Reprinted from COGNITIVE PSYCHOLOGY, Volume 1, Number 4, October 1970
Copyright © 1970 by Academic Press, Inc. *Printed in U. S. A.*

The Grammars of Speech and Language[1]

A. M. LIBERMAN[2]

Haskins Laboratories

The conversion between phonetic message and acoustic signal, *i.e.*, speech, is a grammatical code, similar in interesting ways to syntax and phonology. Being more accessible to experiment, speech should, therefore, be an inviting object of study for those interested in the psychology of grammar. Experiments on speech have already provided some information about the psychological processes associated with the use of grammatical codes.

The implication of the title, and the point of what follows, is that there is a grammar of consonants and vowels no different in principle from the grammar of sentences. More generally, I mean to say that the production and perception of speech sounds is as much a part of language as syntax and often more revealing to those who would understand linguistic processes.

In the conventional wisdom, consonants and vowels are not so highly thought of. To the linguist, these elements serve primarily as a concrete base for abstract concepts. The psychologist seems to find them even less interesting. To him, the sounds of speech are no more than convenient vehicles, much like the letters of the alphabet. They carry linguistic information, but their connection to language does not appear to be organic; they are, therefore, not usually thought to have much to do with psycholinguistics. According to these fairly common views, then, language and its psychology are to be found only at the higher levels; there they enjoy an unspeakably abstract existence, forever safe from the rude interventions of the experimental scientist.

[1] Invited research address, sponsored by Division 3, at the meeting of the American Psychological Association, September 1, 1969, in Washington, D. C. The research on which this talk is based has been supported by the National Institute of Child Health and Human Development, the National Institute of Dental Research, and the Office of Naval Research.

This talk is my interpretation of the research and ideas of my colleagues at the Haskins Laboratories. It derives, in particular, from two papers that have been written recently by members of that institution: Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; and Mattingly and Liberman, 1969.

[2] Also, Department of Psychology, University of Connecticut; Department of Linguistics, Yale University.

GRAMMAR AND SPEECH: COMPLEX CODE OR SIMPLE CIPHER?

Something like the usual view of speech and language is displayed in Fig. 1, where I have tried to show, in the simplest way, what grammar is supposed to be, and how it is that speech lies outside it. My aim is also to provide some basis for the comparisons between speech and language that I will make later in this paper. For those purposes, but not, I hope, for the sake of argument, I have taken an overall view of language similar to that proposed by Chomsky and his colleagues (Chomsky, 1957, 1965; Chomsky & Miller, 1963). But it does not really matter for our purposes exactly which view we take. The point is that language is commonly seen as a structure that contains successive recodings of the linguistic information; accordingly, the structure is arranged in several levels.

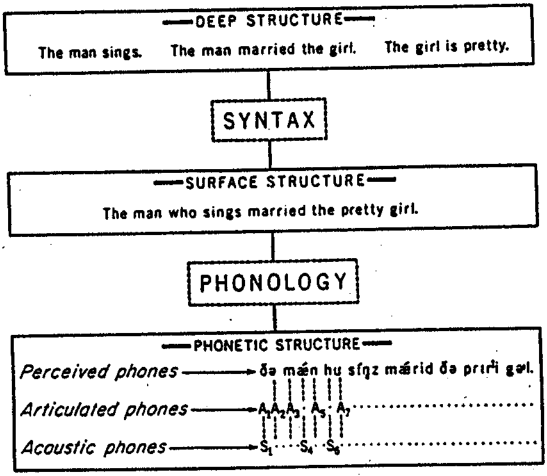

FIG. 1. An illustration of the assumption that the sounds of speech are only an alphabet on the phonetic structure, not another level of language arrived at by grammatical recoding.

In the Chomsky scheme, there are three levels. At the most basic of these, called deep structure, lie meaningful but highly inaccessible segments known as morphemes. You will understand that I have put aside all questions about how abstractly these morphemes are to be represented, and how divided into lexical and grammatical types. I have, instead, simply displayed the deep-structure information, in plain English, as three sentences: *the man sings*; *the man married the girl*; *the girl is pretty.*

The segments at each level are not simply arranged in a linear string but are rather organized into larger units. In Fig. 2 I have used the familiar tree diagram to show how the segments at the deep level are grouped into

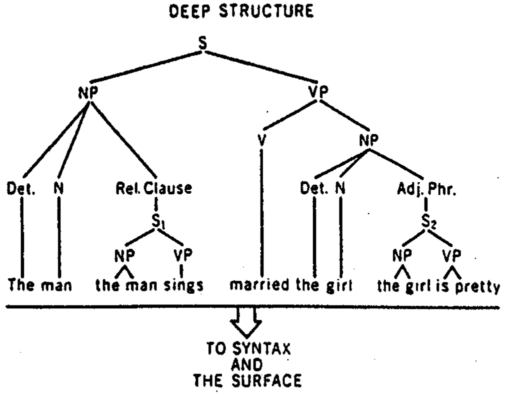

FIG. 2. A tree diagram showing how the segments at one level (deep structure) are organized into larger units.

units larger than the word but smaller than the sentence. I do this for the sake of completeness, and also because I will want later to make comparisons with the way consonants and vowels are organized. For the moment, however, I will simply take note of the fact that there is an organization of the segments at each level; having done that, I will return to the matter of recoding between levels.

You know that we do not ordinarily communicate in deep-structure strings, even though—or, perhaps, because—these are closest to the meaning the speaker intends. Normally, we first recode the message by putting it through the syntactic component, which nests, rearranges, and deletes the unit of the deep structure to produce a new string at the next level down. There, in the surface structure, so called, our three sentences have become one: *the man who sings married the pretty girl.* We are now a greater distance from the true meaning of the sentence, and also from its most inaccessible representation, but still a long way from superficial reality. For the segments at the surface level, roughly equivalent to what were once called phonemes, can still only be spoken of; they cannot be spoken—not, in any case, until they have been recoded once more, this time by rules belonging to the phonological component. Below, comparing speech and grammar, I will omit phonology and speak only about syntax. The comparisons work equally well with phonology, but I want to avoid repetition. I will, therefore, merely note at this point that the segments at the surface must be processed by the phonology before we arrive at the phonetic string, the very end of the whole process.

So far, then, we have three streams of information—deep, surface, and phonetic—related by two conversions, one syntactic, the other phonological. The two conversions, together with the organization of the segments at each level, are said to constitute grammar. I should emphasize about the conversions that neither is a trivial encipherment. Both are true recodings in that they produce a complex reorganization of the information. As a result, the units at one level do not correspond to the units at the next, either in structure or in number.

Concerning the exact nature of grammar, opinion among linguists is, to say the least, divided. But all would agree that grammar is the organization and reorganization of the linguistic segments. They would agree, further, that the grammatical reorganizations end with the phonetic string, the bottom-most box in Fig. 1. All that remains, in any instance, is to decide how to represent those elements. Should it be in terms of perceptual entities, articulatory movements, or sounds? But this is not thought to be an interesting question because it is assumed that these three representations are related in the simplest possible way.

Let me be more explicit. The elements in the first line of the phonetic level are perceptual entities. These "phones," so called, are what we hear when we listen to an utterance like "bag" and know that it consists of three segments—that it differs from "sag" in the first segment, from "big" in the second, and from "bad" in the third. The phones are signalled by the sounds of speech, shown in the third line, and it is usually assumed that the signalling is done as simply as possible, each perceived phone being represented, alphabetically, by a unit sound.

The conversion between phone and sound is thus seen, not as a complex grammatical recoding, but as a simple substitution or encipherment. The linguist needs no elaborate formal apparatus to handle that state of affairs; and the psychologist sees no interesting implications for cognitive processes, since all that is required of the speaker-listener is that he distinguish the unit sounds and associate each one with the name of the appropriate phone.

This is the simple and rather common view of speech that I referred to above. If it were the correct view, then the study of speech would, indeed, include no grammar and only as much psychology as is contained in auditory psychophysics, plus a little paired-associate learning.

But the simple view of speech is wrong, and I will reject it in favor of the more complex view we see in Fig. 3. Here I have meant to represent, though, as yet, only in the most general form, the assertion I made at the start: that the interconversion of phonetic segment and sound is a grammatical recoding, similar in complexity and form to syntax and phonology. I will try to show that this is so and, more generally, that speech is a truly integral part of language, not merely a convenient vehicle for transmitting it.

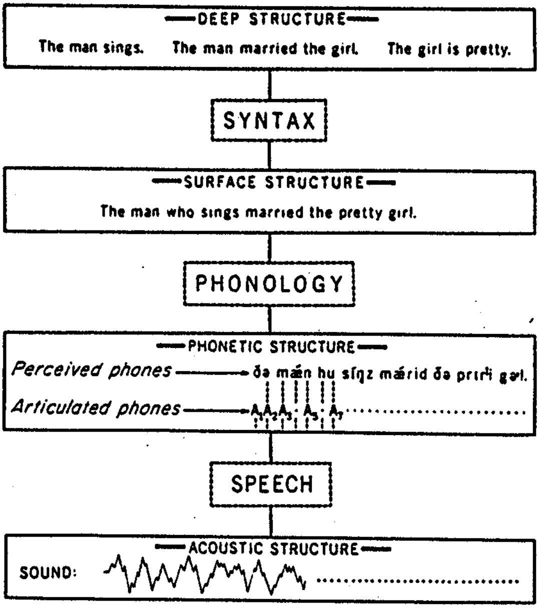

FIG. 3. An illustration of the assumption that the sounds of speech are a separate level of language, connected to the phonetic structure by a grammatical recoding similar to syntax and phonology.

WHY WE SHOULD SUSPECT THAT SPEECH IS NOT A SIMPLE CIPHER

The first step toward a proper understanding of speech is to come to respect it, and to see, apart from research on the speech signal itself, why the simple view is likely to be wrong. Consider that speech is, after all, the only universal vehicle of language. But why, if the sounds of speech are merely a cipher on the phones, should this be so? As a sensory channel, the ear is, by almost any standard, poorer than the eye. Why, then, is it so very difficult to substitute the optical shapes of an alphabet for the acoustical shapes of speech? And why, in particular, should this be even more difficult in the case of congenitally deaf children?

We can hardly suppose that the primacy of speech depends on the absence of earlids or on our ability to hear speech no matter how our heads are turned. Nor does it rest, more generally, on any properties of sound as a medium of communication. We have got to look elsewhere if we are to account for the advantages of speech because we know that sounds other than speech cannot be made to convey language well. That knowledge comes from 55 years of trying to devise nonspeech sounds for use in reading machines for the blind—that is, devices that scan the print and convert it into intelligible sounds. In spite of the most diligent efforts in connection with the development of those machines, no nonspeech acoustic alphabet has yet been contrived that can be made to work more than one-tenth as well as speech.[3]

To understand the difficulty with the nonspeech signals of the reading machines, we have only to consider that in listening to speech, people can perceive as many as 25 or 30 phonetic segments per second, and then to note that 25 or 30 acoustic segments per second would far overreach the resolving power of the ear. If speech were a simple alphabet or cipher on the phonetic message, listeners would often be wholly unable even to separate the individual elements, let alone identify them.

In any case, we have reason to wonder, not only why sound is the uniquely natural vehicle of language, but, even more, why only one set of sounds—those of speech—will work well. The reasons for this can hardly be obvious or superficial. Indeed, they are, as we will see, well hidden, and they go to the heart of man's capacity for language.

HOW WE KNOW THAT SPEECH IS A COMPLEX CODE

To show, now, what is special about the sounds of speech and how intimate is their connection with language, I will draw primarily on the work of my colleagues at the Haskins Laboratories. I will be concerned first, as we were at Haskins, with the perception of speech, though, as you will see, we cannot avoid production very long. For those who would understand the perception of speech, the first task is that which confronts the scientist who sets out to study the perception of anything. It is, of course, to find the cues, the physical stimuli that govern the perception. For that purpose, we built a device that converts hand-painted spectrograms into sound. This provided the basis for what proved to be a convenient method of experimenting with the speech signal: it made it possible to vary those parameters we guessed to be of linguistic importance and then to hear the effects. Thousands of experimental trials later, we had succeeded in identifying most of the acoustic cues for the phonetic segments.[4]

To show in what way speech is a psychologically interesting grammar, I will present a simple example to display the general characteristics of the

[3] The substance of this section of the paper is developed in somewhat greater detail, and with appropriate references, in a recent review (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

[4] Results of research on the acoustic cues for speech perception, together with references to the primary sources, are to be found in Kozhevnikov and Chistovich (1965), Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967), and Stevens and House (in press).

relation between speech sounds and the phonetic message they convey. In Fig. 4 are drastically simplified spectrographic patterns that will, when converted to sound, produce approximations to the syllables [di] and [du]. These patterns do not contain all the cues—hence they are less intelligible than synthetic speech can be—but they fairly exemplify the characteristics I want to talk about.

Each pattern of Fig. 4 consists of two bands of acoustic energy called"formants." At the left, or beginning, of each pattern the formants move rapidly through a range of frequencies. These rapid movements, which consume about 50 msec, are called "transitions." Following the transitions, the formants assume a steady state.

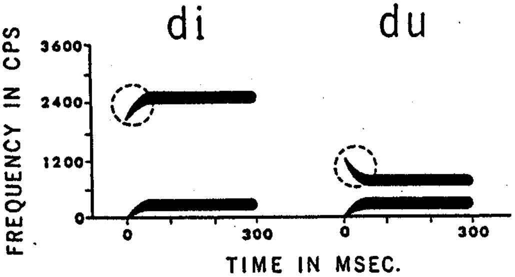

FIG. 4. Simplified spectrographic patterns sufficient to produce the syllables [di] and [du].

The steady-state formants are, by their different positions on the frequency scale, the cues for the vowels [i] and [u]. We can see that for these vowels, as for all others that are of long duration, the relation between acoustic cue and phonetic segment is a simple, alphabetic one. There is a straightforward correspondence in segmentation, in that we can isolate a piece of sound—in this case, the steady-state portion of the pattern—that is, by itself, sufficient to produce a single phonetic segment [i] or [u]. It is also true of these steady-state vowels, though not apparent from the patterns in the figure, that the acoustic cues are the same regardless of context. If all speech were like these vowels, we should have no reason to reject the common assumptions I described earlier; speech would be as simple as most people think it is.

But consider now the stop consonant [d]. To isolate the acoustic cue for that segment, we should first notice the transition of the lower (first) formant. That transition is not specifically a cue for [d]; it rather tells the listener that the segment is one of the voiced stops, [b], [d], or [g]. In all further discussion I will ignore it and talk only about the cue that causes us to know that the phone is [d], not [b] or [g]. To produce [d], instead of [b], or [g], we must add the transitions of the higher (second) formant, the parts of the pattern that are encircled by the dashed lines. These, then, are the acoustic cues for [d].

Perhaps the most obvious characteristics of the cues for [d] is that they are vastly different with the two vowels, [i] and [u]. The one that produces [d] with [i] is an upgoing frequency modulation high in the spectrum; the other one, before [u], is a downgoing modulation low in the spectrum. But in immediate perception, one hears the consonant in its canonical form and is totally unaware of the context-conditioned variation.

A second and less obvious characteristic of the relation between cue and consonant is that there is no segment in the sound stream corresponding to the segment in the phonetic message. That is, there is no way to cut the sound pattern so as to obtain a piece that will produce the phone [d] without also producing the next vowel or some reduced approximation to it, such as the neutral vowel in [də]. Yet listeners do separate the utterance into two distinct segments.

We have seen now two characteristics of the complex relatior etween acoustic cue and stop consonant: the context-conditioned variatıon in the cue and the lack of correspondence in segmentation. Before going on, I should note that both characteristics are found with all the consonants except, perhaps, the fricatives. But the point I would now emphasize is that these characteristics derive from a single one that is more basic than either, namely, that one and the same acoustic cue, such as the second-formant transition, serves more than one phonetic segment. This is a kind of parallel transmission. In the case of our example, we see in what way it occurs by observing that the second-formant transitions are, at every instant, carrying information simultaneously about two successive segments, the stop consonant and the vowel. It follows from this that the transition cue for the consonant must be different with the different vowels, and that there cannot be an acoustic segment corresponding only to the single phone [d]. Thus, we see that the segments in the acoustic signal do not correspond to the segments in the phonetic message, either in structure or in number. In that sense we should say of the acoustic signal that it is not a trivial alphabet on the phonetic message but a complex code.

It is easy to see what this code has to do with the unique efficiency of speech, about which I remarked earlier. Speech evades the limitations set by the temporal resolving power of the ear because it transmits information about successive segments simultaneously, and on the same acoustic cue. Such parallel transmission reduces by a significant factor the number of acoustic segments that must be perceived per unit time. If the phonetic segments are encoded acoustically into units of, say, syllabic size, then the rate at which the sounds of speech merge into a buzz is set by the number of syllables per second, not by the number of phones. But this very great advantage is bought at considerable cost: the relation between acoustic signal and phonetic message is now necessarily a complexly encoded one.

PARALLEL TRANSMISSION: THE SPEECH CODE AND THE SYNTACTIC CODE

Because phonetic segments are linked to sound by a complex but efficient code, we should suppose that speech is of some special interest in its own right. But speech is surely of greater interest if the code and its attendant devices are like the codes and devices of grammar. Let us, then, look again at the speech code and see how like grammar it is.

The formal resemblance of speech to grammar does not end with the fact that both are complex codes. Consider more particularly what we have seen as a most important characteristic of the speech code: that information about successive segments of the message is carried simultaneously by the same part of the signal.

In Fig. 5 I have shown such parallel transmission in a case just a little

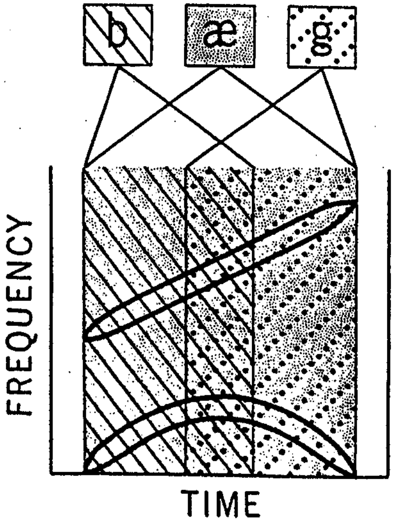

FIG. 5. Parallel transmission of phonetic segments after encoding (by the rules of speech) to the level of sound.

more complicated, and a good deal closer to real life, than the [di] and [du] of Fig. 4. At the top are the phones that form the syllable "bag," at the bottom a schematic spectrogram sufficient to produce that signal or a reasonable approximation to it. The purpose of the figure is to show how the discrete and well-ordered segments at the phonetic level overlap and intermix in the sound stream. Looking at the vowel [æ], we see that it extends the whole length of the acoustic signal. That means that if the vowel were different—if the syllable were [bɔg] or [bcg] instead of [bæg]—then the second formant would, in fact, be different from beginning to end, not just somewhere in the middle. The stop consonant, [b], at the start of the syllable, overlaps with both of the other segments, extending to a point past the middle of the signal. That means simply that if the initial stop were different—if it were, say, [g] instead of [b] (that is, [gæg] instead of [bæg])—then the second formant would be different through the whole length of the section indicated by the diagonal striping. It is of interest, and quite typical of real speech, that there is a region in the middle of this signal in which, as shown, information about all three segments is being delivered simultaneously and on the same piece of sound.

In Fig. 6 I've shown how parallel transmission occurs also in the syntactic conversion between deep and surface structure. At the top are the three deep-structure sentences we saw in an earlier slide, arranged now somewhat artificially into ordered segments. Below them is the single sentence they form after they have been processed by the syntactic component of the grammar. I have here borrowed the graphic device used in the case of [bæg] in order to show over what part of the surface sentence each of the deep-structure segments extends. One of the sentences, "the man married the girl," is spread throughout the length of the single sentence at the surface. Like the vowel in [bæg], it is capable of standing alone; like the stops in that same syllable, the other two sentences exist (at the

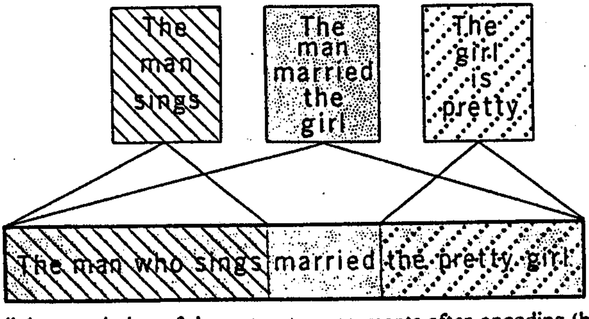

FIG. 6. Parallel transmission of deep-structure segments after encoding (by the rules of syntax) to the level of surface structure.

surface) only as they are encoded into the vowel-like nucleus.

Thus, parallel transmission is as much a characteristic of syntax as of speech. Moreover, it achieves the same purpose and at the same cost: the message is delivered more rapidly but in a highly encoded form.

RATIONALIZING THE SPEECH CODE: THE GRAMMAR OF SPEECH

There are other formal resemblances between speech and the higher levels of language, though none is more basic, I think, than the one I just described. I would return, however, to the very considerable complexity of the speech code, about which I have spoken several times, and consider the problem that such complexity must present to the perceiver. Then I will make such comparisons as I can to the higher levels.

First, to emphasize that the problem of decoding speech is a real one, I would contrast two machines: one designed to perceive print, the other speech. As you know, engineers have succeeded in developing print readers that work quite well, but no one has built a machine that will perceive speech. This difference in achievement directly reflects a great difference in the difficulty of the task. For a machine, print is easy largely because print is a simple alphabet; speech is very hard, because it is a complex code. We ought surely to find it of some psychological interest that for us human beings, the difficulty is just the other way around; the complex speech code is far easier than the simple printed alphabet.

All that is to say that to anyone who would design a speech perceiver, the relation between sound and phonetic message will appear complex, opaque, and eccentric. Yet to every human being who listens to speech, that same relation is simple, transparent, and regular. We should suppose, then, that all speakers of the language have readily available to them a key to the code, a model that rationalizes the otherwise arbitrary connection between the sounds of speech and the phonetic message they convey.

To describe the shape that such a model might take, we must now abruptly shift our attention from perception to production, for the only way to rationalize the code is to describe the rules by which we might arrive at the sound, given the phonetic message. Such rules do not work in the reverse direction, from sound back to phonetic message, and we are unable to find any that do.

To understand the production model, we must first introduce the concepts of syllable and feature. It seems intuitively clear that the string of phones in the word "bagdad" is organized into two syllables, [bæg] and [dæd]. This is very much like saying, as I did earlier when I spoke about grammar, that the strings of words at the higher levels are organized into phrases. To discover the organization of the segments at any of these levels requires initially only paper, pencil, and an intact intuition. But in

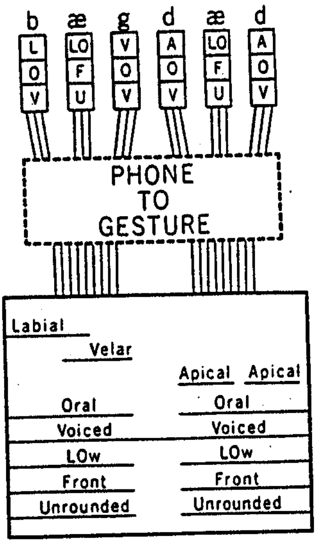

FIG. 7. Schematic diagram showing how phones are organized into syllables by an overlapping of the articulatory features which the phones comprise.

the phonetic case, we can, I think, begin to account for the intuition.

Note, first, that, as I have tried to show in Fig. 7, each phone in the string consists of several features.[5] Take [b] as an example. It has the feature "labial," which means that the closure is made by the lips at the front of the vocal tract, instead of being made by the tip of the tongue at the alveolar ridge, as in the case of [d], or by the back of the tongue in the back of the tract, as in [g]. It also has the feature "oral," which means that the velum—the passage from throat to nose—is closed; it would be open in the case of the nasal counterpart, [m]. And, finally, [b] has the feature "voiced," which means that the vocal cords begin to vibrate simultaneously with the opening of the vocal tract, instead of being delayed about 50 msec, as in the voiceless counterpart, [p].

These features are very real in perception and production and also very independent. I will simply assert that now as an assumption and say that it is strongly supported by many experimental data (Miller & Nicely, 1955; Harris, Lysaught, & Schvey, 1965; Wickelgren, 1966; Klatt, 1968; Halwes, to be published; Studdert-Kennedy & Shankweiler, to be

[5] The term "feature," as it is used here, is considerably narrower and also more concrete than the "distinctive features" of Jakobson and his colleagues (Jakobson, Fant, & Halle, 1952; Jakobson & Halle, 1956). It is nonetheless appropriate to acknowledge indebtedness for the general idea, which is, as will be seen in this paper, central to the speech code and its efficiency.

published). I have, then, supposed that each feature is represented by unitary nervous activity at a relatively high level of the production system.

In the lower half of the figure are horizontal lines that are intended to show two other assumptions that have, even now, some basis in fact. One is that each phonetic feature is represented in articulation by a characteristic muscle gesture (Harris, Lysaught, & Schvey, 1965). Indeed, when I described the features a moment ago, it was in terms of just such gestures. The second assumption is that these component gestures are to a large extent independent and can occur at the same time (Harris, Lysaught, & Schvey, 1965; Fromkin, 1966; Cooper, 1966). Thus, the lip gesture associated with the initial phone [b] can be coarticulated with the tongue gesture appropriate for the vowel [æ].

Between the features in the top half and their articulatory representations in the bottom, is a box, labeled "phone" to "gesture." It has responsibility for timing the articulatory gestures so as to achieve the best possible overlap.

We see that the features of the first and second consonants, [b] and [g], are overlapped with the features of the first vowel. Features belonging to the third and fourth consonants, both of which are [d], are similarly overlapped with the features belonging to the second vowel. Thus, the features that constitute the segments are organized in production into two syllabic bundles, each consisting of overlapped and largely independent articulatory components. But, as in the organization of words into phrases, the formation of the syllable requires only a grouping of the units. There is no encoding any more than there is an encoding in the organization of words into phrases.

The encoding occurs at a later stage, in the conversion from muscle gesture to sound, as I have tried to show in Fig. 8. To provide continuity, I have reproduced the first syllable of the last slide. The general point to be made is that the numerous dimensions at the gesture level are mixed and also greatly reduced in the conversion to sound. As a result, information which was represented independently by the overlapping action of separate muscles is now completely merged. Thus, at the articulatory level, the movement of the lips associated with the consonant [b] is completely overlapped with most aspects of the tongue gesture appropriate for the vowel [æ]. In the conversion to sound the independence of these articulatory features is lost. The two are represented simultaneously and on the same parts of the second formant. They have now been encoded.

Thus we see, at least in the most general terms, how the relation between sound and phone can be rationalized. Details were omitted by the score and, indeed, we don't know all the details, including some very

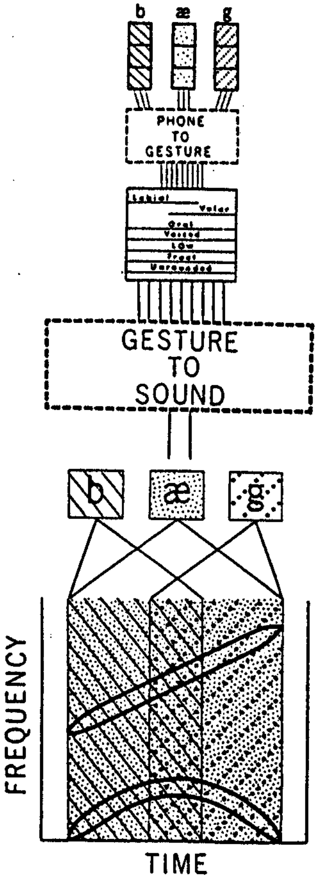

FIG. 8. Schematic diagram showing how the overlap of articulatory features produces encoding in the conversion to sound.

important ones. The claim is only that an articulatory model like the one I described can, in principle, account for the complex relation between phonetic segment and sound.

To return now to the matter of speech and language, I would suggest that the model I just outlined is, quite precisely, a grammar—indeed, a generative grammar—and is to be evaluated as such. It is the proper grammar only if, when fully developed, it most economically accounts for the observed relation between the two streams of information, phonetic and acoustic, that it links. If it is the proper grammar, then it tells us about the intuitive knowledge that is commanded by every person who perceives speech. It does not tell us where that knowledge came from or how it is put into effect. From the standpoint of a psychologist interested in speech perception, our grammar of speech is, at most, a key to the code; it is not a decoder.

RATIONALIZING SYNTAX: THE GRAMMAR OF SENTENCES

At the higher levels of language, we should suppose that the encoded relation between, say, deep and surface structure must also appear eccentric and arbitrary, except to a person who possesses the key. And the key is, again, a model or grammar—a set of rules that rationalizes the code. It has been the business of generative linguists to devise exactly such models.

We should note that the linguist's generative grammar works in one direction only. That is, the grammatical rules enable us to derive the surface structure, for example, when we are given the deep structure as input; the rules won't work in the reverse direction. The grammars of syntax and phonology tell us about the intuitive knowledge that is commanded by every person who perceives language. But they don't tell us where he got that knowledge or how he makes use of it. The grammars are keys to their respective codes; they are not the decoders. You will, I hope, have recognized these statements about the generative grammars of syntax and phonology, if only because they are the same as the statements I made about speech.

To complete the grammar of speech, we will need many hard facts about articulation and about the relation between muscular contraction and vocal-tract shape. Most of the necessary data are unavailable now, but progress in digging them out is being made in several laboratories, including our own. When we have those facts, we will have a proper grammar of speech. That grammar will still be different from the higher grammars, however, in that it will be written in the language of physiology, not in terms of an abstract algebra. That will be a significant step forward for those of us who would understand the mechanisms of language and see them in biological perspective.

ARTICULATION AND THE DECODING OF SPEECH

Although the articulatory grammar I outlined makes some sense of the speech code, it does not yield the design of a device that would decode it. However, we have clues about the decoder, and these prove to be of interest in relation to the grammar.

Consider, again, the syllables [di] and [du] (Fig. 4) and notice, again, how very different are the transition cues for [d], the parts of the patterns encircled by the dashed lines. If you have been saying these syllables quietly to yourselves, you will already have noticed that there is a common element in their production. In both cases the tip of the tongue is moved to a point at the alveolar ridge, so as to close the vocal tract, and then withdrawn. We have just seen in the articulatory model how two such similar gestures can have two such different acoustic consequences. In any case, we have here only one example of a very common occurrence in our results. We often find, as part of the encoding of speech, that the relation between articulatory gesture and sound is very complex; in all these cases perception relates more simply to gesture. It is as if the decoder took the listener back to the neural events that control the articulatory features and thus, if you will, to the speaker's intent. I should guess that the encoder and decoder are somehow related, that man does not have two completely separate mechanisms, one for producing and one for perceiving, but rather a more nearly unitary device that behaves, in all of its operations, as if it somehow embodied the grammar of speech.

EXPERIMENTING WITH SPEECH TO LEARN ABOUT GRAMMATICAL DECODING

If what I have said so far is true, then, as experimental psychologists interested in language, we might feel moved to look more favorably on speech. It is, as we have already seen, more accessible to experiment than the other grammars. But if it is nonetheless part of the language, then we can deal experimentally with many linguistically interesting questions. We have already seen that we can, in the case of speech, learn about the nature and function of the code; we can investigate the explicit and essentially physiological form of the grammar; and we can find clues to the workings of the decoder. Let me add some further remarks about the decoder because the relevant experiments are closer than much of the rest of speech research to the classical tradition of experimental psychology. The thousands of experiments on the acoustic cues, which were squarely in the tradition, showed us what it is that the decoder must do. Now I would offer a few examples of further questions about grammatical decoding that can, in the case of speech, be asked experimentally.

One fairly general question is put this way: If grammar is processed by a special decoder, then there ought to be, in consequence, a special mode of perception, a grammatical mode. At the higher levels it is hard to study that question. But experiments on speech yield relevant data.

First, let me wring yet another observation out of our tired examples [di] and [du] (Fig. 4). Look again at the very different second-formant transitions that are, as we saw, cues for [d]. When we take these transitions out of the speech pattern and present them alone, we hear them in psychophysically sensible ways. They sound very different, which is reasonable given the large acoustic difference between them. Depending on just how the formants are synthesized, the transitions sound either like the chirps of a bird or like glides in pitch. When heard as chirps, the one in front of [i] is high in pitch, the one with [u] is low. When the listener hears the pitch glide, he knows that the high pitch moves upward and the low pitch moves downward. All these observations make sense in terms of auditory psychophysics.

But if we now embed the same formant transitions in a speech context, then we cannot hear anything like pitches or pitch glides. What we hear is speech, not sound. Perception is at several removes from the acoustic stimulus. It does not directly mirror the external physical events that cause it but rather yields a unique linguistic event, called "d," that cannot be described in auditory terms. In that exact sense, perception in the speech mode is abstract.

I will remark the obvious and say that, in a much vaguer way, syntax, phonology, and, indeed, semantics are abstract in the same sense that speech is. Language holds the world at arm's length. When we communicate linguistically we don't portray the dimensions of physical reality; we only talk about them.

There is further evidence for the existence of a speech mode, and for a characteristic—categorical perception—very closely related to abstractness. The basic experimental finding is that, other things equal, our ability to discriminate variations in the cue is much better at the boundary between phone classes than within the phone class (Liberman, Harris, Hoffman, & Griffith, 1957; Liberman, Harris, Kinney, & Lane, 1961; Stevens, Liberman, Ohman, & Studdert-Kennedy, 1969). As a result, we hear the encoded stops, for example, only as categories, [ba], [da], [ga], which is to say that we readily identify the three phones but cannot discriminate physically different tokens of the same phonetic type. Such discontinuities are not normally found in the perception of continuously varied nonspeech sounds (Eimas, 1963).

I hardly need point out that categoricalness is a design feature of language generally, though, as in the case of abstractness, it is harder to study at the higher levels. There are active and passive voices, never something in between. Nouns are singular or plural. Changing one phonetic segment in a word does not produce a similar word, but a completely different one, or none at all.

So much, then, for the question: Is there a speech mode? I would turn now to another, but very closely related, question: Is linguistic decoding done by apparatus that is borrowed for the purpose from other perceptual and cognitive machinery, or does such processing form its own system? At the level of speech perception, the question takes a more specific form:

Is the speech decoder an auditory device or a linguistic one? One relevant answer comes from the dramatic differences between speech and nonspeech that are found in experiments in which competing signals are presented to the two ears. Using this technique, several investigators have found that words presented to the right ear are better heard than those presented to the left, but that the ear advantage is reversed in the case of music and various other nonspeech sounds (Broadbent & Gregory, 1963; Bryden, 1963; Chaney & Webster, 1965; Kimura, 1961, 1964, 1967). Because it is reasonable to suppose that the representation of the ears in the brain is stronger contralaterally than ipsilaterally, these results have been interpreted to mean that the words want to be processed primarily in the left hemisphere and the nonspeech sounds in the right.

Though the early experiments with dichotic stimulation used meaningful words, we know that the right-ear advantage does not depend on meaning because the effect is obtained with dichotically presented stop-vowel nonsense syllables that differ only in the initial stop (Halwes, to be published; Kimura, 1967; Kirstein & Shankweiler, 1969; Shankweiler & Studdert-Kennedy, 1967; Studdert-Kennedy & Shankweiler, to be published.) The results of this bloodless physiology indicate that the encoded sounds of speech are perceived in one part of the brain, and the unencoded sounds of nonspeech in another. We are therefore encouraged to believe that bare phonetic perception is, in some important sense, not entirely auditory.

Of course, we have long known that language functions generally tend to be on the left side of the brain. So the finding that phonetic perception is there, too, is one more piece of evidence that speech in our narrow sense is an integral part of language. But in the speech case we can, by using the dichotic technique, carry out experiments that answer many interesting questions. We can, for example, inquire more specifically into the nature of the stimulus that yields the right-ear (left hemisphere) effect. Several experiments have shown a substantial right-ear effect in the perception of stop consonants, which were cued primarily by formant transitions but a much smaller effect for steady-state vowels (Kirstein & Shankweiler, 1969; Shankweiler & Studdert-Kennedy, 1967; Studdert-Kennedy & Shankweiler, to be published). This has been carried a step forward by Darwin (1969), who recently found a right-ear effect with fricative-vowel syllables when formant transitions were included as cues but a smaller effect when the cues were only the band-limited, steady-state noises. The question is now being pressed still further in an experiment by Shankweiler, Syrdal, Halwes, and Liberman (personal communication, 1969). They are asking whether the special processing on the left side is for all auditory signals having the properties of the second-formant transitions, or whether it is, rather, for those transitions only when they are in a speech context. To that end, they are measuring the laterality of stops (in stop consonant-vowel syllables) cued only by second-formant transitions, and also the laterality of the second-formant transitions when they are presented alone and sound like the chirps of a bird. If it should be found that the second-formant transitions are more strongly lateralized when perceived as cues for speech than when heard in isolation as chirps, then we shall have further evidence that the speech processor is not so much auditory as linguistic.

I certainly do not mean to imply that the auditory system does nothing, or that it merely transmits the sound. The formants must be extracted from the acoustic signal before they can be decoded, and that is not a trivial task, since the formants often constitute a relatively small part of the total acoustic energy. We have much to learn about how these formant transitions are perceived as auditory events; the recent research on nonspeech transitions by Nabelek and Hirsh (1969) and by Pollack (1968), for example, are very relevant to what we need to know. I only want to emphasize that after the auditory system has heard these formants, some device has got to decode them; that is, some device has got to extract the linguistic features from a parametric description in which those features are still quite thoroughly intermixed and merged.

I would remind you, moreover, that however much the speech processor may stand on its own linquistic feet, it is nonetheless firmly attached to the ear. Consider that, in 20 years of trying, no one has really learned to read speech spectrograms (Fant, 1962, p. 4). This is not because the spectrogram fails to extract the right parameters of the signal. The spectrographic transform is quite appropriate for showing the formants, where almost all of the linguistic information is to be found. Spectrograms cannot be read because the eye cannot cope with the codelike complexities of the signal. The fact that we cannot learn to read spectrograms means that we cannot, by training alone, develop an appropriate decoder and make it work in cooperation with the eye. When we read print, we are dealing with a linguistic signal that has already been decoded to the phonetic level; in print, the optical medium *is* the phonetic message.

It is also of interest to ask about grammatical decoding whether all aspects of language stand in equal need of it. At the level of speech we can see very clearly the difference in this regard between consonants and vowels. As I pointed out earlier when I described the patterns [di] and [du] (Fig. 4), the acoustic cues for the vowels bear a simple alphabetic relation to the phonetic message, but in the case of the stops, on the other hand, the relation is a complexly encoded one. It is as if the stops were more deeply linguistic, more different from nonspeech, than the vowels.

There is experimental evidence to indicate that this is a reasonable way to look at the difference. Several experiments have shown that the tendency to categorical perception—a characteristic of the linguistic mode—is significantly greater in the encoded stops than in the unencoded vowels (Fry, Abramson, Eimas, & Liberman, 1962; Stevens, Liberman, Ohman, & Studdert-Kennedy, 1969). Further evidence about this difference comes from "dichotic" experiments, of the kind referred to earlier, which have shown that the stops are more strongly lateralized than vowels (Kirstein & Shankweiler, 1969; Shankweiler & Studdert-Kennedy, 1967; Studdert-Kennedy & Shankweiler, to be published). Still more evidence has come from a recent experiment in which it was found that binaural time differences affect the perception of stops and vowels in opposite directions (Porter, Shankweiler, & Liberman, 1969). All of these findings may be assumed to reflect differences in the processing required to perceive more and less highly encoded segments at the level of speech.

It is tempting to think that at the higher levels, too, some aspects of language are more deeply linguistic—that is, more highly encoded—than others. We have already seen in our example of syntactic encoding how two deep-structure sentences were, like the initial and final consonants of [bæg], encoded into a third sentence, which served as a relatively unencoded, vowel-like nucleus. At the level of speech we can, as we have seen, experiment with differences between the more and less deeply encoded aspects of the system, and we can hope to understand the significance of these differences for the operation of the grammatical decoder. That understanding may, in turn, give us insight into analogous differences at the higher levels.

Before concluding I would say that for behavior as profoundly biological as speech and language, one wants, of course, to see the always interesting interplay of experience and endowment, and to look at the processes from a comparative point of view. Research at the level of speech shows promise of enabling us to do both very well. Cross-language studies of the production and perception of the speech code have already isolated some biologically interesting universals, while also permitting us to see quite clearly the equally interesting effects of experience (Abramson & Lisker, 1970; Lisker & Abramson, 1964; Stevens, Liberman, Ohman, & Studdert-Kennedy, 1969).

Comparative data about speech are available now only for production; we know that primates other than man do not produce strings of phonetic segments and their encoded acoustic correlates (Lieberman, 1968; Lieberman, Klatt, & Wilson, 1969). Unfortunately, nothing is known about the way nonhuman animals perceive speech. If what I have said in this talk is true. 'owever, we should suppose that, lacking the speech-sound decoder, animals would not perceive speech as we do, even at the phonetic level. It is, of course, possible to find out by experimental means how animals do perceive speech, and the appropriate experiments will one day be done. The point I would make here is that questions about the uniqueness of language are more likely to be answered satisfactorily at the level of speech than elsewhere. In terms of purely behavioral manifestations, language is plainly unique to man; other creatures do not use syntax, phonology, or speech when they communicate. But in studying language at the level of speech, we can often go beneath the behavior, down to the mechanisms that underlie it. Then we can hope to see whatever biological continuity there is. We can reasonably expect to discover whether, in developing linguistic behavior, Nature has invented new physiological devices, or simply turned old ones to new ends. We may well find that what is unique about man is not that he alone commands the physiological principles of grammar but that only he is able to use them in the vocal communication called language.

SUMMARY

Experiments in the psychological tradition have shown that the sounds of speech are related to the phonetic message by a complex and efficient code, bearing formal resemblances to the grammatical codes we know as syntax and phonology. Each of these codes speeds communication by delivering the information in parallel. But the gain in speed is achieved at the cost of a considerable complication, since it is in the nature of the codes that they restructure the information; as a result, the levels they link do not correspond in the number or shape of their segments. That complication presents no problem to us human beings because we all have ready access to grammars that rationalize the codes.

Experiments have also uncovered several special characteristics of perception in the speech mode. These characteristics are the more interesting because they are found at the higher levels, too, though it is usually more difficult to investigate them there.As a source of answers to linguistically interesting questions, speech has the advantage over the other grammars that it is more accessible to experiment. It ought, therefore, to be an attractive area of study for experimental psychologists who are interested in language.

REFERENCES

Abramson, A., & Lisker, L. Discrimination along the voicing continuum: cross-language tests. *Proceedings of the 6th International Congress of Phonetic Sciences, Prague 1967*, 569–573 (1970).

Broadbent, D. E., & Gregory, M. Accuracy of recognition for speech presented to right and left ears. *Quarterly Journal of Experimental Psychology*, 1963, 65, 103–105.

BRYDEN, M. P. Ear preference in auditory perception. *Journal of Experimental Psychology*, 1963, 65, 103–105.

CHANEY, R. B., & WEBSTER, J. C. Information in certain multidimensional acoustic signals. Report No. 1339, 1965, United States Navy Electronics Laboratory Reports, San Diego, California.

CHOMSKY, N. *Syntactic structures.* The Hague: Mouton, 1957.

CHOMSKY, N. *Aspects of the theory of syntax.* Cambridge, Mass.: M.I.T. Press, 1965.

CHOMSKY, N., & MILLER, G. A. Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, and E. Galanter, (Eds.), *Handbook of mathematical psychology.* New York: Wiley, 1963. Pp. 269–321.

COOPER, F. S. Describing the speech process in motor command terms. *Journal of the Acoustical Society of America*, 1966, 39, 1221 (Abstract) (Status Report of Speech Research, Haskins Laboratories, SR-5/6, 1966, 2.1–2.27—text).

DARWIN, C. J. Auditory perception and cerebral dominance. Unpublished doctoral dissertation, Cambridge University, 1969.

EIMAS, P. D. The relation between identification and discrimination along speech and non-speech continua. *Language and Speech*, 1963, 6, 206–217.

FANT, C. G. M. Descriptive analysis of the acoustic aspects of speech. *Logos*, 1962, 5, 3–17.

FROMKIN, V. A. Neuro-muscular specification of linguistic units. *Language and Speech*, 1966, 15, 219–242.

FRY, D. B., ABRAMSON, A. S., EIMAS, P. D., & LIBERMAN, A. M. The identification and discrinination of synthetic vowels. *Language and Speech*, 1962, 5, 171–189.

HALWES, T. Effects of dichotic fusion on the perception of speech. Unpublished doctoral dissertation, University of,Minnesota, 1969. (Also reproduced as Supplement to Status Report of Speech Research, Haskins Laboratories; to be published).

HARRIS, K. S., LYSAUGHT, G., & SCHVEY, M. M. Some aspects of the production of oral and nasal labial stops. *Language and Speech*, 1965, 8, 135–147.

JAKOBSON, R., FANT, G., & HALLE, M. Preliminaries to speech analysis. The distinctive features and their correlates. Technical Report No. 13, 1952, Acoustics Laboratory, M.I.T. (Republished, Cambridge, Massachusetts: M.I.T. Press, 1963).

KIMURA, D. Cerebral dominance and perception of verbal stimuli. *Canadian Journal of Psychology*, 1961, 15, 166–174.

KIMURA, D. Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology*, 1964, 16, 335–358.

KIMURA, D. Functional asymmetry of the brain in dichotic listening. *Cortex*, 1967, 3, 163-178.

KIRSTEIN, E., & SHANKWEILER, D. Selective listening for dichotically presented consonants and vowels. Paper read before the 40th Annual Meeting of the Eastern Psychological Association, Philadelphia, 1969.

KLATT, D. Structure of confusions in short-term memory between English consonants. *Journal of the Acoustical Society of America*, 1968, 44, 401–407.

KOZHEVNIKOV, V., & CHISTOVICH, L. Rech' Artikuliasia i vosprilatie. Moscow-Leningrad, 1965 (Trans. in Speech: Articulation and Perception. Washington: Joint Publications Research Service, 1966, 30, 543).

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, 74, 431–461.

LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H., & GRIFFITH, B. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, 54, 358–368.

LIBERMAN, A. M., HARRIS, K. S., KINNEY, J., & LANE, H. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 1961, 61, 379–388.

LIEBERMAN, P. Primate vocalizations and human linguistic ability. *Journal of the Acoustical Society of America*, 1968, 44, 1574–1584.

LIEBERMAN, P., KLATT, D., & WILSON, W. A. Vocal tract limitations on the vowel repertoires of rhesus monkeys and other nonhuman primates. *Science*, 1969, 164, 1185–1187.

LISKER, L., & ABRAMSON, A. A cross-language study of voicing in initial stops: acoustical measurements. *Word*, 1964, 20, 384–422.

MATTINGLY, I. G., & LIBERMAN, A. M. The speech code and the physiology of language. In K. N. Leibovic, (Ed.), *Information processing in the nervous system*. New York: Springer Verlag, 1969.

MILLER, G. A., & NICELY, P. E. Analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 1955, 27, 338–353.

NABELEK, I., & HIRSH, I. J. On the discrimination of frequency transitions. *Journal of the Acoustical Society of America*, 1969, 45, 1510–1519.

POLLACK, I. Detection of rate of change of auditory frequency. *Journal of Experimental Psychology*, 1968, 77, 535–541.

PORTER, R., SHANKWEILER, D. P., & LIBERMAN, A. M. Differential effects of binaural time differences in perception of stop consonants and vowels. Paper presented at annual meeting of the American Psychological Association, Washington, D. C., September 2, 1969.

SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Identification of consonants and vowels presented to left and right ears. *Quarterly Journal of Experimental Psychology*, 1967, 19, 59–63.

STEVENS, K. N. & HOUSE, A. S. Speech perception. In J. Tobias (Ed.), *Foundations of modern auditory theory*, New York: Academic Press, in press.

STEVENS, K. N., LIBERMAN, A. M., OHMAN, S. E. G., & STUDDERT-KENNEDY, M. Crosslanguage study of vowel perception. *Language and Speech*, 1969, 12, 1–23.

STUDDERT-KENNEDY, M., & SHANKWEILER, D. P. Hemispheric specialization for speech perception. Status Report of Speech Research, Haskins Laboratories; to be published.

WICKELGREN, W. A. Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America*, 1966, 39, 388–398.

(Accepted June 1, 1970)