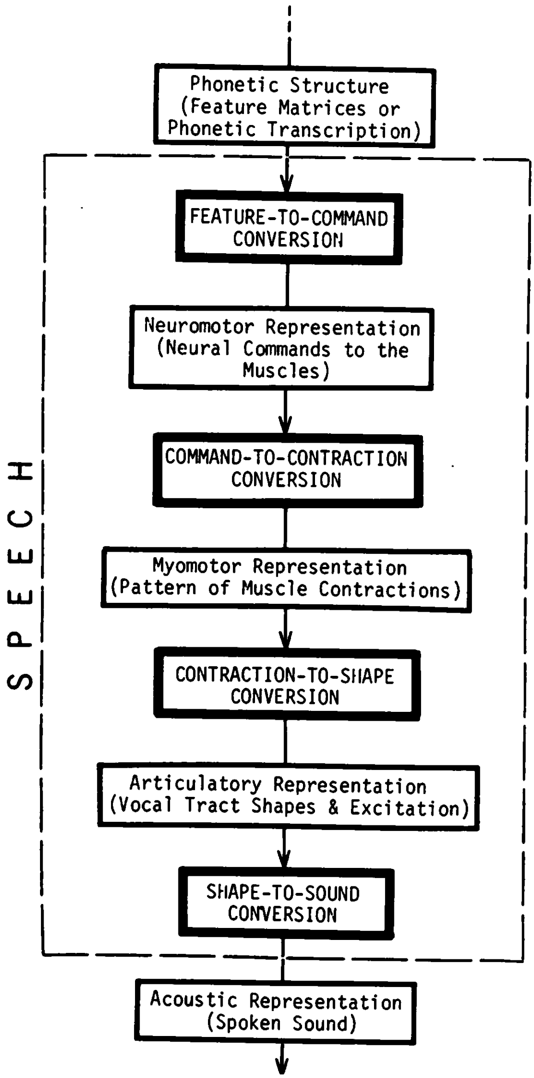
Phonetic Structure (Feature Matrices or Phonetic Transcription)
FEATURE-TO-COMMAND CONVERSION
Neuromotor Representation (Neural Commands to the Muscles)
COMMAND-TO-CONTRACTION CONVERSION
Myomotor Representation (Pattern of Muscle Contractions)
CONTRACTION-TO-SHAPE CONVERSION
Articulatory Representation (Vocal Tract Shapes & Excitation)
SHAPE-TO-SOUND CONVERSION
Acoustic Representation (Spoken Sound)
SPEECH

Figure 4. Internal structure of the speech processor. Again, the message flows from top to bottom through successive processors (the blocks with heavy outlines), with intermediate descriptions given (in the blocks with light outlines).

the reach of present experimental methods. The other problem has to do with the boundary between two kinds of description. A characteristic of this boundary is that the feature matrix (or the phonetic transcription) provided by the phonological processor is still quite abstract as compared with the physiological type of feature that is needed as an input to the feature-to-command conversion. The simple case—and perhaps the correct one—would be that the two sets of features are fully congruent, that is, that the features at the output of the phonology will map directly onto the distinctive components of the articulatory gestures. Failing some such simple relationship, translation or restructuring would be required in greater or lesser degree to arrive at a set of features which are "real" in a physiological sense. The requirement is for features rather than segmental (phonetic) units, since the output of the conversion we are considering is a set of neural commands that go *in parallel* to the muscles of several, essentially independent articulators. Indeed, it is only because the features—and the articulators—operate in this parallel manner that speech can be fast even though the articulators are slow.

The simplistic hypothesis noted above, that there may be a direct relationship between the phonological features and characteristic parts of the gesture, has the obvious advantage that it would avoid a substantial amount of encoding in the total feature-to-command conversion. Even so, two complications would remain. In actual articulation, the gestures must be coordinated into a smoothly flowing pattern of motion which will need the cooperative activity of various muscles (in addition to those principally involved) in ways that depend on the current state of the gesture, that is, in ways that are context dependent. Thus, the *total* neuromotor representation will show some degree of restructuring even on a moment-to-moment basis. There is a further and more important sense in which encoding is to be expected: if speech is to flow smoothly, a substantial amount of preplanning must occur, in addition to moment-by-moment coordination. We know, indeed, that this happens for the segmental components over units at least as large as the syllable and for the suprasegmentals over units at least as large as the phrase. Most of these coordinations will not be marked in the phonetic structure and so must be supplied by the feature-to-command conversion. What we see at this level, then, is true encoding over a longer span of the utterance than the span affected by lower level conversions and perhaps some further restructuring even within the shorter span.

There is ample evidence of encoding over still longer stretches than those affected by the speech processor. The sentence of Figure 2 provides an example—one which implies processor and conversion operations that lie higher in the hierarchical structure of language than does speech. There is no reason to deny these processors the kind of neural machinery that was assumed for the feature-to-command conversion; however, we have very little experimental access to the mechanisms at these levels, and we can infer the structure and operation only from behavioral studies and from observations of normal speech.

In the foregoing account of speech production, the emphasis has been on processes and on models for the various conversions. The same account can also be labeled a grammar in the sense that it specifies relationships between representations of the message at successive stages. It will be important, in the conference discussions on the relationship of speaking to reading, that we bear in mind the difference between the kind of description used thus far—a process grammar—and the descriptions given, for example, by a generative transformational grammar. In the latter case, one is dealing with formal rules that relate successive representations of the message, but there is now no basis for assuming that these rules mirror actual processes. Indeed, proponents of generative grammar are careful to point out that such an implication is not intended; unfortunately, their terminology is rich in words that seem to imply active operations and cause-and-effect relationships. This can lead to confusion in discussions about the *processes* that are involved in listening and reading and how they make contact with each other. Hence, we shall need to use the descriptions of rule-based grammars with some care in dealing with experimental data and model mechanisms that reflect, however crudely, the real-life processes of language behavior.

Perception of Speech

We come back to an earlier point, slightly rephrased: how can perceptual mechanisms possibly cope with speech signals that are as fast and complex as the production process has made them? The central theme of most current efforts to answer that question is that perception somehow borrows the machinery of production. The explanations differ in various ways, but the similarities substantially outweigh the differences.

There was a time, though, when acoustic processing per se was thought to account for speech perception. It was tempting to suppose that the patterns seen in spectrograms could be recognized *as patterns* in audition just as in vision [Cooper, Liberman et al. 1951]. On a more analytic level, the distinctive features described by Jakobson, Fant, and Halle [1963] seemed to offer a basis for direct auditory analysis, leading to recovery of the phoneme string. Also at the analytic level, spectrographic patterns were used extensively in a search for the acoustic cues for speech perception [Liberman 1957; Liberman, Cooper et al. 1967; Stevens and House, 1972]. All of these approaches reflccted, in one way or another, the early faith we have already mentioned in the existence of acoustic invariants in speech and in their usefulness for speech recognition by man or machine.

Experimental work on speech did not support this faith. Although the search for the acoustic cues was successful, the cues that were found could be more easily described in articulatory than in acoustic terms. Even the "locus," as a derived invariant, had a simple articulatory correlate [Delattre, Liberman et al. 1955]. Although the choice of articulation over acoustic pattern as a basis for speech perception was not easy to justify since there was almost always a one-to-one correspondence between the two, there were occasional exceptions to this concurrence that pointed to an articulatory basis, and these were used to support a motor theory of speech perception. Older theories of this kind had invoked actual motor activity (though perhaps minimal in amount) in tracking incoming speech, followed by feedback of sensory information from the periphery to let the listener know what both he and the speaker were articulating. The revised formulation that Liberman [1957, p. 122] gave of a motor theory to account for the data about acoustic cues was quite general, but it explicitly excluded any reference to the periphery as a necessary element: "All of this [information about exceptional cases] strongly suggests . . . that speech is perceived by reference to articulation—that is, that the articulatory movements and their sensory effects mediate between the acoustic stimulus and the event we call perception. In its extreme and old-fashioned form, this view says that we overtly mimic the incoming speech sounds and then respond to the appropriate receptive and tactile stimuli that are produced by our own articulatory movements. For a variety of reasons such an extreme position is wholly untenable, and if we are to deal with perception in the adult, we must assume that the process is somehow short-circuited—that is, that the reference to articulatory movements and their sensory consequences must somehow occur in the brain without getting out into the periphery."

A further hypothesis about how the mediation might be accomplished [Liberman, Cooper et al. 1968] supposes that there is a spread of neural activity within and among sensory and motor networks so that some of the same, interlocking nets are active whether one is speaking (and listening to his own speech) or merely listening to speech from someone else. Hence, the neural activity initiated by listening, as it spreads to the motor networks, could cause the whole process of production to be started up just as it would be in speaking (but with spoken output suppressed); further, there would be the appropriate interaction with those same neural mechanisms—whatever they are—by which one is ordinarily aware of what he is saying when he himself is the speaker. This is equivalent, insofar as awareness of another's speech is concerned, to running the production machinery backward, assuming that the interaction between sensory and motor networks lies at about the linguistic level of the features (represented neurally, of course) but that the linkage to awareness is at some higher level and in less primitive terms. Whether or not such a hypothesis about the role of neural mechanisms in speaking and listening can survive does not really affect the main point of a more general motor theory, but it can serve here as an example of the kind of machinery that is implied by a motor theory and as a basis for comparison with the mechanisms that serve other theoretical formulations.

The model for speech perception proposed by Stevens and Halle [1967; Halle and Stevens 1964] also depends heavily on mechanisms of production. The analysis-by-synthesis procedure was formulated initially in computer terms, though functional parallels with biological mechanisms were also considered. The computer-like description makes it easier to be specific about the kinds of mechanisms that are proposed but somewhat harder to project the model into a human skull.

It is unnecessary to trace in detail the operation of the analysis-by-synthesis model but Figure 5, from Stevens' [1960] paper on the subject, can serve as a reminder of much that is already familiar. The processing within the first loop (inside the dashed box) compares spectral information received from the speech input and held in a temporary store with spectral information generated by a model of the articulatory mechanism (Model I). This model receives its instructions from a control unit that generates articulatory states and uses heuristic processes to select a likely one on the basis of past history and the degree of mismatch that is reported to it by a comparator. The articulatory description that is used by Model I (and passed on to the next loop) might have any one of several representations: *acoustical,* in terms of the normal modes of vibration of the vocal tract; or *anatomical,* descriptive of actual vocal tract configurations; or *neurophysiological,* specifying control signals that would cause the vocal tract to change shape. Most of Stevens' discussion deals with vocal tract configuration (and excitation); hence, he treats comparisons in the second loop as between input configurations (from the preceding loop) and those generated by an articulatory control (Model II) that could also be used to drive a vocal-tract-analog synthesizer external to the analysis-by-synthesis system. There is a second

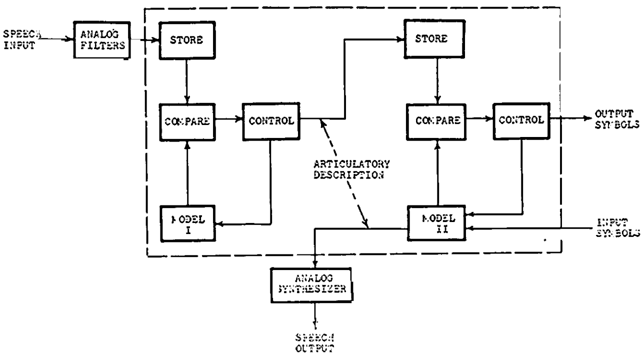

Figure 5. Analysis-by-synthesis model of speech recognition. The acoustic signal enters at upper left and is "recognized" in the form of a string of phonetic symbols that leave at center right. Model I stores the rules that relate articulatory descriptions to speech spectra, and model II stores the rules that relate phonetic symbols to articulatory descriptions. Model II can serve also to generate a speech output from an input of phonetic symbols (Stevens, 1960, p. 52).

controller, again with dual functions: it generates a string of phonetic elements that serve as the input to Model II, and it applies heuristics to select, from among the possible phonetic strings, one that will maintain an articulatory match at the comparator.

A virtue of the analysis-by-synthesis model is that its components have explicit functions, even though some of these component units are bound to be rather complicated devices. The comparator, explicit here, is implicit in a neural network model in the sense that some neural nets will be aroused—and others will not—on the basis of degree of similarity between the firing patterns of the selected nets and the incoming pattern of neural excitation. Comparisons and decisions of this kind may control the spread of excitation throughout all levels of the neural mechanism, just as a sophisticated guessing game is used by the analysis-by-synthesis model to work its way, stage by stage, to a phonetic representation—and presumably on upstream to consciousness. In short, the two models differ substantially in the kinds of machinery they invoke and the degree of explicitness that this allows in setting forth the underlying philosophy: they differ very little in the reliance they put on the mechanisms of production to do most of the work of perception.

The general point of view of analysis-by-synthesis is incorporated in the constructivist view of cognitive processes in general, with speech perception as an interesting special case. Thus, Neisser [1967, p. 10] in the introduction to *Cognitive Psychology,* says, "The central assertion is that seeing, hearing, and remembering are all acts of construction, which may make more or less use of stimulus information depending on circumstances. The constructive processes are assumed to have two stages, of which the first is fast, crude, wholistic, and parallel while the second is deliberate, attentive, detailed, and sequential."

It seems difficult to come to grips with the specific mechanisms (and their functions) that the constructivists would use in dealing with spoken language to make the total perceptual process operate. A significant feature, though, is the assumption of a two-stage process, with the constructive act initiated on the basis of rather crude information. In this, it differs from both of the models that we have thus far considered. Either model can, if need be, tolerate input data that are somewhat rough and noisy, but both are designed to work best with "clean" data, since they operate first on the detailed structure of the input and then proceed stepwise toward a more global form of the message.

Stevens and House (1972) have proposed a model for speech perception that is, however, much closer to the constructivist view of the process than was the early analysis-by-synthesis model of Figure 5. It assumes that spoken language has evolved in such a way as to use auditory distinctions and attributes that are well matched to optimal performances of the speech generating mechanism; also, that the adult listener has command of a catalog of correspondences between the auditory attributes and the articulatory gestures (of approximately syllabic length) that give rise to them when he is a speaker. Hence, the listener can, by consulting his catalog, infer the speaker's gestures. However. some further analysis is needed to arrive at the phonological features, although their correspondence with articulatory events will often be quite close. In any case, this further analysis allows the "construction" (by a control unit) of a tentative hypothesis about the sequence of linguistic units and the constituent structure of the utterance. The hypothesis, plus the generative rules possessed by every speaker of the language, can then yield an articulatory version of the utterance. In perception, actual articulation is suppressed but the information about it goes to a comparator, where it is matched against the articulation inferred from the incoming speech. If both versions match, the hypothesized utterance is confirmed; if not, the resulting error signal guides the control unit in modifying the hypothesis. Clearly, this model employs analysis-by-synthesis principles. It differs from earlier models mainly in the degree of autonomy that the control unit has in constructing hypothesis and in the linguistic level and length of utterance that are involved.

The approach to speech perception taken by Chomsky and Halle [1968] also invokes analysis-by-synthesis, with even more autonomy in the construction of hypotheses; thus, "We might suppose . . . that a correct description of perceptual processes would be something like this. The hearer makes use of certain cues and certain expectations to determine the syntactic structure and semantic content of an utterance. Given a hypothesis as to its syntactic structure—in particular its surface structure—he uses the phonological principles that he controls to determine a phonetic shape. The hypothesis will then be accepted if it is not too radically at variance with the acoustic material, where the range of permitted discrepancy may vary widely with conditions and many individual factors. Given acceptance of such a hypothesis, what the hearer 'hears' is what is internally generated by the rules. That is, he will 'hear' the phonetic shape determined by the postulated syntactic structure and the internalized rules." This carries the idea of analysis-by-synthesis in constructivist form almost to the point of saying that only the grosser cues and expectations are needed for *perfect* reception of the message (as the listener would have said it), unless there is a gross mismatch with the input information, which is otherwise largely ignored. This extension is made explicit with respect to the perception of stress. Mechanisms are not provided, but they would not be expected in a rule-oriented account.

In all the above approaches, the complexities inherent in the acoustic signal are dealt with indirectly rather than by postulating a second mechanism (at least as complex as the production machinery) to perform a straight-forward auditory analysis of the spoken message. Nevertheless, *some* analysis is needed to provide neural signals from the auditory system for use in generating hypotheses and in error comparisons at an appropriate stage of the production process. Obviously, the need for analysis will be least if the comparisons are made as far down in the production process as possible. It may be, though, that direct auditory analysis plays a larger role. Stevens [1971] has postulated that the analysis is done (by auditory property detectors) in terms of acoustic features that qualify as distinctive features of the language, since they are both inherently distinctive and directly related to stable articulatory states. Such an auditory analysis might not yield complete information about the phonological features of running speech, but enough, nevertheless, to activate analysis-by-synthesis operations. Comparisons could then guide the listener to self-generation of the correct message. Perhaps Stevens will give us an expanded account of this view of speech perception in his discussion of the present paper.

All these models for perception, despite their differences, have in common a listener who actively participates in producing speech as well as in listening to it in order that he may compare his internal utterances with the incoming one. It may be that the comparators are the functional component of central interest in using any of these models to understand how reading is done by adults and how it is learned by children. The level (or levels) at which comparisons are made—hence, the size and kind of unit compared—determines how far the analysis of auditory (and visual) information has to be carried, what must be held in short-term memory, and what units of the child's spoken language he is aware of—or can be taught to be aware of—in relating them to visual entities.

Can we guess what these units might be, or at least what upper and lower bounds would be consistent with the above models of the speech process? It is the production side of the total process to which attention would turn most naturally, given the primacy ascribed to it in all that has been said thus far. We have noted that the final representation of the message, before it leaves the central nervous system on its way to the muscles, is an array of features and a corresponding (or derived) pattern of neural commands to the articulators. Thus, the features would appear to be the smallest units of production that are readily available for comparison with units derived from auditory analysis. But we noted also that smoothly flowing articulation requires a restructuring of *groups* of features into syllable-size or word-size units, hence, these might serve instead as the units for comparison. In either case, the lower bound on duration would approximate that of a syllable.

The upper bound may well be set by auditory rather than productive processes. Not only would more sophisticated auditory analysis be required to match higher levels—and longer strings—of the message as represented in production, but also the demands on short-term memory capacity would increase. The latter alone could be decisive, since the information rate that is needed to specify the acoustic signal is very high—indeed, so high that some kind of auditory processing must be done to allow the storage of even word-length stretches. Thus, we would guess that the capacity of short-term memory for purely auditory forms of the speech signal would set an upper bound on duration hardly greater than that of words or short phrases. The limits, *after conversion to linguistic form,* are however substantially longer, as they would have to be for effective communication.

Intuitively, these minimal units seem about right: words, syllables, or short phrases seem to be what we say, and hear ourselves saying, when we talk. Moreover, awareness of these as minimal units is consistent with the reference-to-production models we have been considering, since all of production that lies below the first comparator has been turned over to bone-and-muscle mechanisms (aided, perhaps, by gamma-efferent feedback) and so is inaccessible in any direct way to the neural mechanisms responsible for awareness. As adults, we know how to "analyze" speech into still smaller (phonetic) segments, but this is an acquired skill and not one to be expected of the young child.

Can it be that the child's level of awareness of minimal units in speech is part of his problem in learning to read? Words should pose no serious problem so long as the total inventory remains small and the visual symbols are sufficiently dissimilar. But phonic methods, to help him deal with a larger vocabulary, may be assuming an awareness that he does not have of the phonetic segments of speech, especially his own speech. If so, perhaps learning to read comes second to learning to speak and listen *with awareness.* This is a view that Mattingly will, I believe, develop in depth. It can serve here as an example of the potential utility of models of the speech process in providing insights into relationships between speech and learning to read.

In Conclusion

The emphasis here has been on the processes of speaking and listening as integral parts of the total process of communicating by spoken language. This concentration on speech reflects both its role as a counterpart to reading and its accessibility via experimentation. The latter point has not been exploited in the present account, but it is nonetheless important as a reason for focusing on this aspect of language. Most of the unit processors that were attributed to speech in the models we have been discussing can, indeed, be probed experimentally: thus, with respect to the production of speech, electromyography and cinefluorography have much to say about how the articulators are moved into the observed configurations, and sound spectrograms give highly detailed accounts of the dynamics of articulation and acoustic excitation; examples with respect to speech perception include the use of synthetic speech in discovering the acoustic cues inherent in speech, and of dichotic methods for evading peripheral effects in order to overload the central processor and so to study its operation. Several of the papers to follow will deal with comparable methods for studying visual information processing. Perhaps the emphasis given here to processes and to the interdependence of perception and production will provide a useful basis for considering the linkages between reading and speech.

References

Chomsky, N., 1957. *Syntactic Structures.* The Hague: Mouton.

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

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

Chomsky, N., and G. A. Miller, 1963. Introduction to the formal analysis of natural languages. In *Handbook of Mathematical Psychology,* R. D. Luce, R. R. Bush, and E. Galanter (eds.), New York: Wiley.

Coffey, J. L., 1963. The development and evaluation of the Battelle Aural Reading Device. In *Proceedings of the International Congress on Technology and Blindness,* New York: American Foundation for the Blind.

Cooper, F. S., 1950. Research on reading machines for the blind. In *Blindness: Modern Approaches to the Unseen Environment,* P. A. Zahl (ed.), Princeton: Princeton University Press.

Cooper, F. S., A. M. Liberman, and J. M. Borst, 1951. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proc. Nat. Acad. Sci.* 37:318–328.

Delattre, P. C., A. M. Liberman, and F. S. Cooper, 1955. Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Amer.* 27:769–773.

Fant, C. G. M., 1960. *Acoustic Theory of Speech Production.* The Hague: Mouton.

Flanagan, J. L., 1965. *Speech Analysis Synthesis and Perception.* New York: Academic Press.

Halle, M., and K. N. Stevens, 1964. Speech recognition: A model and a program for research. *IRE Trans. Info. Theory,* 1962, *IT-8,* 155–59. Also in *The Structure of Language,* J. A. Fodor and J. J. Katz (eds.), Englewood Cliffs, N.J.: Prentice-Hall.

Jakobson, R., C. G. M. Fant, and M. Halle, 1963. *Preliminaries to Speech Analysis.* Cambridge, Mass.: M.I.T. Press.

Liberman, A. M., 1957. Some results of research on speech perception. *J. Acoust. Soc. Amer.* 29:117–123.

———, 1970. The grammars of speech and language. *Cogn. Psych.* 1:301–323.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, 1967. Perception of the speech code. *Psych. Rev.* 74:431–461.

Liberman, A. M., F. S. Cooper, M. Studdert-Kennedy, K. S. Harris, and D. P. Shankweiler, 1968. On the efficiency of speech sounds. *Z. Phonetik, Sprachwissenschaft u. Kommunikationsforschung* 21:21–32.

Miller, G. A., 1956. The magical number seven, plus or minus two, or, some limits on our capacity for processing information. *Psych. Rev.* 63:81–96.

Neisser, U., 1967. *Cognitive Psychology.* New York: Appleton-Century-Crofts.

Orr, D. B., H. L. Friedman, and J. C. C. Williams, 1965. Trainability of listening comprehension of speeded discourse. *J. Ed. Psych.* 56:148–156.

Stevens, K. N., 1960. Toward a model for speech recognition. *J. Acoust. Soc. Amer.* 32:47–55.

———, 1971. Perception of phonetic segments: Evidence from phonology, acoustics, and psychoacoustics. In *Perception of Language,* D. L. Horton and J. J. Jenkins (eds.), Columbus, Ohio: Merrill.

Stevens, K. N., and M. Halle, 1967. Remarks on analysis by synthesis and distinctive features. In *Models for the Perception of Speech and Visual Form,* W. Wathen-Dunn (ed.), Cambridge, Mass.: M.I.T. Press.

Stevens, K. N., and A. S. House, 1972. Speech perception. In *Foundations of Modern Auditory Theory,* Vol. 2, J. Tobias (ed.), New York: Academic Press.

Studdert-Kennedy, M., and F. S. Cooper, 1966. High-performance reading machines for the blind: Psychological problems, technological problems, and status. In *Proceedings of the International Conference on Sensory Devices for the Blind,* R. Dufton (ed.), London: St. Dunstan's.