Sonderdruck aus

Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung

Band 21, Heft 1/2, 1968

AKADEMIE-VERLAG · BERLIN

A. M. Liberman*, F. S. Cooper, M. Studdert-Kennedy**, K. S. Harris, and D. P. Shankweiler, Haskins Laboratories New York

On the Efficiency of Speech Sounds

Summary

The sounds of speech are uniquely efficient vehicles for the transmission of phonemic information. Serious attempts to communicate language by sounds other than speech have repeatedly failed; no acoustic cipher has yet come within an order of magnitude of matching the performance of speech sounds. When we compare the requirements of speech communication to the known properties of the ear, we see, indeed, that we should not expect to transmit a phonemic system by a sound alphabet or cipher – by any sounds, that is, that stand in a one-to-one relation to the phonemes they represent. We are not surprised, then, to discover that speech sounds are not a cipher on the phonemic structure of the language, but a special and complex code: the primary acoustic cues for a given phoneme may differ widely from one context to another, and it is, in general, impossible to assign the observable acoustic boundaries to phoneme segments. The purpose of this paper is to examine some of the characteristics of that code. We shall pay particular attention to those properties that enable the sounds of speech, alone among acoustic signals and in spite of the limitations of the ear, to work so well.

Our concern is with the question: how is it that on hearing the acoustic stream of speech a listener perceives phonemes? The question is reasonable only if we assume that phonemes *are* perceived. Our purpose here is not to justify that assumption[1], but to accept it and then go on to ask how such perception might occur. There are two compelling reasons for an interest in the question. First, there are several indications that phonemes could not be communicated by a simple cipher or sound alphabet. Second, there is good evidence that phonemes are not, in fact, communicated in this simple way, but are, rather, recovered from a special code.

The ear is not well suited to the reception of phoneme segments represented by an acoustic alphabet. Speech can be perceived at rates that require the listener to take in as many as 30 such segments per second. But the temporal resolving power of the ear is limited, and 30 acoustic segments per second would merge,

* Also, the University of Connecticut. ** Also, the University of Pennsylvania.

[1] There is a considerable weight of evidence for the psychological reality of the phoneme; the ingenious experiments described by Koževnikov and Čistovič [Koževnikov, V. A., *Reč. artikuljacija i vosprijatie* (Moskva-Leningrad, 1965)] provide another and recent example of such evidence. We are not assuming that the phoneme is always perceived, or that speech perception is always phonemic, only that the phoneme can be perceived and often must be perceived. The term "phoneme" is used here in a linguistic sense and to denote the perceptual unit that is the nearest counterpart of the linguistic entity.

perceptually, into an unanalyzable buzz[2]. Furthermore, an acoustic alphabet would call for a large number of identifiable acoustic shapes: there are approximately 40 phonemes in English, more in some other languages. The literature on auditory perception does not encourage the belief that it would be possible to find 40 or more highly identifiable acoustic signals of short duration that could be made to stand for those phonemes in a simple sound alphabet[3].

Direct evidence that a cipher or sound alphabet on the phonemes is not likely to work comes from experience with non-speech ciphers on the language. These include not only the familiar case of International Morse, but also the several ciphers devised over more than fifty years of research and development in an attempt to build reading machines for the blind[4]. The difficulty has not been to transform print into sound, but to find a set of nonspeech sounds that can be identified rapidly and accurately. Many people have practiced for years, even decades, with a variety of sounds, yet top speeds are scarcely a tenth of the rate we can attain with speech.

We should not be surprised, then, to discover that the sounds of speech are not an alphabet or cipher, but a special code. Speech sounds are a code in the sense that they represent a restructuring of the phonemic message: several units of the original are replaced by a single unit of the encoded message, that is, several successive phonemes by a single syllable[5]. The evidence from experiments

[2] R. H. Stetson, *Motor Phonetics*. 2nd Edition. North Holland Publishing Co. (Amsterdam, The Netherlands, 1951); M. Studdert-Kennedy and A. M. Liberman. Psychological considerations in the design of auditory displays for reading machines. *Proc. of the International Congress on Technology and Blindness, I*, 289–304 American Foundation for the Blind (New York, 1963).

[3] I. Pollack, The information of elementary auditory displays. *J. Acoust. Soc. Amer.* **24**, 745–749 (1952); I. Pollack and L. Ficks, *J. Acoust. Soc. Amer.* **26**, 155–158 (1954); G. A. Miller, The magical number seven, plus-or-minus two, or, some limits on our capacity for processing information. *Psychol. Rev.* **63**, 81–96 (1956); P. W. Nye, Aural recognition time for multidimensional signals. *Nature*, **196**, 1282–1283 (London, 1962).

[4] F. S. Cooper, Research on reading machines for the blind. In *Blindness*, P. A. Zahl, Ed. (Princeton, 1950); J. Freiberger and E. F. Murphy. Reading machines for the blind. *IRE Professional Group on Human Factors in Electronics* (March, 1961); J. L. Coffey, The development and evaluation of the Battelle Aural Reading Device. *Proceedings of the International Congress on Technology and Blindness, I*, 343–360. American Foundation for the Blind (New York, 1963); P. W. Nye, Reading aids for blind people – a survey of progress with the technological and human problems. *Med. Electron. Biol. Engng.*, **2**, 247–264 (1964); P. W. Nye, An investigation of audio outputs for a reading machine. National Physical Laboratory (Autonomics Division), Teddington, England (February, 1965).

[5] The encoded nature of speech may not be limited to the segmental phonemes. There is evidence that stress and intonation may present similarly complex relations between acoustic cue and perceived language. See: B. Lindblom, Spectrographic study of vowel reduction. *J. Acoust. Soc. Amer.*, **35**, 1773–1781 (1963); K. Hadding-Koch and M. Studdert-Kennedy, An experimental study of some intonation contours. *Phonetica*, **11**, 175–185 (1964); and P. Lieberman, Intonation and syntactic processing of speech. *Proc. AFCRL Symposium on Models for the Perception of Speech and Visual Form*. AFCRL (In press). The abstracts of Lieberman

on the acoustic basis of speech perception points to at least two general conclusions indicating that speech is a code. First, there is, in general, no way to cut the acoustic signal along the time dimension so as to recover segments that will be perceived as separate phonemes[6]. This is because the acoustic cues for successive phonemes are overlapped to form units of approximately syllabic size. The relevant facts have emerged over the past fifteen years in many publications. Among the more interesting treatments is that contained in the recent monograph by Koževnikov and Čistovič[7].

The other general and related conclusion is that for many phonemes, including in particular those consonants that seem to carry the heaviest information load, there is no way to define the acoustic cues so as to have an invariant relation with the phoneme or with phoneme perception. The acoustic cue for the same phoneme (as perceived) is often vastly different in different contexts[8]. Moreover, the invariance problem may be further complicated by variations in the rate of articulation[9].

We should emphasize that not all phonemes are so thoroughly restructured. In slow articulation the acoustic cues for the fricatives and the vowels seem to be largely independent of context, and the cues can be reasonably well isolated as phonemic segments. In rapid articulation, however, some restructuring tends to occur, even with these phonemes.

The intermixing and overlapping of the acoustic representations of the phonemes to form syllables is the essence of the code. This reduces by a factor of three or four the number of discrete acoustic segments that must be perceived

(On the Structure of Prosody) and Öhman and Lindquist (Analysis-by-Synthesis of Prosodic Pitch Contours) suggest that some aspects of this are to be presented at this symposium.

[6] C. M. Harris, A study of the building blocks of speech. *J. Acoust. Soc. Amer.*, **25**, 962–969 (1953); G. Peterson, W. Wang and E. Sivertsen. Segmentation techniques in speech synthesis. *J. Acoust. Soc. Amer.* **30**, 739–742 (1958); and A. M. Liberman, F. Ingemann, L. Lisker, P. Delattre and F. S. Cooper. Minimal rules for synthesizing speech. *J. Acoust. Soc. Amer.*, **31**, 1490–1499 (1959).

[7] V. A. Koževnikov and L. A. Čistovič, *Op. cit.*

[8] A. M. Liberman, P. Delattre and F. S. Cooper, The role of selected stimulus variables in the perception of the unvoiced stop consonants. *Am. J. Psychol.*, **65**, 497–516 (1952); A. M. Liberman, P. Delattre, F. S. Cooper and L. Gerstman, The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, **68**, No. 8, 1–13 (1955); A. M. Liberman, Some results of research on speech perception. *J. Acoust. Soc. Amer.* **29**, 117–123 (1957); F. S. Cooper, A. M. Liberman, K. S. Harris and P. M. Grubb, Some input-output relations observed in experiments on the perception of speech. *Proc. of 2nd Intl. Cong. of Cybernetics*, 930–941, Namur, Belgium (1958); K. N. Stevens, Toward a model for speech recognition. *J. Acoust. Soc. Amer.*, **32**, 47–55 (1960); L. Lisker, F. S. Cooper and A. M. Liberman, The uses of experiment in language description. *Word*, **18**, 82–106 (1962); A. M. Liberman, F. S. Cooper, K. S. Harris, P. F. MacNeilage and M. Studdert-Kennedy, Some observations on a model for speech perception. *Proc. AFCRL Symposium on Models for the Perception of Speech and Visual Form.* AFCRL (In press).

[9] B. Lindblom, Spectrographic study of vowel reduction. *J. Acoust. Soc. Amer.*, **35**, 1773–1781 (1963).

per unit time, and thus enables the listener to evade the limitations on rate of segment perception set by the temporal resolving power of the ear. In effect, the code transmits the phonemes in parallel in the sense that, at every instant, the acoustic signal provides information about more than one phoneme. We begin, then, to understand why the code is efficient. But the fact of the code poses an important question: by what mechanism does the listener decode the signal and recover the phoneme segments?

We will return to this question later. To appreciate that the speech code is special in some interesting sense, we must first consider the evidence for the existence of a distinctive mode of perception for speech. In several different kinds of experiments it has been found that the same or similar acoustic stimuli are perceived differently when, in the one case, they are important parts of speech signals and when, in some non-speech context, they are not [10]. Related to this is the evidence from studies of consonant perception that continuous variations in the acoustic cue are perceived discontinuously – that in these cases the phonemes are categorical, not only in the abstract linguistic sense, but as immediately given in perception [11]. This, we should think, is the speech mode. Continuous variations in non-speech sounds and, indeed, in isolated steady-state vowels, are perceived continously, which is, presumably, the normal non-speech mode [12]. The abstract of K. N. Stevens indicates that vowels in proper dynamic context are perceived more nearly in the fashion of the stop consonants [13]. These, then, would appear to be in the speech mode, and, according to the Stevens abstract, for theoretically interesting reasons.

[10] A. Liberman, K. S. Harris, J. Kinney and H. Lane, The discrimination of relative onset-time of the components of certain speech and non-speech patterns. *J. Exptl. Psych.*, **61**, 379–388 (1961); J. Bastian, P. Eimas and A. M. Liberman, Identification and discrimination of a phonemic contrast induced by silent interval. *J. Acoust. Soc. Amer.*, **33**, 842 (1957) (A). A. M. Liberman, K. S. Harris, P. Eimas, L. Lisker and J. Bastian, An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance. *Language and Speech*, **4**, 175–195 (1961); A. House, K. N. Stevens, T. Sandel and J. Arnold, On the learning of speech-like vocabularies. *J. Verbal Learn. and Verbal. Behav.*, *I*, 133–143 (1962).

[11] See all but the last of the papers referenced in Fn. 10 and also: A. M. Liberman, K. S. Harris, H. Hoffman and B. Griffith, The discrimination of speech sounds within and across phoneme boundaries. *J. Exptl. Psych.*, **54**, 358–368 (1957); B. Griffith, A study of the relation between phoneme labeling and discriminability in the perception of synthetic stop consonants. Unpubl. Ph. D. dissertation, Univ. of Conn., 1958.

[12] D. Fry, A. S. Abramson, P. Eimas and A. M. Liberman, The identification and discrimination of synthetic vowels. *Language and Speech*, **5**, 171–189 (1962); P. Eimas, The relation between identification and discrimination along speech and non-speech continua. *Language and Speech*, **6**, 206–217 (1963); K. N. Stevens, S. E. G. Öhman and A. M. Liberman, Identification and discrimination of rounded and unrounded vowels. *J. Acoust. Soc. Amer.*, **35**, 1900 (1963) (A).

[13] K. N. Stevens, On the relations between speech movements and speech perception. Abstract of paper to be presented at the Symposium on the Perception of Speech and Speech Mechanisms, XVIIIth Intl. Cong. of Psychol., Moscow, August, 1966.

Also relevant to the existence of these two modes are the now-emerging facts on lateral differences in auditory perception. Several investigators have in recent years found small but reliable differences in the response to various acoustic stimuli depending on the ear to which the stimuli are presented[14]. (Due to the greater efficacy of the crossed neural pathways, inferences concerning the relative contribution of the two temporal lobes in processing different kinds of auditory input can be drawn from such experiments.) When different speech materials are presented to the two ears simultaneously, i. e., dichotically, thus creating binaural rivalry, the input to the right ear is more accurately perceived than that to the left ear. With brief melodies and sonar signals, the reverse is found. To discover, thus, that speech and non-speech sounds may be more effectively processed in different parts of the brain accords well with the hypothesis that such sounds are somehow processed differently. It also fits well with clinical evidence regarding the very different consequences of damage to the left and right cerebral hemispheres.

In our own laboratory we have recently found, in experiments with dichotically presented materials, that perception of stop consonants is more strongly lateralized (in the left hemisphere) than is the perception of vowels[15]. We interpret this to mean that stops are more readily processed by the speech decoder, which is in the left hemisphere, while the vowels can be, and presumably sometimes are, processed as if they were nonspeech. The different tendencies to lateralization of stops and vowels fit well with the evidence referred to earlier concerning their differing tendencies to categorical and continuous perception.

We have so far been concerned to establish that speech is some kind of code on the phonemic structure of the language. We should like to turn our attention now to the perceptual process by which the phonemes might be recovered from the coded acoustic signal. For that purpose it will be useful to consider, first, where and how the encoding might have occurred.

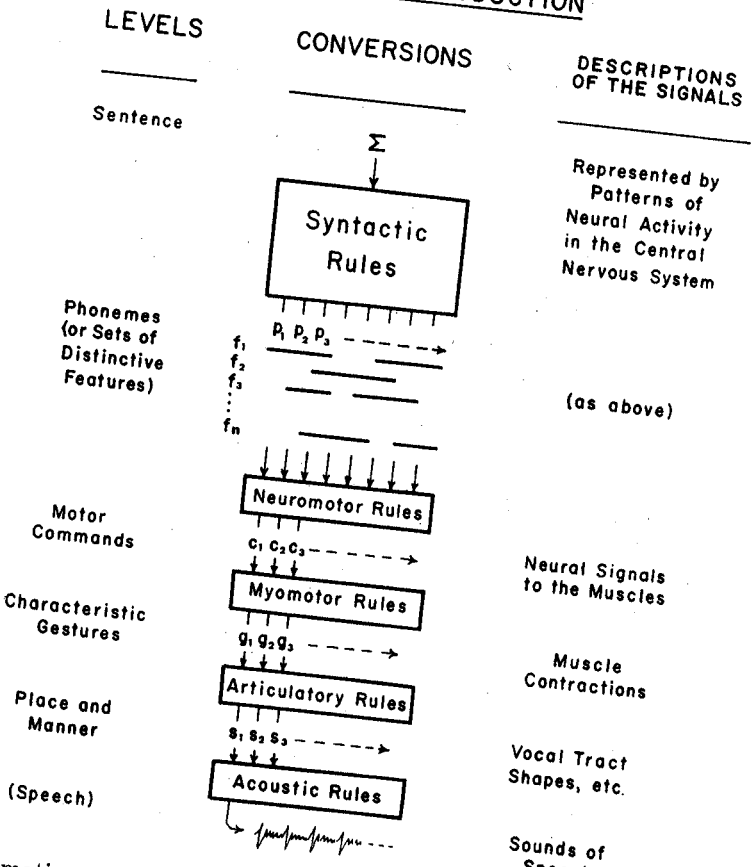
In thinking about the successive steps in the process of producing speech, we can proceed – as speech itself presumably does – from meaningful units of

[14] D. KIMURA, Cerebral dominance and the perception of verbal stimuli. *Canad. J. Psychol.*, **15**, 166–171 (1961); *Idem.* Some effects of temporal-lobe damage on auditory perception. *Canad. J. Psychol.* **15**, 156–165 (1961); B. MILNER, Laterality effects in audition. In V. B. MOUNTCASTLE (Ed.). *Interhemispheric Relations and Cerebral Dominance.* Johns Hopkins Press (Baltimore, 1962); M. P. BRYDEN, Ear preference in auditory perception. *J. Exptl. Psychol.*, **65**, 103–105 (1963); D. KIMURA, Left-right differences in the perception of melodies. *Quart. J. Exptl. Psychol.* **16**, 355–358 (1964); D. E. BROADBENT and M. GREGORY, Accuracy of recognition for speech presented to the right and left ears. *Quart. J. Exptl. Psychol.* **16**, 359–360 (1964); J. C. WEBSTER and R. B. CHANEY, Jr., Information and complex signal perception. *Proc. AFCRL Symposium on Models for the Perception of Speech and Visual Form*, AFCRL. (In press); D. SHANKWEILER. Effects of temporal-lobe damage on perception of dichotically presented melodies. *J. Comp. Physiol. Psychol.* **62**, 115–119 (1966).

[15] D. SHANKWEILER and M. STUDDERT-KENNEDY, Identification of consonants and vowels presented to left and right ears. *Quart. J. Exptl. Psychol.* **19**, 59–63 (1967). *Idem.* An analysis of perceptual confusions in identification of dichotically presented CVC syllables. *J. Acoust. Soc. Amer.* (1967), in press. (Abstract).

sentence size through smaller, sementically empty units at the phonological level to the final acoustic waveform. This is the sequence shown in Figure 1. Here the initial operations at the syntactic level, currently the subject of widespread study and discussion, have been grouped under a single heading[16]. A more detailed diagram would distinguish such components as phrase structure rules, trans-

SCHEMA FOR PRODUCTION



formational rules, and morphophonemic rules. Since our concern is primarily with the phonological phase, however, we shall skip the syntactic operations and start with the message in the form of a phoneme sequence, taking this sequence as the input to successive converters that operate by neuromotor rules, myomotor rules, articulatory rules, and acoustic rules to yield, eventually, an acoustic stream.

In the first of these operations, the neuromotor rules serve to convert the ordered string of phonemes into a temporal sequence of neural signals to the

[16] For a recent treatment and reference, see: N. CHOMSKY, *Aspects of the Theory of Syntax*. M. I. T. Press (Cambridge, Mass., 1965).

muscles of articulation. The extent to which the subphonemic features, which collectively comprise the phoneme at each instant, are uniquely represented by these neural signals is a point to which we shall return, for it is central to a working hypothesis that is guiding a major part of our own experimental work.

The relationships seem simpler and clearer at the second stage in the production process, namely, the conversion of neural signals into muscle contractions on the basis of myomotor rules. The signals map directly onto the muscles and control their contractions, whether via the intrafusal system or directly; moreover, the muscular events are observable by electromyographic techniques.

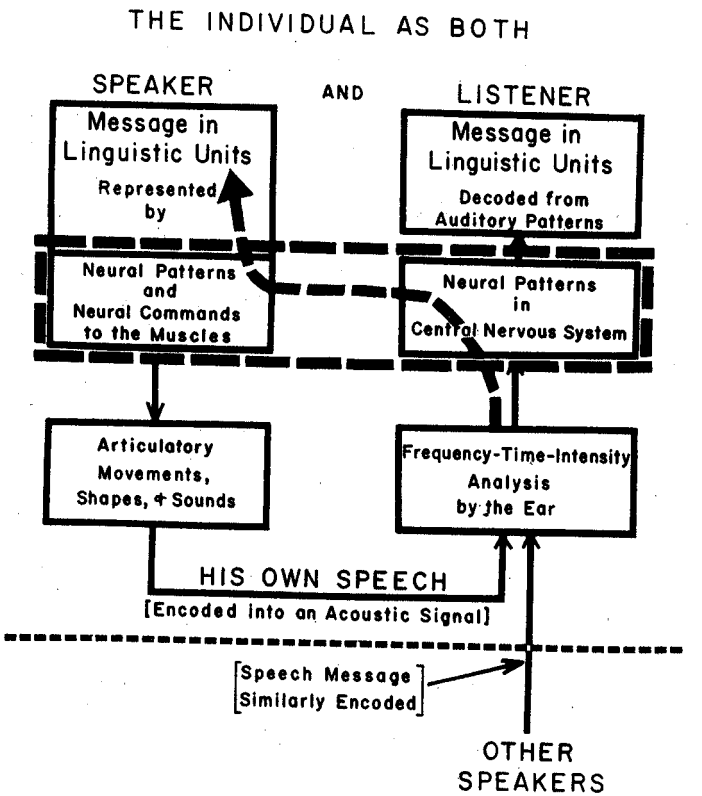
But simplicity has disappeared again at the next stage: the conversion from muscle contractions to vocal tract shapes (and related expiratory movements) by way of articulatory rules. The complexity introduced by this conversion is of two kinds. One follows from the bone and muscle structure of the articulatory system, with its intricate mechanical linkages and the spatial overlap of its muscular actuators. The shape that the tract will take in response to a particular set of muscle contractions is, one supposes, fully predictable, though extraordinarily difficult to compute. But what will happen if a new set of contractions (for the next phoneme) begins before the last set has had its full effect? Clearly, this introduces another kind of complexity, one that arises from temporal overlap of the incoming instructions. The shapes that result will no longer stand in one-to-one correspondence with the phonemes, but will reflect at each instant the interacting influences of several phonemes. This merging of effects constitutes a true encoding (or restructuring) of the message: this restructuring will be especially complex when temporal and spatial overlap occur together. Since overlap is the rule in speech, we can summarize the effects of the conversion from contraction to shape by saying that it is complex at best and almost always introduces an encoding of the sequential input units into output units of about syllabic size.

The final conversion, from continuously changing shape to a modulated acoustic stream, is by now rather well understood, thanks in considerable part to the able efforts of one of the organizers of this conference, Dr. GUNNAR FANT[17]. The application of the acoustic rules is complex in a computational sense and may give unequal acoustic prominence to various aspects of the changes in shape; nevertheless, the rules operate on an instant-by-instant basis and yield (for the most part) one-to-one relations between shape and sound, so that it is appropriate to consider this step from shape to sound as an enciphering rather than an encoding operation.

There are, then, between the phonological input and the acoustic output, at least four distinguishable conversions. We can say with assurance that the third step necessarily introduces an encoding of about the kind we observe in the output sound stream. And so we see how and why it is that speech, as it exists out in the air, must be a coded message.

[17] FANT's monograph on this subject offers a comprehensive treatment and references to other related research [C. G. M. FANT, *Acoustic Theory of Speech Production*. Mouton ('s-Gravenhage, 1960)].

But how does the listener decode it? Let us consider the total process of speaking and listening as it is sketched in Figure 2. Please ignore for the moment the dashed lines and consider only the U-shaped sequence. The schematic account it gives of events in reception – the righthand side of the U – is just what one would except if the perception of speech had no special connection with its



production. This is, indeed, the obvious and seemingly simple view of how speech is perceived, namely, that acoustic units of appropriate size are learned as separate neural patterns which are stored in the central nervous system, with as many neural patterns as there are words (or perhaps syllables) in the language. This would solve the invariance problem, albeit by an extravagant use of storage, but it leaves unexplained the listener's known ability to perceive on a phoneme-by-phoneme basis. As an alternative, and to explain how the listener retrieves the phonemes, we might postulate an auditory decoder (i. e., one that operates directly on the acoustic signal) with functions that include, at the very least, an inverse of the encodings imposed by bone and muscle on the acoustic stream.

But can we not account for the perception in some less extravagant way? We think so. We think there is a mode of perception uniquely fitted to speech and responsible for the high efficiency of its signals. In the most general terms, this

mode takes advantage of readily available mechanisms that allow perception to operate by reference to the motivating events of articulation[18].

A possible model[19] finds its basis in overlapping activity of the neural networks that supply control signals to the articulatory apparatus and those that process incoming neural patterns from the ear. We know that temporal overlap of these activities exists as an ever-present consequence of the fact that people listen while they speak. If we assume also (a) that there is functional overlap at the neural level so that *both* motor and sensory networks respond (in ways that are characteristic of the activating event) when *either* is activated, and (b) that information can be passed in either direction through these neural mechanisms, then there exists a path from ear to perceived message that is not dependent on an *auditory* decoder or require a vast store of auditory patterns. It is this pathway, and the assumed areas of overlap in neural structure and function, that are implied by the dashed lines of the figure.

Clearly, it will be important to know the linguistic level at which the message units are recovered, and the model we have described does not speak to this point. Before we turn to one that does, it may be useful to make some further observations about this very general model for speech perception by reference to production: (1) Even so general a model as this permits useful inferences, since it implies recovery of the speaker's own message – or his analysis as a listener of the messages of other speakers – in terms of the same linguistic units that enter into production. Thus we can see, in a general way, how the listener is able to decode complexly encoded messages into their linguistic components without having to make use of a special auditory decoder. (2) The reference to motor activity does not – as in some older motor theories – call for reference to the peripheral muscle activity and its proprioceptive consequences. This kind of operation is not excluded by the model, but is no part of it. (3) The model deals primarily with the nature of the decoding mechanism, not with how it was

[18] Some of the bases for this view and its development into a motor theory of speech perception are to be found in earlier publications from our laboratory. See especially: A. M. LIBERMAN, P. C. DELATTRE and F. S. COOPER, The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Am. J. Psychol.*, **65**, 497–516 (1952); A. M. LIBERMAN, Some results of research in speech perception. *J. Acoust. Soc. Amer.* **29**, 117–123 (1957); F. S. COOPER, A. M. LIBERMAN, K. S. HARRIS and P. M. GRUBB. Some input-output relations observed in experiments in the perception of speech. *Proc. 2nd Intl. Cong. Cybernetics.* 930–941 (Namur, Belgium, 1958); L. LISKER, F. S. COOPER, A. M. LIBERMAN, The uses of experiment in language description. *Word*, **18**, 82–106 (1962); A. M. LIBERMAN, F. S. COOPER, K. S. HARRIS and P. F. MACNEILAGE, A motor theory of speech perception. *Proc. Speech Comm. Seminar*, Royal Institute of Technology (Stockholm, 1963); A. M. LIBERMAN, F. S. COOPER, K. S. HARRIS, P. F. MACNEILAGE and M. STUDDERT-KENNEDY, Some observations on a model for speech perception. *Proc. AFCRL Symposium on Models for the Perception of Speech and Visual Form*, AFCRL (In press).

[19] The model sketched out in this paper is intentionally non-restrictive as to physiological mechanisms. The reader will detect, however, an obligation to HEBB and MILNER that the authors are glad to acknowledge. [D. O. HEBB, *The Organization of Behavior*. Wiley (New York, 1949); P. M. MILNER, The cell assembly: Mark II. *Psychol. Rev.*, **64**, 242–252 (1957).]

acquired by the species or the individual. The relative contributions of intrinsic structure and learning, as well as their interaction, raise interesting but separate questions. (4) Reference to production provides a pathway for perception, but not one that is obligatory – that is, the existence of this pathway does not preclude direct auditory processing of speech patterns by the same means that are used for recognizing animal cries, traffic noises, and the like. The special pathway would be used, we suppose, whenever it facilitates perception, as it would in recovering linguistic units that lack invariant acoustic counterparts, but when it is not needed, it may not be used. (5) Finally, the special processes that permit reference to production are not necessarily restricted to reference at the phonological level or, for that matter, to only one level at a time.

So much for this very general model. It is, as we have noted, noncommittal as to the level at which reference and recovery occur. Since our present concern is with the phoneme and how it is perceived so well and so fast, we shall try to make the model more specific in ways that bear on that question. One possibility is to postulate that the reference occurs at a level in production where the neural patterns represent the individual phonemes (or the sets of subphonemic features) i. e., at the input to the first of the phonological converters of Fig. 1, the one that operates by neuromotor rules. This would "explain" why it is the phonemes that we perceive – but it accounts for very little else.

A more productive assumption is that the neural overlap is at, or just above, the output of this neuromotor converter. Here we would expect to find separate neural patterns that send actuating signals to the separate muscles (or closely related groups of muscles) of the articulatory apparatus. But independent control of the component parts of the mechanism means that the information flowing through it becomes multidimensional in physiological coordinates; that is, the serially ordered phonemes come to be represented by muscle activities that proceed in parallel – as do their component features – and that can persist and overlap as segmental phonemes could not. These muscle activities will leave their traces in the sound stream, though here the original dimensions are no longer represented in independent form. If, now, the neural patterns of reception, which likewise contain the "traces" of the original muscle activities, overlap and link into the motor patterns that actuate the muscle groups, then reception, too, will be multidimensional, and in the original terms.

The original terms, however, were those of multidimensional motor control signals that had somehow come to represent the subphonemic features or, collectively, the phonemes. But let us be more explicit about how this happens. The strongest assumption would be that each individual phoneme of a language has its own characteristic set of neural patterns all the way down to the motor nerves that go to the muscles. This is, in fact, the assumption we are testing in some of our experimental work.

The working model containing this assumption does not imply, however, that the *total* neural and muscular activity will be in one-to-one correspondence with the phoneme, but implicates only one or a few component parts, perhaps even the contraction of a single specific muscle for each of the component features.

The neural signals for this characteristic component of the total activity are what we have referred to as motor commands[20]; the term was chosen to distinguish these characteristic, or invariant, signals from all the other neuromotor signals (needed in well coordinated gestures) that may be present at the same time. The objective of the experiments is, then, to find muscle contractions (from which motor commands can be inferred) that are present whenever a particular phoneme is present in a message and are not present when the phoneme is not[21].

It may turn out that Nature will not endorse so simple a model. She provides many examples of intricate motor coordination in skilled movements, and it may be true of speech that significant reorganization of the neural patterns occurs above the level at which nerve impulses are sent to the muscles. A likely case would be the production of clusters that overlap spatially; these might come to be treated as "ligatures" in motor command terms. The model we have proposed is not too inflexible for such eventualities; it would, though, be less useful in helping us to understand the processes and to predict their consequences if the neural reorganizations were so extensive and so far upstream in the nervous system as to be inaccessible to experiment. But even if this should happen, the more general model for perception by reference to production would remain, with much to recommend it.

We began these observations by asking how speech can be so rapid – how it can evade the rate limitations of the ear – and how segmental phonemes can be recovered as well – perceived segments from the unsegmented and usually encoded acoustic stream. We have found an explanation in terms of a model[22] that refers the acoustic and auditory features of an incoming message back to the multidimensional events of actual or implicit articulation. Simultaneous tracking of the features can then lead to recovery of the equivalent multidimensional descriptions, and so of the phoneme string. Thus, the phonemes are disassembled

[20] Motor commands are essentially the same as the "action patterns" we discussed in an earlier description of this model. (F. S. COOPER, A. M. LIBERMAN, K. S. HARRIS and P. M. GRUBB, *Op. cit.*)

[21] Little can be said here of experimental methods and early results. See, however, K. S. HARRIS, M. M. SCHVEY and G. F. LYSAUGHT, Component gestures in the production of certain final clusters. *J. Acoust. Soc. Amer.* **35**, 461–463 (1963); P. F. MACNEILAGE and G. M. SHOLES, An electromyographic study of the tongue during vowel production. *J. Speech and Hear. Res.* **7**, 209–232 (1964); F. S. COOPER, Research techniques and instrumentation: emg. *Proc. of the Conference on Communicative Problems in Cleft Palate, ASHA Reprints*, No. 1 (April, 1965); K. S. HARRIS, G. F. LYSAUGHT and M. M. SCHVEY, Experimental studies of the production of oral and nasal labial stops. *Language and Speech*. Electromyography is a most useful tool since it permits very direct inferences about the neural commands to the muscles, but X-rays and a variety of physiological measures are valuable also, even though the information they give directly is about the encoded consequences of the neural commands. ÖHMAN has described a model that computes vocal tract shapes resulting from VCV coarticulation. [S. E. G. ÖHMAN, Numerical model for coarticulation, using a computer-simulated vocal tract. *J. Acoust. Soc. Amer.* **36**, 1038 (1964) (A).]

[22] Both similarities and differences may be noted between the model discussed here and STEVEN's model for analysis-by-synthesis. [K. N. STEVENS, Toward a

and reassembled during production and reception, repectively, and are transmitted in a parallel mode as features[23] that refer to neuromotor events.

The transmission is in terms of several slowly changing features, and so falls easily within the speed limitations of the ear-brain system. But how do we recover the high-speed timing and proper sequencing of implicit motor events, as we must since correct timing is as important in production as the correct selection of actuators? Perhaps this is the role of phonological constraints and a justification of the phoneme's existence[24], for if the neuromotor events are allowed to occur only in particular combinations and sequences, then there is a basis for re-synchronizing – and so for imposing phoneme boundaries on – the incoming multidimensional message stream.

In short, we have assumed that the encoding of the phonemes occurs below the level of the neural patterns that send actuating signals to the articulatory muscles. By referring the incoming sound to those patterns, the listener can track the features and recover the invariant relation to the phoneme. One can suppose, moreover, that the many and distinctively different dimensions of the neuromotor events give the listener a basis for identifying the phonemes absolutely, and for doing this far better than he can with an equal number of acoustic signals that are not in the speech mode. Finally, we see that by encoding the message so as to put the phoneme segments through in parallel (as features), we avoid the limitations on rate of discrete segment perception that are set by the temporal resolving power of the ear. In general, then, we find that the sounds of speech are uniquely well perceived because they are a special and especially efficient code, processed by a special and readily available mechanism.

model for speech recognition. *J. Acoust Soc. Amer.* **32**, 47–55 (1960).] Both involve interaction between productive and receptive processes, with a comparison of the two sets of signals and linguistic choices determined by this comparison. The orientations are rather different, with more of the imagery of electronics and computation in the one case and of neurophysiology and adaptive networks in the other; also, the analysis-by-synthesis model (at least in its early forms) implies that comparison is done on the receptive side, i. e., that the incoming speech patterns are matched with implicit auditory patterns generated by a motor mechanism, whereas the model presented here calls for a comparison of implicit motor patterns, hence, on the other side of the system. It is obvious, of course, that neither model will be found intact under the skull, and it may be that even the attempt to localize the comparison function on the one side or the other has no meaning in neural reality. Preference, then, will rest on predictive power – and so on further research.

[23] These might have been called "distinctive features" (in the Prague School sense) and this usage would also have acknowledged our deep obligation to ROMAN JAKOBSON and his colleagues. We have not done so just to avoid confusions between JAKOBSON's usage and the usage here. Whatever the overlap may prove to be, the defining operations that we use are different from those used by JAKOBSON et al., and so call for different terms. (See: ROMAN JAKOBSON, MORRIS HALLE and C. GUNNAR M. FANT, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. M. I. T. Press, 2nd Ed., 1963.)

[24] In this view, the segmental phoneme is a unit that specifies in quantal form *both* selection (i. e., the features) and relative timing, and so facilitates – at the very least – the conversion from the parallel mode employed in transmission to the serial order of message units as they exist on the phonological level.