Some Assumptions about Speech and How They Changed*

Alvin M. Liberman

My aim is to provide a brief account of the research on speech at Haskins Laboratories as seen from my point of view. In pursuit of that aim, I will scant most experiments and their outcomes, putting the emphasis, rather, on the changing assumptions that, as I understood them, guided the research or were guided by it. I will be particularly concerned to describe the development of those assumptions, hewing as closely as I can to the order in which they were made, and then either abandoned or extended.

My account is necessarily inaccurate, not just because I must rely in part on memory about my state of mind almost 50 years ago when, in the earliest stages of our work, we did not always make our underlying assumptions explicit, but, even more, for reasons that put a proper account beyond the reach of any recall, however true. The chief difficulty is in the relation between the theoretical assumptions and the research they were supposed to rationalize. Thus, it happened only once, as I see it now, that the assumptions changed promptly in response to the results of a particular experiment; in all other cases, they lagged behind, as reinterpretations of data that had accumulated over a considerable period of time. Moreover, theory was influenced, not only by our empirical findings, but equally, if even more belatedly, by general considerations of plausibility that arose when, stimulated by colleagues and by the gradual broadening of my own outlook, I began to give proper consideration to the special requirements of phonological communication and to all that is entailed by the fact that speech is a species-typical product of biological evolution. The consequence for development of theory was that, with the one exception just noted, I never changed my mind abruptly, but rather segued from each earlier position to each later one, leaving behind no clear sign to mark the time, the extent of the change, or the reason for making it. Faced now with having to describe a theory that was in almost constant flux, I can only offer what I must, in retrospect, make to seem a progress through distinct stages.

My account is also necessarily presumptuous, because I will sometimes say 'we' when I probably should have said 'I', and vice versa. In so doing, I am not trying to avoid blame for bad ideas that were entirely mine, nor to claim credit for good ideas that I had no part in hatching. It is, rather, that everything I have done or thought in research on speech has been profoundly influenced by my colleagues at the Laboratories. In most cases, however, I can't say with certainty just how or by whom. A consequence is that, of all the words I use in this chronicle, the most ambiguous by far are 'I' and 'we'.

'I' began to be confused with 'we' on a day in June, 1944, when I was offered a job at the Laboratories to work with Frank Cooper on the development of a reading machine for the blind, a device that would convert printed letters into intelligible sounds. Of course, we appreciated from the outset that the ideal machine would render the print as speech that the blind user had already mastered. Though that is done quite routinely now, it was, in 1944, far beyond the science and technology that was available to us. There were no optical character readers to identify the letters, and no rules for synthesis that would have enabled us to produce speech from their outputs. But, reassured by our assumptions about the relation of speech to language, of which more...
later, we did not think it critical that the machine should speak. Rather, we supposed that it had only to produce distinctive sounds of some kind that the blind would then learn to associate with the consonants and vowels of the language, much as we supposed they had done at an earlier stage of their lives with the sounds of speech. Thus conceived, our enterprise lay in the domains of auditory perception and discrimination learning, two subjects I had presumably mastered as a graduate student at Yale. Indeed, my dissertation had been based on experiments about discrimination learning to acoustic stimuli, and I was thoroughly familiar with the neobehaviorist, stimulus-response theory that was thought by my professors and me to account for the results. In the course of my education in psychology, I had learned nothing about speech, but I didn’t think that mattered, because the theory I had absorbed was supposed, like most other theories in psychology, to apply to everything. I was, therefore, enthusiastic about the job, confident that I knew exactly how to make the reading machine serve its intended purpose and so put the theory to practical use. In the event, the theory, and, indeed, virtually everything else I had learned, proved to be, at best, irrelevant and, at worst, misleading. I think it unlikely that I would ever have discovered that had it not been for two fortunate circumstances. One was the collaboration, from the outset with Frank Cooper, whose gentle but nonetheless insistent prodding helped me to accept that the theory might be wrong, and to see what might be more nearly right. The other was that speech lay constantly before me, providing an existence proof that language could be conveyed efficiently by sound, and thus setting the high standard by which I was bound to evaluate the performance of the acoustic substitutes for speech that our early assumptions led us to contrive. But for Frank Cooper, on the one hand, and speech on the other, I might still be massaging those substitutes and modifying the conditions of learning, satisfied to achieve trivial improvements in systems that were, by comparison with speech, hopelessly inadequate. As it was, experience in trying to find an alternative set of sounds brought Frank and, ultimately, me to the conclusion that speech is uniquely effective as an acoustic vehicle for language. It remained only to find out why.

But I get ahead of the story. To set out from the proper beginning, I should say why we initially believed there was nothing special about speech or its underlying processes, and where that belief led us, not only in the several stages of the reading machine work, but also in the development of, and early work with, a research synthesizer we called the Pattern Playback.

The assumptions about speech that have been made by us and others differ in many details. As I see it now, however, there is one question that stands above those details, dividing theories neatly into two categories: does speech recruit motor and perceptual processes of a general sort, processes that cut horizontally across a wide variety of behaviors; or does it belong to a special phonetic mode, a distinct kind of action and perception comprising a vertical arrangement of structures and mechanisms specifically and exclusively adapted for linguistic communication? I will use this issue as the basis for organizing my account of the assumptions we made, calling them ‘horizontal’ or ‘vertical’ according to the side of the divide on which they fall.

THE HORIZONTAL VIEW OF SPEECH AND THE DESIGN OF READING MACHINES

As it pertains to the perceptual side of the speech process, the horizontal view, which is how we saw speech at the outset, rests on three assumptions: (1) The constituents of speech are sounds. (2) Perception of these sounds is managed by processes of a general auditory sort, processes that evoke percepts no different in kind from those produced in response to other sounds. (3) The percepts evoked by the sounds of speech, being inherently auditory, must be invested with phonetic significance, and that can be done only by means of a cognitive translation. Accordingly, the horizontal view assumes a second stage, beyond perception, where the purely auditory percepts are given phonetic names, measured against phonetic prototypes, or associated with ‘distinctive features’. Seen this way, perceiving speech is no different in principle from reading script; listener and reader alike perceive one thing and learn to call it something else.

These assumptions were-and, perhaps, still are—so much a part of the received view that we could hardly imagine an alternative; it simply did not occur to us to question them, or even to make them explicit. At all events, it was surely these conventional assumptions that gave us confidence in our assumption that nonspeech sounds could be made to work as well as speech, for what they clearly implied was that the processes by which blind users would learn to ‘read’ our sounds would
differ in no important way from those by which they had presumably learned to perceive speech. Our task, then, was simply to contrive sounds of sufficient distinctiveness, and then design the procedures by which the blind would learn to associate them with language.

Auditory discriminability is all it takes

As for distinctiveness, I was, at the outset, firmly in the grip of the notion that it was just so much psychophysical discriminability, almost as if it were to be had simply by separating the stimuli one from another by a sufficient number of just-noticeable-differences. This notion fit all too nicely with our ability, even then, to engineer a print-to-sound machine that would meet the discriminability requirement, for such a machine had only to control selected parameters of its acoustic output according to what was seen by a photocell positioned behind a narrow scanning slit. Provided, then, that we chose wisely among the many possible ways in which the sound could be made to vary according to what the photocell saw, the auditory results would be discriminable, hence learnable.

In all these machines the sound would be controlled directly by the amount or vertical position of the print under the scanning slit. We therefore called them ‘direct translators’ to distinguish them from an imaginable kind we called ‘recognition machines’, having in mind a device that might one day be available to identify each letter, as present-day optical character readers do, and produce in response a preset sound of any conceivable type. Since we were in no position to build a recognition machine, we set our sights initially on a direct translator.

We were aware that a reading machine of the direct-translation kind had been constructed and tested in England just after World War I. This machine, called the Optophone, scanned the print with five beams of light, each modulated at a different audio frequency. When a beam encountered print, a tone was made to sound at a frequency (hence, pitch) equal to the frequency at which that beam was modulated. Thus, as the scanning slit and its beams were moved across the print, the user would hear a continuously changing musical chord, the composition of which depended, instant by instant, on the modulation frequencies of the beams that struck print. The user’s task, of course, was to learn to interpret the changing chords as letters and words. I recall not ever knowing what tests, if any, had been carried out to determine just how useful the machine was, only that a blind woman in Scotland had been able to make her way through simple newspaper text. We did know that the machine was not in use anywhere in 1944, so we could only assume that it had been found wanting. At all events, the lesson I took from the Optophone was not that an acoustic alphabet is no substitute for speech, but that the particular alphabet produced by the Optophone was not a good one. Moreover, it seemed likely, given the early date at which the machine had been made, that the signal was less than ideally clear. I, perhaps more than Frank, was sure we could do better. (Indeed, I think that Frank, even at this early stage, had reservations about any machine that read letter by letter, but he shared my belief that we could design a letter-reading machine good enough to be of some use.)

Doing better, as we conceived it then, simply meant producing more learnable sounds. Unfortunately, there was nothing in the literature on auditory perception to tell us how to do that, so we set out to explore various possibilities. For that purpose, Frank built a device with which we could simulate the performance of almost any direct translator, with the one limitation that the subjects of our experiments were not free to choose the texts or to control the rate of scan. (This was because the print had to be scanned by passing a film negative of the text across a fixed slit.) Otherwise, the simulator was quite adequate for producing the outputs we would want to screen before settling on those that deserved further testing.

For much of the initial screening, Frank and I made the evaluations ourselves after determining roughly how well we could discriminate the sounds for a selected set of letters and words. On this basis, we quickly came to the conclusion that there was, perhaps, some promise in a system that took from the Optophone was not that an acoustic alphabet is no substitute for speech, but that the particular alphabet produced by the Optophone was not a good one. Moreover, it seemed likely, given the early date at which the machine had been made, that the signal was less than ideally clear. I, perhaps more than Frank, was sure we could do better. (Indeed, I think that Frank, even at this early stage, had reservations about any machine that read letter by letter, but he shared my belief that we could design a letter-reading machine good enough to be of some use.)

Doing better, as we conceived it then, simply meant producing more learnable sounds. Unfortunately, there was nothing in the literature on auditory perception to tell us how to do that, so we set out to explore various possibilities. For that purpose, Frank built a device with which we could simulate the performance of almost any direct translator, with the one limitation that the subjects of our experiments were not free to choose the texts or to control the rate of scan. (This was because the print had to be scanned by passing a film negative of the text across a fixed slit.) Otherwise, the simulator was quite adequate for producing the outputs we would want to screen before settling on those that deserved further testing.

For much of the initial screening, Frank and I made the evaluations ourselves after determining roughly how well we could discriminate the sounds for a selected set of letters and words. On this basis, we quickly came to the conclusion that there was, perhaps, some promise in a system that used the print to modulate the frequency of the output signal. (Certainly, this was very much better than modulating the amplitude, which had been tried first.) Having made this early decision, we experimented with different orientations and widths of the slit, the shape of the sound wave, and the range of modulation, selecting, finally, a vertical orientation of a slit that was somewhat narrower than, say, the vertical bar of a lower case 't', and a sine wave with a modulation range of 100 to 4000 Hz.

Further informal tests with this frequency-modulation (FM) system showed that it failed to provide a basis for discriminating certain letters (for example, n and u), which led us to conclude that it would likely prove in more thorough tests
to be not good enough. This conclusion did not discourage me, however, for I reckoned that the difficulty was only that the signal variation, being one-dimensional, was not sufficiently complex. Make the signals more complexly multidimensional, I reasoned, so that somewhere in the complexity the listener would find what he needed as the basis for associating that signal with the letter it represented. I recall that here, too, Frank was skeptical, but, open minded as always, he agreed to try more elaborate arrangements. In one, the Dual FM, the slit was divided into an upper and lower half (each with its own photocell), so that two tones, one in the upper part of the 4000-Hz range, the other in the lower, would be independently modulated. In another (Super FM), the slit was divided into thirds (each with its own photocell), and the difference between the output of the middle cell and the sum of the upper and lower cells controlled the (relatively slow) frequency modulations of a base tone, while the sum of all three cells controlled an audio frequency at which the base tone was itself frequency modulated. The effect of this latter modulation was to create side bands on the base frequency at intervals equal to the modulation frequency, and thus cause frequent, sudden, and usually gross changes in timbre and pitch as a consequence of the amount and position of the print under the slit. At all events, the signals of the Super FM seemed to me to vary in wonderfully complex ways, and so to provide a fair test of my assumption about the necessary condition for distinctiveness and learnability. We also simulated the Optophone, together with a variant on it in which, in a crude attempt to create consonant-like effects, we had the risers and descenders of the print produce various hisses and clicks. We even tried an arrangement in which the print controlled the frequency positions of two formant-like bands, giving an impression of constantly changing vowel color.

We tested all of these systems, together with three or four others of similar type, by finding how readily subjects could learn to associate the sounds with eight of the most common four-letter words. To determine how well these systems did in relation to speech, we applied the same test to an artificial language that we spoke simply by transposing phonological segments. Thus, vowels were converted into vowels, stops into stops, fricatives into fricatives, etc., so the syllable structure remained unchanged, hence easy for our human reading machine to pronounce. We called this new language 'Wuhzi', in honor of one of the transposed words.

The results of the learning test were simple enough: of all the nonspeech signals tested, the original, simple FM and the original Optophone were the most nearly learnable; all others trailed far behind. Worst of all, and by a considerable margin, were the extremely complex signals of the Super FM, the system for which I had entertained such high hopes. As for Wuhzi, the transposed speech, it was in a class of its own. It was mastered (for the purposes of the test) in about 15 trials, in contrast to the nonspeech sounds, on which subjects were, after 24 trials, performing at a level of only 60% with the simple FM and Optophone signals, and at 50% or lower for the others, and all seemed at, or close to, their asymptotes. More troubling was the fact that learnability of the nonspeech signals went down appreciably as the rate of scan was increased, even at modest levels of rate. What should have been most troubling, however, were the indications that learning would be rate specific. Though we did not pursue the point, it seemed clear in my own experience with the FM system that what I had learned to 'read' at a rate of, say, 10 words per minute, did not transfer to a rate of 50. My impression, as I remember it now, was not that things sounded much the same, only speeded up, but rather that the percept had changed entirely, with the result that I could no longer recognize whatever I had learned to respond to at the slower speed. To the extent that this observation was correct, users would have had to learn a different nonspeech 'language' for every significantly different rate of reading, and that alone would have made the system impractical. It also became clear that at rates above 50 or so words per minute, listeners could identify the words only on the basis of some overall impression—I am reluctant to call it a pattern—in which the individual components were not readily retrievable. The consequence of this, though we did not take proper account of it at the time, was that perception of the nonspeech signals would have been, at best, logophonic, as it were, not phonologic, so users would have had to learn to identify as many sounds as there were words to be read, and would not have been able to fall back on their phonologic resources in order to cope with new words.

At this point it had become evident that, as Frank had been saying for some time, perception of the acoustic alphabet produced by any letter-
reading machine would be severely limited by the
temporal resolving power of the ear—that is, by
its poor ability to segregate and properly order
acoustic events that are presented briefly and in
rapid succession. Taking account of the number of
discrete frequency peaks in each alphabetic sound
and the average number of alphabetic characters
per word, we estimated the maximum rate to be
around 50 words per minute, which I now believe
to have been very much on the high side of what
might have been achieved. Still, 50 words per
minute would be quite useful for many purposes,
and we had a considerable investment in the
letter-reading direct-translator kind of system, so
we felt obliged to see how far a subject could go,
given control of the scanning rate, a great variety
of printed material, and much practice. Frank
therefore built a working model of a direct­
translating FM reading machine, and we set
several subjects to work trying to master it. After
90 hours of practice, our best subject was able to
manage a fifth-grade reader (without correction)
at an average speed of about four words per
minute. Moreover, the subject seemed to have
reached her peak performance during the first 45
hours. Comparing performance at the end of that
period with performance after 90 hours, we found
no increase in rate and no decrease in errors.
Tests of the same system conducted by John Flynn
at the Naval Medical Research Laboratory yielded
results that gave no more grounds for optimism.

I must not omit from this account a reference to
our experience with a sighted subject named (dare
I say, appropriately?) Eve, who, having first
produced a properly shaped learning curve,
atained a reading speed of over 100 words per
minute, and convincingly demonstrated her skill
before the distinguished scientists who were
sponsoring our research. In that demonstration,
as in all of her work with the machine, she wore
blacked-out welder's goggles so she would not
suffer the discomfort of having to keep her eyes
shut tight. We all remember the day when one of
our number, acting on what seemed an
ungentlemanly impulse, put an opaque screen
between her and the print she was 'reading', at
which point she bounced from her chair and fled
the lab, confessing later to the young man who
had been monitoring her practice sessions that, by
leaning her cheek against her fist, she had, from
the very beginning, been raising the bottom of the
goggles and so seeing the print. In designing the
training and testing procedures, I had controlled
for every possibility except that one. Obviously, I
was the wrong kind of psychologist.

Having concluded at last that, given the prop­
terties of the auditory system, letter-by-letter read­ing machines were destined to be unsatisfactory,
we experimented briefly with a system in which
the sound was controlled by information that had
been integrated across several letters. This was
intended to reduce the rate at which discrete
acoustic events were delivered to the ear, and so
circumvent the limitation imposed by its temporal
resolving power. I cannot now imagine what we
thought about the consequences of having to as­
ociate holistically different sounds with linguisti­
cally arbitrary conjunctions of letters. But what­
ever it was that we did or did not have in mind,
this integrating type of reading machine failed ev­
ery test and was quickly abandoned.

One other scheme we tested deserves mention if
only because it now seems incredible that we
should ever have considered it. I suppose we felt
obliged to look forward to the time when optical
character readers would be available, and so to
test some presumably appropriate output of the'
recognition' machine that would then be possible.
Prerecorded 'phonemes' seemed an obvious choice.
After all, the acoustic manifestations of phonemes
were known to be distinctive, and people had, on
the horizontal view, already learned to connect
them to language. As for difficulty with rate, we
must have supposed that these sounds carried
within them the seeds of their own integration
into larger wholes. At all events, we recorded
various 'phonemes' on film—'duh' for /d/, 'luh' for
/l/, etc.—and then carefully spliced the pieces of
film together into words and sentences. The result
was wholly unintelligible, even at moderate rates,
and gave every indication of forever remaining so.

Perhaps auditory patterning is the answer

From all this experience with nonspeech we
drew two conclusions, one right, one wrong. The
right conclusion was that an acoustic alphabet is
no way to convey language. Of course, it was
humbling for me to realize that I might have
reached that conclusion without all the work on
the reading machines, simply by measuring the
known temporal-resolving characteristics of the
ear against the rate at which the discrete sounds
would have to be perceived. As I have already
suggested, Frank must have thought this through,
though he might not have been sufficiently pes­
simistic about the limits, and that would account
for his early opinion that an acoustic alphabet
might be at least marginally useful. I, on the other
hand, took the point only after seeing our early re­
results and having my nose rubbed in them.
The wrong conclusion—wrong because it was still uncompromisingly horizontal in outlook—was that what we needed was not discriminability but proper auditory patterning. Speech succeeds, we decided, because its sounds conform to certain Gestalt-like principles that yield perceptually distinctive patterns, and cause the phonetic elements to be integrated into the larger wholes of syllables and words. Apparently, it did not occur to me that phonological communication would be impossible if the discrete and commutable elements it requires were to be absorbed into indissoluble Gestalten. Nor did I bother to wonder how speakers might have managed their articulators so as to force all of the many covarying but not-independently-controllable aspects of the acoustic signal to collaborate in such a way as to produce just those sounds that would form distinctive wholes.

In any case, nobody knew what the principles of auditory pattern perception were. Research in audition, unlike the corresponding effort in vision, had been quite narrowly psychophysical, having been carried out with experimental stimuli that were not sufficiently complex to reveal such pattern perception as there might be. So if, for the purposes of the reading machine, we were going to make its sounds conform to the principles that underlie good auditory patterns, we had first to find the principles. We therefore decided to turn away from the reading machine work until we should have succeeded in that search. We were the more motivated to undertake it, because the outcome would not only advance the construction of a reading machine, but also count, more generally, as a valuable contribution to the science of auditory perception.

THE PATTERN PLAYBACK: MAKING THE RIGHT MOVE FOR THE WRONG REASON

It is exactly true, and important for understanding how horizontal our attitude remained, to say that our aim at this stage was to study the perception of patterned sounds in general, not speech in particular. As for the reading machine, we did not suppose we would use our results to make it speak, only that we would have it produce sounds that would be as good as speech because they would conform to the same principles of auditory perception. These would be sounds that, except for the accidents of language development, might as well have been speech. Obviously, we would look to speech for examples of demonstrably well-patterned sounds. Indeed, we would be especially attentive to speech, but only because we did not know where else to search for clues. At this stage, however, our interest lay in auditory perception.

Since we barely knew where to begin, we expected to rely, especially at the outset, on trial and error. Accordingly, we had to have some easy way to produce and modify a large number and wide variety of acoustic patterns. The sound spectrograph and the spectrograms it produced were no longer a wartime secret, and, like many others, we were impressed with the extent to which spectrograms made sense to the eye. Beyond a doubt, they provided an excellent basis for examining speech; it seemed to us a small step to suppose that they would serve equally well for the purpose of experimenting with it. The experimenter would have immediately available the patterned essence of the sound, and thus easily grasp the parameters that deserved his attention. These would be changed to suit his aims, and he would then have only to observe the effects of the experimental changes on the sound as heard. But that required a complement to the spectrograph—that is, a device that would convert spectrograms, as well as the changes made on them, into sound.

It was in response to these aims and needs that the Pattern Playback was designed and, in the case of its most important components, actually built by Frank Cooper. My role was simply to offer occasional words of encouragement, and, as the device took shape, to appreciate it as a triumph of design, a successful realization of Frank's determination to provide experimental and conceptual convenience to the researcher. The need for that convenience is, I think, hard to appreciate fully, except as one has lived through the early stages of our speech research, when little was known about acoustic phonetics, and progress depended critically, therefore, on the ability to test guesses at the rate of dozens per day.

Seven years elapsed between the start of our enterprise and the publication of the first paper for which we could claim significance. I would therefore here record that the support of the Haskins Laboratories and the Carnegie Corporation of New York was a notable and, to Frank and me, essential exercise of faith and patience. Few universities or granting agencies would then have been, or would now be, similarly supportive of two young investigators who were trying for such a long time to develop an unproved method to investigate a question—what is special about speech that it works so well?—that was not otherwise being asked. I managed to survive in academe by changing universities frequently, thus
confusing the promotion-tenure committees, and also by loudly trumpeting two rat-learning studies that I published as extensions of my thesis research.

There were many reasons it took so long to get from aim to goal. One was the failure, after two years of work, of the first model of the device that was to serve as our primary research tool. It was the second and very different model (hereafter the Pattern Playback or simply the Playback) that was, for our purposes, a success. In the form in which it was used for so many years, this device captured the information on the spectrogram by using it to control the fate of frequency-modulated light beams. Arrayed horizontally so as to correspond to the frequency scale of the spectrogram, there were 50 such modulated beams, comprising the first 50 harmonics of a 120 Hz fundamental. Those beams that were selected by the pattern of the spectrogram were led to a phototube, the resulting variations of which were amplified and transduced to sound. In effect, the device was, as someone once said, an optical player piano.

When the Pattern Playback was conceived, we thought we might not be able to work from hand-painted copies of spectrograms, but only from real ones. Therefore, we needed spectrograms on film so that, given photographic negatives, the phototube would see the beams that passed through the film. (When operating from hand-painted patterns, the phototube would receive the beams that were reflected from the paint.) For convenience in manipulating the patterns, we also wanted the frequency and time scales to be of generous dimensions. Moreover, we thought at this stage that we would need spectrograms with a large dynamic range. None of these needs was fully met by the spectrograph that had been built originally at the Bell Telephone Laboratories, and it was not available for sale in any case, so Frank set about to design and build our own. Unfortunately for the progress of our research, that took time. As for the Playback itself, it was, as our insurance inspector remarked when first he saw it, 'homemade'. And, indeed it was. Only the raw materials were store-bought. Everything else was designed and fashioned at the Laboratories, including, especially, the tone wheel, the huge circular film with 50 concentric rings of variable density used to produce the light beams that played on the spectrograms.

Once the special-purpose spectrograph and the Playback were, at last, in working order, we had to determine that speech could survive the rigors of the transformations wrought first by the one machine and then by the other. I well remember the relief I felt—Frank probably had more faith and therefore experienced correspondingly less relief—when, operating from a negative of a spectrogram, the Playback produced, on its first try, a highly intelligible sentence.

But that was only to pass the first test. For using these 'real' spectrograms as a basis for experimenting with speech would have been awkward in the extreme, since it would have been very hard to make changes on the film. We therefore began immediately to develop the ability to use hand-painted spectrograms to control the sound. For that purpose, we had first to find a paint that would wet the acetate tape (i.e., not bead up), be easily erased (without leaving scars), and dry quickly. (The last requirement became nearly irrelevant when, quite early in the work, we acquired a hair-dryer.) Unable to find a commercially available paint that met our specifications, we became paint manufacturers, making our way by trial and error to a satisfactory product. That much we had to do. But there was time spent in other preliminaries that proved to be quite unnecessary. Thus, overestimating the degree of fidelity to the original spectrogram we would need, we assumed we would have to control the relative intensities of the several parts of the pattern, and so devoted considerable effort to preparing and calibrating paints of various reflectances. We also supposed that we would have to produce gradations of a kind that could best be done with an airbrush, so we fiddled for a time with that.

But then, having decided that the price of fidelity was too high, we simply began, with an artist's brush and our most highly reflecting paint, to make copies of spectrographic representations of sentences, using for this purpose a set of twenty that had been developed at Harvard for work on speech intelligibility. Taking pains with our first copies to preserve as much detail as possible, we succeeded in producing patterns that yielded what seemed to us a level of intelligibility high enough to make the method useful for research. We were then further encouraged about the prospects of the method when, after much trial and error, we discovered that we could achieve even greater intelligibility—about 85% for the twenty sentences—with still simpler and more highly schematized patterns.

Having got this far, we ran a few preliminary experiments no different in principle from those, very common at the time, in which intelligibility was measured as a function of the filtering that the speech signal had been subjected to. But
instead of filtering, we presented the formants one at a time and in all combinations. The grossly statistical result seemed of no great interest, even then, so we did not pursue it. There was, however, one unintended and unhappy result of our interest in the relative contributions of the several formants. In the course of giving a talk about the Playback at a meeting of the Acoustical Society of America, Frank played several copy-synthetic sentences, first with each formant separately, and then with the formants in various combinations. It was plain from the reaction of the audience that everyone was greatly surprised each time Frank read out the correct sentence. Apparently, people had formed wrong hypotheses as they tried to make sense of the speech produced by the individual formants, and subsequently had trouble correcting those hypotheses when presented with the full pattern. So the sentences got nowhere near the 85% intelligibility they deserved, and Frank got little recognition on that occasion for a promising research method.

AN EXCURSION INTO NONSPEECH AND THE INFINITELY HORIZONTAL HYPOTHESIS

We were, at this stage, strongly attracted to the notion that spectrograms of speech were highly readable because, as tranformations of well-patterned acoustic signals, they managed still to conform to basic principles of pattern perception. Implicit in this notion was the assumption that principles of pattern perception were so general as to hold across modalities, provided proper account was taken of the way coordinates were to be transformed in moving from the one modality to the other. As applied to the relation between vision and audition, this assumption could be tested by the extent to which patterns that looked alike could be so transformed as to make them sound alike. To apply this test to the spectrographic transform, we varied the size and orientation of certain easily identifiable geometric forms, such as circles, squares, triangles, and the like, converted them to sound on the Playback, and then asked whether listeners would categorize the sounds as they had the visual patterns. Under certain tightly constrained circumstances, the answer was yes, they would; otherwise, it was no, they would not. At all events, we were so bold as to publish the idea, together with a description of the Playback in the Proceedings of the National Academy of Sciences. Fortunately for us, that journal is not widely read in our field, so our brief affair with the mother of all horizontal views has remained till now a well-kept secret.

We also experimented briefly with a phenomenon that would, today, be considered an instance of 'streaming', though I thought of it then as the auditory equivalent of the phi phenomenon that had for so long occupied a prominent place in the study of visual patterns. Using painted lines so thin that each one activated only a single harmonic, we observed that when they alternated between a lower and a higher frequency, the listener heard the alternation only if the resulting sinusoids were sufficiently close in frequency and sufficiently long in duration; otherwise, the impression was of two streams of tones that bore no clear temporal relation to each other. In pursuing this effect, we looked for that threshold of frequency separation and duration at which the subject could not distinguish a physically alternating pattern from one in which the two streams of tones came on and went off in perfect synchrony. We did, in fact, succeed in finding such thresholds and in observing that they were sensitive to frequency separation and duration, as our hypothesis predicted, but, after exhibiting a certain threshold for a while, the typical subject would quite suddenly begin to hear the difference and to settle down, if only temporarily again, at a new threshold. Discouraged by this apparent lability, we abandoned the project and began to put our whole effort on speech.

EARLY (AND STILL HORIZONTALLY ORIENTED) RESEARCH ON SPEECH

It was at about this point that Pierre Delattre came to the lab to visit for a day and, fortunately for us, stayed for ten years. Initially, his interest was in discovering the acoustic basis for nasality in French vowels, but once he found what the Playback was capable of, his ambition broadened to include any and all elements of phonetic structure. So, while continuing to work on nasality, Pierre applied himself (and us) to producing two-formant steady-state approximations to the cardinal vowels of Daniel Jones. (Pierre supplied the criterial ear.) The result was published in Le Maitre Phonetique, written in French and in a narrow phonetic transcription. Thus, in our second paper, we continued the habit that had been established in the first of so publishing as to guarantee that few among our colleagues would read what we wrote.

Meanwhile, we had begun to put our attention once again on the copy-synthetic sentences, but now, instead of inquiring into the contribution of
each formant to overall intelligibility, we sought, more narrowly, to find the acoustic bases—the cues—for the perception of individual phones. I recall working first on the word 'kill' as it appeared in simplified copy we had made from a spectrogram of the utterance, 'Never kill a snake with your bare hands.' Looking for the phone [l], we held the pattern stationary at closely spaced temporal intervals, thus creating at each resting point a steady-state signal controlled by the static positions of the formants at that instant. The result, to our ears, was a succession of vowel-like sounds, but nothing we would have called an [l]. Yet, running the tape through at normal speed, and thus capturing the movement of the formant, produced a reasonable approximation to that phone. I don't remember what we made of this (to me) first indication of the importance of the dynamic aspects of the signal, but we did not immediately pursue it, turning instead to the [k] at the beginning of the same word.

**Context-conditioned variability and the horizontal version of the motor theory**

The advantage of [k] from our point of view as experimenters was that it appeared, in the copy-synthetic word 'kill', to be carried by a clearly identifiable and nicely delimited cue: a brief burst of sound that stood apart from the formants. Since we knew that [k] was a voiceless stop, in the same class as [p] and [t], we reckoned that all three stops might depend on such a burst, according to its position on the frequency scale. So, after a little trial and error, we carried out what must be our first proper experiment. Bursts (about 360 Hz in height, 15 msec in duration at their widest, and of a shape that caused Pierre to call them 'blimps') were centered at each of 12 frequencies covering the full range of our spectrograms. Each burst was placed in front of seven of Pierre's steady-state cardinal vowels, and the resulting stimuli were randomized for presentation to listeners who would be asked to identify the consonant in each case as [p], [t], or [k].

Of all the synthetic patterns ever used, these burst-plus-steady-state-vowel 'syllables' were undoubtedly the farthest from readily recognizable speech. Indeed, they so grossly offended Pierre's phonetic sensibilities that he agreed only reluctantly to join Frank and me as a subject in the experiment. After discharging my own duty as a subject, I thought, as did Frank, that Pierre's reservations were well taken, and that our responses would reveal little about stop consonants or, indeed, anything else. We were therefore surprised and pleased when, on tabulating the results, we saw a reasonably clear and systematic pattern, and then equally surprised and pleased when a group of undergraduates, known for technical purposes as 'phonetically naive subjects', did much as we had done, if just a little more variably. For us and for them, the modal [k] response was for bursts at or slightly above the second formant of the vowel, wherever that was; [t] was assigned most often to bursts centering at frequencies above the highest of the [k] bursts; and a [p] response was given to most of the bursts not identified as [t] or [k].

This experiment was, for me, an epiphany. It is the one, referred to earlier, that changed my thinking within hours or days after its results were in. What caused the change was the finding that the effect of an acoustic cue depended to such a very large extent on the context in which it appeared, and, more to the point, that perception accorded better with articulatory gesture than with sound.

The effect of context was especially apparent in the fact that the burst most frequently perceived as [k] was the one lying at or slightly above the second formant of the following vowel, even though that formant sampled a range that extended from 3000 Hz (for [i]) at the one end, to 700 Hz (for [u]), at the other. Moreover, this evidence that the same phonetic percept was cued by stimuli that were acoustically very different was but one side of the coin, for the results also showed that, given the right context, different percepts could be evoked by stimuli that were acoustically identical. This was the case with the burst at 1440 Hz, which was perceived predominantly as [p] before [i], as [k] before [a], and then again as [p], though weakly, before [u].

As an empirical matter, then, this first, very crude experiment demonstrated a kind and degree of context-conditioned variability in the acoustic cues that subsequent research has shown to be pervasive in speech. As for its relevance to our earlier work on reading machines, it helped to rationalize one of the conclusions we had been brought to, which was that speech cannot be an acoustic alphabet; for what the context effects showed was that the commutable acoustic unit is not a phone, but rather something more like a syllable.

From a theoretical point of view, the results revealed the need to find, somewhere in the perceptual process, an invariant to correspond to the invariant phonetic unit, and they strongly suggested
that the invariant is in the articulation of the phone. Thus, in the case of the [k] burst we noted that it was the articulatory movement—raising and lowering the tongue body at the velum—that remained reasonably constant, regardless of the vowel, and, further, that coarticulation of the consonant with the variable vowels accounted for the extreme variability in the acoustic signal. As for the other side of the coin—the very different perception of the burst before different vowels—which resulted, it should be noted, from a fortuitous combination of circumstances probably unique to this experiment, we supposed that, in order to produce something like a burst at 1440 Hz in front of [i] or [u], one had to close and open at the lips, while in front of [a], closing and opening at the velum was required.

Taking all this into account, we adopted a notion—the Early Motor Theory—that I now believe to have been partly right and partly wrong. It was right, I think, in assuming that the object of perception in phonetic communication is to be found in the processes of articulation. It was wrong, or at least short of being thoroughly right, because it implied a continuing adherence to the horizontal view. As earlier indicated, that view assumes a two-stage process: first an auditory representation no different in kind from any other in that modality, followed, then, by linkage to something phonetic, presumably as a result of long experience and associative learning. The phonetic thing the auditory representation becomes associated with is, in the conventional view, a name, prototype, or distinctive feature. As we described the Early Motor Theory in our first papers, we made no such explicit separation into two stages, but only because our concern was rather to emphasize that perception was more closely associated with articulatory processes than with sound, and then to infer that this was because the listener was responding to the sensory feedback from the movements of the articulators. However, we offered no hint that phonetic perception takes place in a distinct modality, thus omitting to make explicit the assumption that is, as will be seen, the heart of the vertical view. Indeed, we implied, to the contrary, that the effects we were concerned with were, potentially at least, perfectly general, and so would presumably occur for any perceptual response to a physical stimulus, given long association between the stimulus and some particular muscular movement, together with its sensory feedback. What was special about speech was only that it provided the par excellence example of the opportunity for precisely that association to be formed. In any case, my own view, as I remember it now, did comprehend two more or less distinct stages: an initial auditory representation that, as a result of associative learning, ultimately gave way to the sensory consequences of the articulatory gesture that had always been coincident with it. I had, I must now suppose, not spent much time wondering exactly what 'gave way' might mean. Had I been challenged on this point, I think I should have said that in the early stages of learning there surely was a proper auditory percept, no different in kind from the response to any other acoustic stimulus and equally available to consciousness, but that later, when the bond with articulation was secure, this representation would simply have ceased to be part of the perceptual process. But however I might have responded, the Early Motor Theory was different from the standard two-stage view in a way that had an important effect on the way we thought about speech, and also on the direction of our research. It mattered greatly that we took the object of perception, and the ultimate constituent of phonetic structure, to be an articulatory gesture, not a sound (or its auditory result), for this began a line of thinking that would in time be seen to eliminate the need for the horizontalists' second stage, and so permit us to exorcise the linguistic ghosts—the phonetic names or other cognitive entities—that haunted it. As for the direction of our research, the Early Motor Theory caused us to turn our attention to speech production, and so to initiate the inquiry into that process that has occupied an ever larger and more important place in our enterprise.

Some mildly interesting assumptions that underlay the methods we used

Given the extreme unnaturalness of the stimuli in many of our experiments, and the difficulty subjects reported in hearing them as speech, we had to assume that there would be no important interaction between the degree of approximation to natural speech quality and the way listeners perceive the phonetic information. I therefore note here that this assumption has proved to be correct: no matter how unconvincing our experimental stimuli as examples of speech, those listeners who were nevertheless able to hear them as speech provided results that have held up remarkably well when the experiments were carried out subsequently with stimuli that were closer approximations to the real thing. This was the first indication we had of the theoretically impor-
tant fact that accurate information about the relevant articulator movements, as conveyed, however unnaturally, by the acoustic signal, is sufficient for veridical phonetic perception, provided only that the listener is not too put off by the absence of an overall speech-like quality.

We also had to assume that the validity of our results would be little affected by a practice, followed throughout our search for the cues, of investigating, in any one experiment, only a single phonetic feature in a single phonetic class, and instructing the subjects to limit their responses to the phones in that class. Obviously, this procedure left open the possibility that cues sufficient to distinguish phones along, say, the dimension of place in some particular condition of voicing or manner would not work when voicing or manner was changed, or when the set of stimuli and corresponding response choices was enlarged. In fact, the results obtained in the limited contexts of our experiments were not overturned in later research when, for whatever reason, those limits were relaxed. I take this to be testimony to the independence in perception of the standard feature dimensions.

Finally, given early indications that there were several cues for each phonetic element, and given that, to make our experiments manageable, at least in the early stages, we typically investigated one at a time, we had to assume the absence of strong interactions among the cues. In fact, as we later discovered, there are such interactions—specifically, the trading relations that bespeak a perceptual equivalence among the various cues for the same phone and that are, therefore, of some theoretical interest, as I will say later—but these occur only within fairly narrow limits, and they only change the setting of a cue that is optimal for the phone; they do not otherwise affect it. Therefore, working on only one cue at a time did not cause us to be seriously misled.

More context-conditioned variability and the dynamic aspects of the speech signal

The most cursory examination of a spectrogram of running speech reveals nothing so clearly as the almost continuous movement of the formants; even the band-limited noises that characterize the fricatives seem more often than not to be dynamically shaped. Indeed, Potter, Kopp, and Green, in their book, Visible Speech, had remarked these movements, but they considered the effect to proceed from (constant) consonant to (variable) vowel, at least in the case of stop consonant-vowel syllables; and since their interest was primarily in how these transitional influences might help people to 'read' spectrograms, they simply called attention to the direction of the movement, up or down, and did not speculate about the role of these movements in speech perception. Martin Joos, on the other hand, wrote explicitly about the consonant-vowel transitions, as well as their context-conditioned variability, and showed, by cutting out properly selected parts of magnetic-tape recordings, that these transitions conveyed information about the consonants. He also made the important observation that there was, therefore, no direct correspondence between the segmentation of the acoustic signal and the segmentation of the phonetic structure. But Joos could not vary the transition for experimental purposes, so his conclusions could not be further refined. We therefore thought it a reasonable next step to make those variations, and so learn more about the role of the transitions in perception, choosing, first, to study place of production among stop and nasal consonants.

Our research to that point had prepared us to deal only with two-formant patterns, and since inspection of spectrograms indicated that transitions of the first formant did not vary with place, we chose to experiment with transitions of the second. To that end, we varied the starting point, hence the direction and extent, of these transitions by starting them at each of a number of frequencies above and below the steady state of the following vowel. In one condition, the first formant had a fixed transition that rose to the steady state from the very lowest frequency (120 Hz); the resulting patterns were intended to represent voiced stops, and it was our judgment that they did that reasonably well. In a second condition, the first formant had a zero transition—that is, it was straight. We hoped that these patterns would sound voiceless, but, in fact, they did not. They were used nevertheless. The stimuli in each of these conditions were presented for identification as [b], [d], [g] to one group of listeners and as [p], [t], [k] to another. The results showed clearly that the transitions do provide important information about the place dimension of the voiced and voiceless stops, and, also, that this information is independent of voicing. Indeed, it mattered little whether the listeners were identifying a particular set of synthetic stops as voiced or voiceless; the same transitions were associated with the same place, and there was, at most, only slightly less variability in the condition with the rising first formant, where the stimuli sounded voiced to us and the subjects were asked to judge them so. In a
third condition, we strove, fairly successfully we thought, for nasal consonants by using a straight first formant, together with what we considered at the time to be an appropriate (fixed) nasal resonance. Here, too, the second-formant transitions provided information about place, and that information was in no way affected by the change in manner, even though we had reversed the patterns so as to make the nasals syllable final, and so to take account of the fact that the velar nasal never appears initially in the syllable in English. As for context effects, they were large and systematic. Thus, the best transition for [d] (or [t] or [n]) fell from a point considerably above the steady state of the vowel with [u], but with [i] it rose from a point below the vowel’s steady state, and similar effects were evident with the transitions for other phones. We thought it supportive of a motor theory that the highly variable transitions for the same consonant were produced by a reasonably constant articulatory gesture as it was merged with the gesture appropriate for the following vowel. Equally supportive, in our view, was the fact that mirror-image transitions in syllable-initial and syllable-final positions nevertheless yielded the same consonantal percept, for surely these would sound very different in any well-behaved auditory system. From an articulatory point of view, however, these transitions are seen as the acoustic consequences of the opening and closing phases of the same gesture. As with the bursts, then, perception cued by the transitions accorded better with an articulatory process than with sound.

Despite the demonstration, by us and others, of the extreme context sensitivity of the acoustic cues, some researchers have been concerned for many years to show that there are, nevertheless, invariant acoustic cues, implying, then, that no special theoretical exertions are necessary in order to account for invariant phonetic percepts. My own view of this matter has always been that, whatever the outcome of the seemingly never-ending search for acoustic invariants, the theoretical issue will remain largely untouched; for there is surely no question that the highly context-sensitive transitions do supply important information for phonetic perception—they can, indeed, be shown to be quite sufficient in many circumstances—and that incontrovertible fact must be accounted for.

A brief flirtation with binary decisions

In our first attempt to interpret the significance of the burst and transition results, we took seriously, if only for a short time, the possibility that the two kinds of cues collaborated in such a way that two binary decisions resolved all perceptual ambiguity. Our data had shown that the bursts were identified as [t], if they were high in frequency and as [p] or [k], if low; the second-formant transitions evoked [t] or [k], if they were falling, and [p], if rising. So a low burst and a rising transition would be an unambiguous [p]; a high burst and falling transition would be [t]; and a low burst coupled with a falling transition would be [k]. We were, of course, influenced to this conclusion not just by our data but also, if only indirectly, by the then prevailing fashion for binary arrangements. At all events, we made no attempt to link our notion about binary decisions with the Early Motor Theory, perhaps because that would have been hard to do.

Acoustic loci: Rationalizing the transitions and their role in perception

It required only a little further reflection, combined with an examination of the results of our experiments on the second-formant transitions, to see that perception was sensitive to something more than whether the transition was rising or falling. Since those transitions reflect the cavity changes that occur as the articulators move from the consonant position to the vowel, and, since the place of production for each consonant is more or less fixed, we saw that we should expect to find a correspondingly fixed position—or ‘locus’, as we chose to call it—for its second formant. More careful examination of the results of our experiments suggested that, for each position on the dimension of place, the transition might, indeed, have originated at some particular frequency, and then made its way to the steady state of the vowel, wherever that was. To refine this notion, we carried out a two-step experiment. In the first, we put steady-state second formants at each of a number of frequency levels, from 3600 Hz to 720 Hz, and paired each with one of a number of first formants in the range 720 Hz to 2400 Hz. The first formants had rising transitions designed to evoke the effect of a voiced stop consonant. Careful listening revealed that [d] was heard when the second formant was at 1800 Hz, [b] at 720 Hz, and [g] at 3000 Hz, so we settled on these frequencies as the loci for the places of production of the three stops.

The second step was to prepare two-formant patterns in which the second formant started at each of these loci, and then rose or fell to the steady state of the vowel, wherever that was. With
these patterns, the consonant appropriate to the locus was not evoked clearly. For [d], indeed, starting the transitions at the locus produced, for some steady states, [b] and [g]. To get good consonants, we had in all cases to ‘erase’ the first half of the transition so as to create a silent interval between the locus and the actual start of the transition. We noted, further, that in the case of [g], this maneuver worked only with second-formant steady states from 3000 Hz to about 1200 Hz, which is approximately where the vowel shifts from spread to rounded; below 1200 Hz, no approximation to [g] could be heard.

The concept of the locus, together with the experimental results that defined it, made simple sense of the transitions. It also tempted me to temper the emphasis on context-conditioned variability by assuming that the perceptual machinery—by which I might have meant the auditory machinery—‘extrapolated’ backward from the start of the transition to the locus, and so arrived at a ‘virtual’ acoustic invariant. Fortunately, I yielded to this temptation only briefly, and there is, I think, no written record of my lapse. In any case, we began early to take the opposite tack, using the locus data to strengthen our conclusions about the role of context by emphasising the un-toward consequences of actually starting the transitions at the locus, and by pointing to the sudden shift in the [g] locus when the vowel crosses the boundary from spread to rounded.

Stop vs. semivowel, or once more into the auditory breach

It was apparent on the basis of articulatory considerations, and also by inspection of spectrograms, that an acoustic correlate of the stop-semivowel distinction was the duration or rate of the appropriate transitions. To find out how this variable actually affected perception, we carried out the obvious experiment. We particularly wanted to know whether it was rate or duration, and also where on the rate or duration continuum the phonetic boundary was. By varying the positions of the vowel formants, and hence the extent of the transition, we were able to separate the two variables and find that duration seemed to be doing all the work. We also found that the boundary was at about 50 msec.

I recall thinking that 50 msec might be critical for some kind of auditory integration. As I have already said, it seemed reasonable to me at the time to suppose that phonetic distinctions had accommodated themselves to the properties of the auditory system as revealed at the level of psychophysical relations, since the (implicit) motor activity that was ultimately perceived was itself initially evoked by some kind of first-stage auditory representation. I was therefore naturally attracted to the possibility that transition excursions of less than 50 msec duration were, perhaps, so integrated by the auditory system as to produce a unitary impression like that of a stop consonant, while transitions with excursions longer than that would evoke the impression of gradual change that characterizes the semivowels. I well remember the excitement I felt when, having decided that a 50-msec duration might well be the auditory key, I appreciated how easy it would be to find out if, indeed, it was. The experimental test required only that I draw on the Playback a series of rising and falling isolated transitions in which the duration was varied over a wide range. What I expected and hoped to find was that a duration of 50 msec would provide a boundary between perception of a unitary stop-like burst of sound on the one side, and a semivowel kind of glide on the other. So far as I could tell, however, there appeared to be no such boundary at 50 msec and thus no evidence of an auditory basis for the results of our experiment on the distinction between stop and semivowel. I was disappointed, but not enough to abandon my horizontal attitude about the role of auditory representations in the ontogenetic development of phonetic perception.

Categorical perception: the right prediction from the wrong theories

According to just those aspects of the Early Motor Theory that I now believe to be mistaken, the auditory percept originally evoked by the speech signal was supposed to give way to the sensory consequences of the articulatory gesture, and it was just those consequences that were ultimately perceived. In arriving at this theory, I had, of course, been much influenced by the behaviorist stimulus-response tradition in which I had been reared. It was virtually inevitable, then, that I should take the next step and consider the consequences of two processes—‘acquired distinctiveness’ and ‘acquired similarity’—that were part of the same tradition. The point was simple enough: if two stimuli become connected, through long association, to very different responses, then the feedback from those responses, having become the end-states of the perceptual process, will cause the stimuli to be more discriminable than they had originally been; conversely, if these stimuli become connected to the same response, then, for
the same reason, they will be less discriminable. In fact, there was not then, and is not now, any evidence that such an effect occurs. But that did not trouble me, for I supposed that investigators had not thought to look in the right place, and I could not imagine a better place than speech perception. Neither was I troubled by what seems to me now the patently implausible assumption, basic to the concepts of acquired distinctiveness and acquired similarity, that the normal auditory representation of a suprathreshold acoustic stimulus could be wholly supplanted, or even significantly affected, by the perceptual consequences of some motor response just because the acoustic stimulus and the motor response had become strongly associated. It was for me compelling that listeners had for many years been making different articulatory responses to stimuli that happened to lie on either side of a phonetic boundary, so, by the terms of the Early Motor Theory and the theory of acquired distinctiveness, that difference should have become more discriminable. On the other hand, those listeners had been making the same articulatory response to equally different stimuli that happened to lie within the phonetic class, so, by the same theories, those stimuli should have been rendered less discriminable.

To test the theories, I thought it necessary only to get appropriate measures of acoustic-cue discriminability. As I know now, one can easily get the effect I sought simply by listening to voiced stops, for example, as the second-formant transition is changed in relatively small and equal steps, for what one hears is, first, several consonants that are almost identical [b]'s, then a rather sudden shift to [d], followed by several almost identical [d]'s, and then, again, a sudden shift, this time to [g]. Though we had the means to make this simple and quite convincing test, we did not think to try it. Instead, I put together two wrong theories and produced what my professors had taught me to strive for as one of the highest forms of scientific achievement: a real prediction.

The test of the prediction was initially undertaken by one of our graduate students, Belver Griffith, who, with my enthusiastic approval, elected to do the critical experiment on steady-state vowels. We know now that the effect we were looking for does not occur to any significant extent with such vowels, so it was fortunate that Griffith, by nature very fussy about the materials of his experiments, was unsuccessful in producing vowels he was willing to use. The happy consequence was that he, together with the rest of us, decided to move ahead, in parallel, with stop consonants.

It was not our purpose to obtain discrimination thresholds, but only to measure discriminability of a constant physical difference at various points on the continuum of second-formant transitions. To that end, we synthesized a series of 14 syllables in which the starting point of the second formant was varied in steps of 120 Hz, from a point 840 Hz below the steady state of the following vowel to a point 720 Hz above it. We then paired stimuli that were one, two, and three steps apart on the continuum, and for each such pair measured discriminability by the ABX method (A and B were members of the pair, X was one or the other, and the subject's task was to match X with A or B). The result was that there were peaks in the discrimination functions at positions on the continuum that corresponded to the phonetic boundaries as earlier determined by the way the subjects had identified the stimuli as [b], [d], or [g] when they were presented singly and in random order. This is to say that, other things equal, discrimination was better between stimuli to which the subjects habitually gave different articulatory responses than it was between stimuli to which the responses were the same. Thus, the prediction was apparently confirmed by this instance of what we chose to call 'categorical perception'.

In the published paper, we included a method, worked out by Katherine Harris, for computing from the absolute identification functions what the discrimination function would have been if, indeed, perception had been perfectly categorical—that is, if listeners had been able to perceive differences only if they had assigned the stimuli to different phonetic categories. Applying this calculation to our results, we found that, in this experiment at least, perception was rather strongly categorical, but not perfectly so.

To this point in our research, psychologists, including even those interested in language, had paid us little attention. Requests for reprints, and such other tokens of interest as we had received, had come mostly from communication engineers and phoneticians, and there were few references to our work in the already considerable literature of psycholinguistics, a field that had been established, seemingly by fiat, by a committee of psychologists and linguists who had met for a summer work session at Cornell. The result of their deliberations was a briefly famous monograph in which they defined the new
discipline, constructed its theoretical framework, and posed the questions that remained to be answered. That done, they officially launched the field at a symposium held during the next national convention of the American Psychological Association. At the end of the symposium, I asked a question that provoked one of the founding fathers to inform me, icily, that speech had nothing to do with psycholinguistics. He did not say why, but then, as one of the inventors of the discipline, he was entitled to speak ex cathedra. It is, however, easy to appreciate that in looking at speech horizontally, as he and the other members of the committee surely did, one sees nothing that is linguistically interesting, only a set of unexceptional noises and equally unexceptional auditory percepts that just happen to convey the more invitingly abstract structures where anyone who would think deeply about language ought properly to put his attention.

The categorical perception paper seemed, however, to touch a psycholinguistic nerve. Perhaps this was because, if taken seriously, it showed that the phonetic units were categorical, not only in their communicative role, but also as immediately perceived; and this nice fit of perceptual form to linguistic function must have seemed at odds with the conventional horizontal assumption that the auditory percepts assume linguistic significance only after a cognitive translation, not before. Our results could be taken to imply that no such translation was necessary, and that there might, therefore, be something psycholinguistically interesting and important about the precognitive—that is, purely perceptual—processes by which listeners apprehend phonetic structures.

Most psychologists seemed unwilling to accept that implication, though not all for the same reason. Some argued that categorical perception, as we had found it, was of no consequence because it was merely an artifact of our method. This criticism boiled down to an assertion, perfectly consistent with the standard horizontal view, that the memory load imposed by the ABX procedure made it impossible for the subject to compare the initial auditory representations, forcing him to rely, instead, on the categorical phonetic names he assigned to the rapidly fading auditory echoes as they were removed from the everchanging sensory world and elevated, for safer keeping, into short-term memory. It is, of course, true that reducing the time interval between the stimuli to be compared does raise the troughs that appear in the within-category parts of the discrimination function, and thus reduces the approximation to categorical perception. But the Early Motor Theory did not require that the articulatory responses within a phonetic category be identical, so it did not predict that perception had to be perfectly categorical. (The degree to which the articulatory responses within a category are similar presumably varies according to the category and the speaker; it is, therefore, a matter for empirical determination.) Neither did the theory in any way preclude the possibility that perceptual responses would be easier to discriminate when fresh than when stale. In any case, it surely was relevant that the peaks one finds with various measures of discrimination merely confirm the quantal shifts a listener perceives as the stimuli are moved along the physical continuum so rapidly as to make the memory load negligible.

The other criticism, which seemed almost opposite to the one just considered, was that categorical perception is not relevant to psycholinguistics because it is so common, and, more particularly, because those boundaries that the discrimination peaks mark are simply properties of the general auditory system, hence not to be taken as support for the view that speech perception is interesting, except, perhaps, within the domain of auditory psychophysics. As for the criticism that categorical perception is common, it seemed to have been based on the misapprehension that we had claimed categorical perception to be unique to speech, but in fact we had not, having merely observed (correctly) that, given stimuli that lie on some definable physical continuum, observers commonly discriminate many more than they can identify absolutely. Our claim about phonetic perception was only that there is a significant, if nevertheless incomplete, tendency for that commonly observed disparity to be reduced. On the other hand, the claim by our critics that the boundaries are generally auditory, not specifically phonetic, was important and deserved to be taken seriously.

It has led to many experiments on perception of nonspeech control stimuli and on perception of speech by nonhuman animals, leaving us and the other interested parties with an issue that is still vexed. In fact, I think the weight of evidence, taken together with arguments of plausibility, overwhelmingly favors the conclusion that the boundaries are specific to the phonetic system, but I reserve the justification for that conclusion to the last section of the paper.

There was yet another seemingly widespread misapprehension about categorical perception, which was that it had served us as the primary
basis for the Early Motor Theory. In fact, we (or, at least, I) have long believed that the facts about this phenomenon are consistent with the theory, but they do not by any means provide its most important support; indeed, they were not available until at least five years after we had been persuaded to the theory, as I earlier indicated, by the very first results of our search for the acoustic cues.

Finally, in the matter of categorical perception, I will guess that consonant perception is likely, when properly tested, to prove more nearly categorical than experiments have so far shown it to be. The problem with those experiments is that they have used acoustic synthesis, so it has been prohibitively difficult in any single experiment to make proper variations in more than one of the many aspects of the signal that are perceptually relevant. But when only one cue is varied, as in the experiments so far done, then, as it is changed from the form appropriate for one phone to the form appropriate for the next, it leaves all the other relevant information behind, as it were, creating a situation in which the listener is discriminating, not just the phonetic structure, but also increasingly unnatural departures from it. I suspect that, with proper articulatory synthesis, when the acoustic signal will change in all relevant aspects—at least for the cases that are produced by articulations that can be said to vary continuously—the discrimination functions will come much closer to being perfectly categorical.

The concept of ‘cue’ as a theoretically relevant entity

At the very least, ‘cue’ is a term of convenience, useful for the purpose of referring to any piece of signal that has been found by experiment to have an effect on perception. We have used the word in that sense, and continue to do so. But there was a time when cue had, at least in my mind, a more exalted status. I supposed that there was, for each phone, some specifiable number of particulate cues that combined according to a scientifically important principle to evoke the correct response. It was this understanding of cues that was implicit in the ‘binary’ account of their effects that I referred to earlier. The same understanding was more explicit in a dissertation on cue combination that I had urged on Howard Hoffman. Finally, and perhaps most egregiously, it became the centerpiece of our interpretation of an experiment on the effects of third-formant transitions on perception of place among the stops. Having found there that, in enhancing the perception of one stop, any particular transition does not do so equally at the expense of the other two, we concluded that a cue not only tells a listener what a speech sound is, but also which of the remaining possibilities it is not. It was almost as if we were supposing that the third-formant transition had been designed, by nature or by the speaker, just to resolve an ambiguity that the more important second-formant transition had overlooked. At all events, we went on to speculate that the response alternatives exist in a multidimensional phonetic space, and, though we were not perfectly explicit about this in the published paper, that a cue has a magnitude and a direction, just like a vector, with the result that the final position of the percept in the phonetic space is determined by the sum of the vectors. Such a conception is, of course, at odds with all the data now available that indicate how exquisitely sensitive the listener is to all the acoustic consequences of phonetically significant gestures, for what those data mean is that any definition of an acoustic cue is always to some extent arbitrary. Surely, it makes little sense to wonder about the rules by which arbitrarily defined cues combine to produce a perceptual result.

The voicing distinction; an exercise in not seeing that which is most visible

We discovered very early how to produce stops that were convincingly voiced, but we had been frustrated for five years or more while seeking the key to synthesis of their voiceless counterparts. In our quest, we had examined spectrograms, sought advice from colleagues in other laboratories, and, by trial and error on the Playback, tried every trick we could think of. We varied the strength of the burst relative to the vocalic section of the syllable, drew every conceivable kind of first-formant transition, and substituted various intensities of noise for the harmonics through varying lengths of the initial parts of the formant transitions.

In fact, there was no noise source in the Pattern Playback, only harmonics of a 120 Hz fundamental, but we had been able in research on the fricatives to make do by placing small dots of paint where noise was supposed to be. In isolation, patches of such dots sounded like a twittering of birds, but in syllabic context they produced fricatives so acceptable that, when we used them in experiments, we got results virtually identical to those obtained later when, with a new synthesizer called Voback, we were able to deploy
real noise. Before Voback was available, however, we had, in our attempts at voiceless stops, to rely on the trick that had worked for the fricatives. When it did not help, we concluded that, unlike the fricatives, the voiceless stops needed true noise and, accordingly, that our inability to synthesize them was to be attributed to the noise-producing limitations of the Playback.

That we were wrong to blame the Playback became apparent one day as a consequence of a discovery by Andre Malecot, one of Pierre's graduate students. While working to synthesize the syllable-final releases of stops, he omitted the first formant of the short-duration syllable that constituted the release. I believe that he did this inadvertently. But whether by inadvertence or by design, he produced a dramatic effect: we all heard a stop that was quite clearly voiceless. Encouraged by this finding, we adapted it to stops in syllable initial position, and carried out several related experiments. In each, we varied one potential cue for the voicing contrast for all three stops, paired with each of the vowels [i], [ae], and [u]. Our principal finding was that, with all else equal, simply delaying the onset of the first formant relative to the second and third was sufficient to cause naïve listeners to shift their responses smartly between voiced and voiceless.

Then, recognizing that in so delaying the onset of the first formant we were, at the same time, starting it at a higher frequency, we reconfirmed the observation we had made in our earlier experiments, which was that starting the first formant at a very low frequency was important in creating the impression of a voiced stop, but that starting it higher, at the steady state of that formant, did not, by itself, make much of a contribution to voicelessness. On the other hand, delaying the onset of the first formant without at the same time raising its starting frequency did prove to be a very potent cue. Indeed, it appeared from the responses of our listeners to be about as potent as the original combination of delay and starting point. (For the purpose of this experiment, we varied the delay alone by contriving a synthetic approximation to the vowel [o] in which the first formant was placed as low on the pattern as it could go; it was, then, just this straight formant that was delayed.) Next, we took advantage of the newly available synthesizer, Voback, which, as I earlier said, had a proper noise source, to experiment with the effect of noise in place of harmonics during the transitions. What we found was that substituting noise in all three formants was, by itself, ineffective, but that substituting it for harmonics in the second and third formants for the duration of the delay in first-formant onset did somewhat strengthen the impression of voicelessness. In connection with this last conclusion, we noted that when, in an attempt to produce an initial [h], which is, of course, the essence of aspiration, we replaced the harmonics of all the formants with noise for the first 50 or 100 msec, we did not get [h], but rather the impression of a vowel that changed from whispered to fully voiced; to get [h], we had to omit the first formant. To explain all this, we advanced a suggestion, made to us by Gunnar Fant, that the vocal cords were open during the period of aspiration, and that it was this circumstance that reduced the intensity of the first formant, thus effectively delaying its onset.

We emphasized, then, that all the acoustically diverse cues were consequences of the same articulatory event, and therefore led, in accordance with the Motor Theory, to a percept that was perfectly coherent.

This early work on the voicing distinction was subsequently refined and considerably extended by Arthur Abramson and Leigh Lisker. In particular, it was they who established how the acoustic boundaries for the voicing distinction vary with different languages, and thus provided the basis for the great volume of later research by other investigators who exploited these differences in pursuit of their interests in the ontogenesis of speech. And it was Abramson and Lisker who accurately characterized the relevant variable as voice-onset-time (VOT), defined as the duration of the interval between the consonant opening (in the oral part of the tract) and the onset of voicing at the larynx. Unfortunately, some of the researchers who later used the voicing distinction for their own purposes ignored the fact that the VOT variable is articulatory, not acoustic, and therefore failed to take into account in their theoretical interpretations that its acoustic manifestations are complexly heterogeneous.

As for our initial discovery of the acoustic cues for the voicing distinction, I note an irony in the long search that preceded it, for once one knows where to look in the spectrogram, the delay in the first formant onset can be measured more easily, and with greater precision, than almost any of the other consonant cues we had found. Consider, for example, how important to perception is the frequency at which a formant transition starts, and then how hard it is to specify that frequency precisely from an inspection of its appearance on a spectrogram. Yet our tireless examination of
spectrograms had, in the case of the voicing distinction, availed us nothing; we simply had not seen what we now know to be so plainly there.

**Synthesis by rule and a reading machine that speaks**

In this very personal chronicle of our early research, I have chosen to write of just those experiments that best illustrate certain underlying assumptions I now find interesting. I have said nothing about the many other experiments that were carried out during roughly the same period of time. I would now partly repair that omission by recognizing the existence of those others, and by emphasizing that they provided a collection of data sufficient as a basis for synthesizing speech from a phonetic transcription, without the need to copy from a spectrogram. Unfortunately, all the relevant data had been brought together only in Pierre's head. Relying only on the experience he had gained from participation in our published research, and also from the countless unpublished experiments he had carried out in his unflagging effort to refine and extend, Pierre could 'paint' speech to order, as it were. But the knowledge that Pierre had in his head was, by its nature, not so public as science demands.

We therefore recommended to Frances Ingemann, when she came to spend some time at the Laboratories, that she write a set of rules for synthesis, making everything so explicit that someone totally innocent of knowledge about acoustic phonetics could, simply by following the rules, draw a spectrogram that would produce any desired phonetic structure. Accepting this challenge, she decided to rely entirely on the papers we had published in journals or in lab reports; she did not use the synthesizer to test and improve as she went along, nor did she attempt to formalize what Pierre and other members of the staff might know but had never written down. She nevertheless succeeded very well, I think, in producing what must count as the first rules for synthesis. I don't recall that we ever formally assessed the intelligibility of the speech produced by these rules, but I know that we found it reasonably intelligible. At all events, an outline of the work he had done previously with Holmes and Shearme in England, and also on all that had been learned about the cues and rules for synthesis at the Laboratories. By 1968 the job was done. Accepting an input of a phonetic string, the system would speak. The intelligibility of the speech was tested on several occasions and in several different ways. Thus, it was tested informally by having blind veterans listen, for example, to rather long passages from Dickens and Steinbeck. It was evident that they understood the speech, even at rates of 225 words per minute, but we had no measure of exactly how hard they found it to do so, and they did complain, not without reason we thought, of what they called the machine's 'accent'. In more formal tests, the rule-generated synthetic speech came off quite well by comparison with 'real' speech, but, not unexpectedly, there was evidence of a price exacted by the extra cognitive effort that was required to overcome its evident imperfections, a price that had to be paid, presumably, by the processes of comprehension.

At that point, we had in hand a principal component of a reading machine that would convert text to speech, and thus avoid all the problems we had encountered in our earlier work with nonspeech substitutes. What was needed, in addition, was an optical character reader to convert the letters into machine-readable form, and also, of course, some way of translating spelled English into a phonetic transcription appropriate to the synthesizer. Given our history, it was inevitable that we should have been impatient to acquire these other components and see (or, more properly, hear) what a fully automatic system could do. So, Frank, Patrick Nye, and others cobbled together just such a system, using an optical character reader we bought with money given us for the purpose by the Seeing Eye Foundation, a phonetic dictionary...
made available by the Speech Communications Research Laboratory of Santa Barbara, and, of course, our own computer-controlled synthesizer. Tests revealed that the speech produced in this fully automatic way was almost as good as that for which the phonetic transcription had been hand-edited. But we were concerned about the evidence we had earlier collected concerning the probable consequences for ease of comprehension that arose out of the shortcomings of the speech. We therefore put together a plan to evaluate the machine with blind college students who would use it to read their assignments; having found its weaknesses, we would then try to correct them. I assumed that various federal agencies would compete to see which one could persuade us to accept their support for this undertaking. We could, after all, show that a reading machine for the blind was not pie-in-the-sky, but a do-able thing that stood in need of just the kinds of improvement that further research would surely bring. Yet, though we tried very hard with several agencies, and for several years, we failed utterly to get support, and were forced finally to abandon our plans. Still, we had the satisfaction of having proved to ourselves that a reading machine for the blind was close to being a reality. The basic research was largely complete; what remained was just the need for proper development.

ON BECOMING VERTICAL,
OR HOW I RIGHTED MYSELF

To this point, my concern has been to describe the various forms of the horizontal view that my colleagues and I held during our work on non-speech reading machines and in the early stages of the research on speech to which it led. Now I mean to offer a more detailed account of the important differences between that view and the vertical view I now hold. In so doing, I draw freely, and without specific attribution, on a number of theoretically oriented papers that were written in close collaboration with various of my colleagues. Among the most relevant of these are several reviews by Ignatius Mattingly and me in which we hammered out the vertical view as I (we) see it now.

All these theoretically oriented papers deal, at least implicitly, with questions about speech to which the horizontal and vertical views give different, sometimes diametrically opposed, answers. Such questions serve well, therefore, to define the two positions, and to explain how I came to abandon the one for the other; for those reasons, I will organize this section of the paper around them.

The issue that unites the questions pertains to the place of speech in the biological scheme of things. That I should have come to regard that issue as central is odd, given the habits of mind I had brought to the research, for, as I earlier implied, my education in psychology had been unremittingly abiological. I had, to be sure, studied a little physiology, narrowly conceived, and it cannot have escaped my notice that in the physiological domain things were not of a piece, having been formed, rather, into distinct systems for correspondingly distinct functions. At the level of behavior, however, I saw only an overarching sameness, a reflection of my attachment to principles so general as to apply equally to a process as natural as learning to speak and as arbitrary as memorizing a list of nonsense syllables.

I think I was moved first, and most generally, to a different approach by scientists who work, not on speech, but on other forms of species-typical behavior. Thus, it was largely under the influence of people like Peter Marler, Nobuo Suga, Mark Konishi, and Fernando Nottbohm that I came to see myself as an ethologist, very much like them, and to appreciate that I would be well advised to begin to think like one. They helped me to understand that speech is to the human being as echolocation is to the bat or song is to the bird—to see, that is, that all these behaviors depend on biologically coherent faculties that were specifically adapted in evolution to function appropriately in connection with events that are of particular ecological significance to the animal. To the horizontalist that I once was, this was heresy; but to the verticalist I was in process of becoming, it was the beginning of wisdom.

Meanwhile, back at the Laboratories there were biological stirrings on the part of Michael Studdert-Kennedy, who is nevertheless not a committed verticalist, and Ignatius Mattingly, who is. Michael has been a constructive critic in regard to virtually every biologically relevant notion I have dared to entertain. As for the biological slant of the vertical view (including the Revised Motor Theory), that is as much Ignatius's contribution as mine. Indeed, the view itself is the result of a joint effort, though, of course, he bears no responsibility for what I say about it here.

Among the influences of a somewhat different sort, there was the growing realization that my early horizontal view did not sit comfortably even with the results of the early research it was
designed to explain. That will have been seen in what I have already said about my attempts to account for those results, and, especially, about the patch on the horizontal view that I have here called the Early Motor Theory. Heavily loaded as it was with untested and wholly implausible assumptions—for example, that auditory percepts could, as a result of learning, be replaced by sensations of movement—it had begun to fall of its own weight.

Contributing further to the collapse of the Early Motor Theory was the research, pioneered by Peter Eimas and his associates, in which it was found that prelinguistic infants had a far greater capacity for phonetic perception than a literal reading of the theory would allow.

My faith was further weakened by the work of Katherine Harris and Peter MacNeilage, who, as the first of the Laboratories' staff to work on speech production, were busily finding a great deal of context-conditioned variability in the peripheral articulatory movements (as reflected in electromyographic measures), and thus disproving one of the assumptions of the Early Motor Theory, which was that the articulatory invariant was in the final-common-path commands to the muscles.

At the same time, Michael Turvey was pointing the way to an appropriate revision by showing how, given context-conditioned variability at the level of movement, it is nevertheless possible, indeed necessary, to find invariance in the more remote motor entities that Michael called 'coordinative structures'. In any case, Michael and Carol Fowler were strongly encouraging me to persevere in the aspect of the Early Motor Theory that took gestures to be the objects of speech perception, while simultaneously heaping scorn on the idea that perception was a second-order translation of a sensory representation, as the horizontal version of the theory required. I began, therefore, to take more seriously the possibility that there is no mediating auditory percept, only the immediately perceived gestures as provided by a system—the phonetic module—that is specialized for the ecologically important function of representing them.

Not that Michael and Carol or, indeed, any of the other 'ecological' psychologists in the Laboratories, are verticalists. They most certainly are not, because they do not accept (yet) that there is a distinct phonetic mode, preferring, rather, to take speech perception as simply one instance of the way all perception is tuned to perceive the distal objects; in the case of speech, these just happen to be the articulatory gestures of the speaker. Thus, I have been in the happy position of taking advantage of the best of what my ecological friends have had to offer, while freely rejecting the rest, and, as an important bonus, being stimulated by our continuing disagreements to correct weaknesses and repair omissions in my own view.

It was also relevant to the development of my thinking that Isabelle Liberman, Donald Shankweiler, and Ignatius Mattingly—followed later by such younger colleagues as Benita Blachman, Susan Brady, Anne Fowler, Hyla Rubin, and Virginia Mann—had begun to see in our research how to account for the fact that speech is so much more natural (hence easier) than reading and writing, and thus to be explicit about what is required of the would-be reader/writer that mastery of speech will not have taught him. As I will say later, their insights and the results of their empirical work illuminated aspects of the vertical view that I would otherwise not have seen.

I was affected, too, by the results of experiments on duplex perception, trading relations, and integration of cues, experiments that went beyond those, referred to earlier, that merely isolated the cues and looked for discontinuities in the discrimination functions. These later experiments, done (variously) in close collaboration with Virginia Mann, Bruno Repp, Douglas Whalen, Hollis Fitch, Brad Rakerd, Joanne Miller, Michael Dorman, and Lawrence Raphael (few of whom are admitted verticalists) provided data that spoke more clearly than the earlier findings to some of the shortcomings of the horizontal position, and therefore inclined me ever more strongly to the vertical alternative.

Finally, I should acknowledge the profound effect of Fodor's provocative monograph, 'The Modularity of Mind', which, in the early stages of my conversion, enlightened and stimulated me by its arguments in favor of the most general aspects of the vertical view.

That I should finally have asked the following questions, and answered them as I do, reflects the influences I have just described, and fairly represents the theoretical position to which they moved me.

In the development of phonological communication, what evolved?

Defined as the production and perception of consonants and vowels, speech, as well as the phonological communication it underlies, is plainly a species-typical product of biological
evolution. All neurologically normal human beings communicate phonologically; no other creatures do. The biologically important function of phonologic communication derives from the way it exploits the combinatorial principle to generate vocabularies that are large and open, in contrast to the vocabularies of nonhuman, nonphonologic systems, which are small and closed. Thus, phonological processes are unique to language and to the human beings who command them. It follows that anyone who would understand how speech works must answer the question: what evolved? Not when, or why, or how, or by what progression from earlier-appearing stages. The first question is simply: what?

The answer given by the horizontal view is clear: at the level of action and perception, nothing evolved; language simply appropriated for its own purposes the plain-vanilla ways of acting and perceiving that had developed independently of any linguistic function. Thus, those horizontalists who put their attention rather narrowly on the perceptual side of the process argue that the categories of phonetic perception simply reflect the way speech articulation has accommodated itself to the production of sounds that conform to the properties of the auditory system, a claim that I will evaluate in some detail later. A recent and broader, but still horizontal, take on the same issue distributes the emphasis more evenly between production and perception, arguing that phonetic gestures were selected by language on the basis of constraints that were generally motor, as well as generally auditory. However, the important point in this, as in the narrower view, is that the constraints are independent of a phonetic function, hence in no way specific to speech. Put forth as an explicit challenge to the vertical assumption, the broader view has it that there is no reason to assume a special mode for the production and perception of speech, if, with the proper horizontal orientation, one can see that the units of speech are optimized with respect to motor and perceptual constraints that are biologically general.

But the question is not whether language somehow developed out of the biology that was already there; surely, it could hardly have done otherwise. The question, to put it yet again, asks, rather, what did that development produce as the basis for a unique mode of communication? When the horizontalists say that the development of this mode was accomplished merely by a process of selection from among the possibilities offered by general faculties that are independent of language, they are giving an account that applies as well to the development of, say, a cursive writing system. Was not the selection of the cursive gestures similarly determined by motor and perceptual constraints that are independent of language? Yet, what that selection produced were not the biologically primary units of speech, but only a set of optical artifacts that had then to be connected to speech in a wholly arbitrary way. Of course, this is merely to say the obvious about the relation between speech and a writing system, which is that the evolution of the one was biologically, the other, not. That is surely a critical difference, but one that the horizontal view must have difficulty comprehending.

If pressed further to answer the question about the product of evolution, the horizontalists would presumably have to say that, while nothing evolved at the level of perception and action, there must have been relevant developments at a higher cognitive level. Thus, it would have been evolution that produced the phonetic entities of a cognitive type to which the nonphonetic acts and percepts of speech must, on the horizontal view, be associated. Being neither acts nor percepts, these cognitive entities—or ideas, as they might be—would presumably be acceptable within the horizontal framework as genetically determined adaptations for language, hence special in a way that speech is not allowed to be. In itself, this seems an unparsimonious, not to say biologically implausible, assumption. And it can be seen to be the more unparsimonious and implausible once the horizontalist tries to explain how the phonetically neutral acts and percepts got connected to the specialized cognitive entities in the first place. In the case of a script, to wring one more point out of that tired example, the obviously nonphonetic motor and visual representations of the writer and reader were connected to language by agreement among the interested parties. Can we seriously propose a similar account for speech?

If the horizontalists should reject the notion that phonetic ideas were the evolutionary bases for speech, there remains to them the most thoroughly horizontal view of all, which is that what evolved was a large brain. In that case, they might suppose either that phonological communication was an inevitable by-product of the cognitive power that such a brain provides, which seems unlikely, or that phonological communication was an invention, created by large-brained people who were smart enough to have appreciated the immense advantages for communication of the combinatorial principle, which seems absurd.
The vertical view is different on all counts. What evolved, on this view, was the phonetic module, a distinct system that uses its own kind of signal processing and its own primitives to form a specifically phonetic way of acting and perceiving. It is, then, this module that makes possible the phonological mode of communication.

The primitives of the module are gestures of the articulatory organs. These are the ultimate constituents of language, the units that must be exchanged between speaker and listener if linguistic communication is to occur. Standing apart as a class from the nonphonetic activities of the same organs—for example, chewing, swallowing, moving food around in the mouth, and licking the lips—these gestures serve a phonetic function and no other. Hence they are marked by their very nature as exclusively phonetic in character; there is no need to make them so by connecting them to linguistic entities at the cognitive level. As part of the larger specialization for language, they are, moreover, uniquely appropriate to other linguistic processes. Thus, the syntactic component is adapted to operate on the specifically phonetic representations of the gestures, not on representations of an auditory kind. Indeed, it is precisely this harmony among the several components of the language specialization that makes the epithet ‘vertical’ particularly apposite for the view I am here promoting.

Of course, the gestures constitute only the phonetic structures that the perceptual process extracts from the speech signal. Such aspects of the percept as, for example, those that contribute to the perceived quality of the speaker’s voice are not part of the phonetic system. Indeed, these are presumably auditory in the ordinary sense, except as they may figure in speaker identification, for which there may be a separate specialization.

It is not only the gestures themselves that are specifically phonetic, but also, presumably, their control and coordination. Surely, there is in speech production, as in all kinds of action, the need to cope with the many-to-one relations between means and ends, and also to reduce degrees of freedom to manageable proportions. In these respects, then, the management of speech and nonspeech movements should be subject to the same principles. But there is, in addition, something that seems specific to speech: the grossly overlapped and smoothly merged movements at the periphery are controlled by, and must preserve information about, relatively long strings of the invariant, categorical units that speech cares about but other motor systems do not. And, certainly, it is relevant to the claim about a specialized mode of production that speech, in the very narrowest sense, is species-specific: given every incentive and opportunity to learn, chimpanzees are nevertheless unable to manage the production of simple CVC syllables. (The fact that the dimensions of their vocal tracts presumably do not allow a full repertory of vowels should not, in itself, preclude the articulation of syllables with whatever vowels their anatomy permits.)

As for the evolution of the phonetic gestures, I should think an important selection factor was not so much the ease with which they could be articulated, or the auditory salience of the resulting sound, but rather how well they lent themselves to being coarticulated. For it is coarticulation that, as I will have occasion to say later, makes phonological communication possible.

But it is also this very coarticulation that, as we saw earlier, frustrates the attempt to find the phonetic invariant in the acoustic signal or in the peripheral movements of the articulators. Still, such motor invariants must exist, not just for the aspect of the Motor Theory that explains how phonetic segments are perceived, but for just any theory that presumes to explain how they are produced; after all, speech does transmit strings of invariant phonological structures, so these invariants must be represented in some way and at some place in the production process. But how are they to be characterized, and where are they to be found? Having accepted the evidence that they are not in the peripheral movements, as the Early Motor Theory assumed, Mattingly and I proposed in the Revised Motor Theory that attention be paid instead to the configurations of the vocal tract as they change over time and are compared with other configurations produced by the same gesture in different contexts. As for the invariant causes of these configurations, they are presumably to be found in the more remote motor entities—something like Turvey’s coordinative structures—that control the various articulator movements so as to accomplish the appropriate vocal-tract configurations. It is, I now think, structures of this kind that represent the phonetic primitives, providing the physiological basis for the phonetic intentions of the speaker and the phonetic percepts of the listener. Unfortunately for the Motor Theory, we do not yet know the exact characteristics of these motor invariants, nor can we adequately describe the processes by which they control the movements of the articulators. My colleagues, including especially Cathe
Browman, Louis Goldstein, Elliot Saltzman, and Philip Rubin, are currently in search of those invariants and processes, and I am confident that they will, in time, succeed in finding them. Meanwhile, I will, for all the reasons set forth in this paper, remain confident that motor invariants do exist, and that they are the ultimate constituents of speech, as produced and as perceived.

According to the Revised Motor Theory, there is a phonetic module, part of the larger specialization for language, that is biologically adapted for two complementary processes: one controls the overlapping and merging of the gestures that constitute the phonetic primitives; the other processes the resulting acoustic signal so as to recover, in perception, those same primitives. On this view, one sees a distinctly linguistic way of doing things down among the nuts and bolts of action and perception, for it is there, not in the remote recesses of the cognitive machinery, that the specifically linguistic constituents make their first appearance. Thus, the Revised Motor Theory is very different from its early ancestor; the two remain as one only in supposing that the object of perception is the invariant gesture, not the context-sensitive sound.

How is the requirement for parity met?

In all communication, whether linguistic or not, sender and receiver must be bound by a common understanding about what counts: what counts for the sender must count for the receiver, else communication does not occur. In the case of speech, speaker and listener must perceive, or otherwise know, that, out of all possible signals, only a particular few have linguistic significance. Moreover, the processes of production and perception must somehow be linked; their representations must, at some point, be the same. Though basic, this requirement tends to pass unnoticed by those who look at speech horizontally, and, especially, by those whose preoccupation with perception leaves production out of account. However, vertical Motor Theorists like Ignatius Mattingly and me are bound to think the requirement important, so we have given it a name—'parity'—and asked how, in the case of speech communication, it was established and how maintained.

Horizontalists must, I think, find the question very hard. For if, as their view would have it, the acts of the speaker are generally motor and the percepts of the listener generally auditory, then act and percept have in common only that neither has anything to do with language. The horizontalist is therefore required to assume that these representations are linked to language and to each other only insofar as speaker and listener have somehow selected them for linguistic use from the indefinitely large set of similarly nonphonetic alternatives, and then connected them at a cognitive level to the same phonetic name or other linguistic entity. Altogether, a roundabout way for a natural mode of communication to work.

For the verticalists, on the other hand, the question is easy. On their view, it was specifically phonetic gestures that evolved, together with the specialized processes for producing and perceiving them, and it is just these gestures that provide the common currency with which speaker and listener conduct their linguistic business. Parity is thus guaranteed, having been built by evolution into the very bones of the system; there is no need to arrive at agreements about which signals are relevant and how they are to be connected to units of the language.

How is speech related to other natural modes of communication?

I noted earlier that human beings communicate phonologically but other creatures do not, and, further, that this difference is important, because it determines whether the inventory of 'words' is open or closed. Now, in the interest of parsimony, I ask whether either view of speech allows that there is, nevertheless, something common to two modes of communication that are equally natural.

On the horizontal view, the two modes must be seen as different in every important respect. Nonhuman animals, the horizontalists would presumably agree, communicate as they do because of their underlying specializations for producing and perceiving the appropriate signals. I doubt that anyone would seriously claim that these require to be translated before they can take on communicative significance. For the human, however, the horizontal position, as we have seen, is that the specialization, if any, is not at the level of the signal, but only at some cognitive remove. I find it hard to imagine what might have been gained for human beings by this evolutionary leap to an exclusively cognitive representation of the communicative elements, except, perhaps, the smug satisfaction they might take in believing that they communicate phonologically, and the nonhuman animals do not, because they have an intellectual power the other creatures lack, and that even in the most basic aspects of communication they can count themselves broad generalists, while the others must be seen as narrow specialists.
On the vertical view, human and nonhuman communication alike depend on a specialization at the level of the signal. Of course, these specializations differ one from another, as do the vehicles—acoustic, optical, chemical, or electrical—that they use. And, surely, the phonetic specialization differs from all the others in a way that is, as we know, critical to the openness or generativity of language: Still, the vertical view permits us to see that phonetic communication is not completely off the biological scale, since it is, like the other natural forms of communication, a specialization all the way down to its roots.

What are the (special) requirements of phonological communication, and how are they met?

If phonology is to use the combinatorial principle, and so serve its critically important function of building a large and open vocabulary out of a small number of elements, then it must meet at least two requirements. The more obvious is that the phonological segments be commutable, which is to say discrete, invariant, and categorical. The other requirement, which is only slightly less obvious, concerns rate. For if all utterances are to be formed by stringing together an exiguous set of commutable elements, then, inevitably, the strings must run to great lengths. There is, therefore, a high premium on rapid communication of the elements, not only in the interest of getting the job done in good time, but also in order to make things reasonably easy, or even feasible, for those other processes that have got to organize the phonetic segments into words and sentences.

Consider how these requirements would be met if, as the horizontal view would have it, the elements were sounds and the auditory percepts they evoke. If it were these that had to be commutable, then surely it would have been possible to make them so, but only at the expense of rate. For sounds and the corresponding auditory percepts to be discrete, invariant, and categorical would require that the segmentation be apparent at the surface of the signal and in the most peripheral aspects of the articulation. How else, on the horizontal view, could commutability be achieved, except as each discrete sound and associated auditory percept were produced by a correspondingly discrete articulatory maneuver? Of course, the sounds and the percepts might be joined, as are the segments of cursive writing, and that might speed things up a bit, but, exactly as in cursive writing, the segmentation would nevertheless have to be patent. The consequence would be that, to say a monosyllabic word like 'bag', the speaker would have to articulate the segments discretely, and that would produce, not the monosyllable ‘bag’, but the trisyllable [boɪ əʊɡ]. To articulate the syllable that way is not to speak, but to spell, and spelling would be an impossibly slow and tedious way to communicate language.

One might imagine that if production had been the only problem in the matter of rate, nature might have solved it by abandoning the vocal tract, providing her human creatures, instead, with acoustic devices specifically adapted to producing rapid-fire sequences of sound. That would have taken care of the production problem, while, at the same time, defeating the ear. The problem is that, at normal rates, speech produces from eight to ten segments per second, and, for short stretches, at least double that number. But if each of those were a unit sound, then rates that high would strain the temporal resolving power of the ear, and, of particular importance to phonetic communication, also exceed its ability to perceive the order in which the segments had been laid down. Indeed, the problem would be exactly the one we encountered when, in the early work on reading machines, we presented acoustic alphabets at high rates.

According to the vertical view, nature solved the rate problem by avoiding the acoustic-auditory (horizontal) strategy that would have caused it. What evolved as the phonetic constituents were the special gestures I spoke of earlier. These serve well as the elements of language, because, if properly chosen and properly controlled, they can be coarticulated, so strings of them can be produced at high rates. In any case, all speakers of all languages do, in fact, coarticate, and it is only by this means that they are able to communicate phonologically as rapidly as they do.

Coarticulation had happy consequences for perception, too. For coarticulation folds information about several successive segments into the same stretch of sound, thereby achieving a parallel transmission of information that considerably relaxes the constraint imposed by the temporal resolving properties of the ear. But this gain came at the price of a relation between acoustic signal and phonetic message that is complex in a specifically phonetic way. One such complication is the context-conditioned variability in the acoustic signal that I identified as the primary motivation for the Early Motor Theory, presenting it then as if it were an obstacle that the processes postulated by the theory had to overcome. Now, on the Revised
Motor Theory, we can see that same variability as a blessing, a rich source of information about phonetic structure, and, especially, about order. Consider, again, the difficulty the auditory system has in perceiving accurately the ordering of discrete and brief sounds that are presented sequentially. Coarticulation effectively gets around that difficulty by permitting the listener to apprehend order in quite another way. For, given coarticulation, the production of any single segment affects the acoustic realization of neighboring segments, thereby providing, in the context-conditioned variation that results, accurate information about which gesture came first, which second, and so on. Hence, the order of the segments is conveyed largely by the shape of the acoustic signal, not by the way pieces of sound are sequenced in it. For example, in otherwise comparable consonant-vowel and vowel-consonant syllables, the listener is not likely to mistake the order, however brief the syllables, because the transitions for prevocalic and postvocalic consonants are mirror images. But these will have the proper perceptual consequence only if the phonetic system is specialized to represent the strongly contrasting acoustic signals, not as similarly contrasting auditory perceptions, but as the opening and closing phases of the same phonetic gesture. Accordingly, order is given for free by processes that are specialized to deal with the acoustic consequences of coarticulated phonetic gestures. Thus, we see that a critical function of the phonetic module is not so much to take advantage of the properties of the general motor and auditory systems—a matter that was briefly examined earlier—as it is to find a way around their limitations.

Could the assignment of the stimulus information to phonetic categories plausibly be auditory?

Many of the empirically based arguments about the two theories of speech, including, especially, most of those that have been advanced against the vertical position, come from experiments, much like those described in the first section of this essay, that were designed very simply to identify the information that leads to perception of phonetic segments. The results of these experiments have proved to be reliable, so there is quite general agreement about the nature of the relevant information. Disagreement arises only, but nonetheless fundamentally, about the nature of the event that the information is informing about. Is it the sound, as a proper auditory (and horizontal) view would have it, or the articulatory gesture, which is the choice of the vertically oriented Motor Theory.

The multiplicity of acoustic-phonetic boundaries and cues. Research of the kind just referred to has succeeded in isolating many acoustic variables important to perception of the various phonetic segments, and in finding for each the location of the boundary that separates the one segment from some alternative—for example, [ba] from [da]. The horizontalists take satisfaction in further experiments on some of these boundaries in which it has been found that they are exhibited by nonhuman animals, or by human observers when presented with nonspeech analogs of speech, for these findings are, of course, consistent with the assumption that the boundaries are auditory in nature. In response, the verticalists point to experiments in which it has been found that the boundaries differ between human and nonhuman subjects, and, in humans, between speech and nonspeech analogs, arguing, in their turn, that these findings support the view that the boundaries are specifically phonetic. Indeed, for some parties to the debate it has been in the interpretation of these boundaries that the difference between the two views has come into sharpest focus. The issue therefore deserves to be further ventilated.

It is now widely accepted that the location of the acoustic-phonetic boundary on every relevant cue dimension varies greatly as a function of phonetic context, position in the syllable, and vocal-tract dimensions. It is now also known, and accepted, that some vary with differences in language type, linguistic stress, and rate of articulation. For at least one of these—rate of articulation—the variation is possibly continuous. From all this it follows that the number of acoustic-phonetic boundaries is indefinitely large, far too large, surely, to make reasonable the assumption that they are properties of the auditory system. How would these uncountably many boundaries have been selected for as that system evolved? Surely, not just against the possibility that language would come along and find them useful. Indeed, as auditory properties, they would presumably be dysfunctional, since they are perceptual discontinuities of a sort, and would, therefore, cause continuously varying acoustic events to be perceived discontinuously, thereby frustrating veridical perception.

The matter is the worse confounded for the auditory theory when proper account is taken of the fact that, for every phonetic segment, there
are multiple cues, and that phonetic perception uses all of them. For if, in accounting for the perception of certain consonantal segments, we attribute an auditory basis to all the context-variable boundaries on, say, the second-formant transitions—already a dubious assumption, as we've seen—then what do we do about the third-formant transitions and the bursts (or fricative noises)? These various information-bearing aspects of the signal are not independently controllable in speech production, so one must wonder about the probability that a gesture so managed as to have just the 'right' acoustic consequences for the second-formant transition would happen, also, to have just the right consequences in all cases for the other, acoustically very different cues. On its face, that probability would seem to be vanishingly small.

Nor does it help the horizontal position to suggest, as some have, that the acoustic-phonetic boundaries exhibited by nonhuman animals served merely as the auditory starting points—the protoboundaries, as it were—to which all the others were somehow added. For this is to suppose that, out of the many conditions known to affect the acoustic manifestation of each phonetic segment, some one is canonical. But is it plausible to suppose that there really are canonical forms for vocal-tract size, rate of articulation, condition of stress, language type, and all the other conditions that affect the speech signal? And what of the further implications? What, for example, is the status of the countless other boundaries that had then to be added in order to accommodate the noncanonical forms? Did they infect the general auditory system, or were they set apart in a distinct phonetic mode? If the former, then why does everything not sound very much like speech? If the latter, then are we to suppose that the listener shifts back and forth between auditory and phonetic modes depending on whether or not it is the canonical form that is to be perceived?

None of this is to say that natural boundaries or discontinuities do not exist in the auditory system—I believe there is evidence that they do—but rather to argue that they are irrelevant to phonetic perception.

All of the foregoing considerations are simply grist for the Motor Theory mill. For on that theory, the phonetic module uses the speech stimulus as information about the gestures, which are the true and immediate objects of phonetic perception, and so finds the acoustic-phonetic boundaries where the articulatory apparatus happened, for its own very good reasons, to put them. The auditory system is then free to respond to all other acoustic stimuli in a way that does not inappropriately conform their perception to a phonetic mold.

Integrating cues that are acoustically heterogeneous, widely distributed in time, and shared by disparate segments. Having already noted that there are typically many cues for a phonetic distinction, I take note of the well-known fact that these many cues are, more often than not, acoustically heterogeneous. Yet, in the several cases so far investigated, they can, within limits, be traded, one for the other, without any change in the immediate percept. That is, with other cues neutralized, the phonetic distinction can be produced by any one of the acoustically heterogeneous cues, with perceptual results that are not discriminably different. Since these perceptual equivalences presumably exist among all the cues for each such contrast, the number of equivalences must be very great, indeed. But how are these many equivalences to be explained? From an acoustic or auditory-processing standpoint, what do such acoustically diverse, but perceptually equivalent, cues have in common? Or, in the absence of that commonality, how plausible is it to suppose that they might nevertheless have evolved in the auditory system in connection with its nonspeech functions? In that regard, one asks the same questions I raised about the claim concerning the auditory basis of the boundary positions. What general auditory function would have selected for these equivalences? Would they not, in almost every case, be dysfunctional, since they would make very different acoustic events sound the same? And, finally, what is the probability that speakers could so govern their articulatory gestures as to produce for each particular phonetic segment exactly the right combination of perceptually equivalent cues?

The Revised Motor Theory has no difficulty dealing with the foregoing facts about stimulus equivalence. It notes simply that the acoustically heterogeneous cues have in common that they are products of the same phonetically significant gesture. Since it is the gesture that is perceived, the perceptual equivalence necessarily follows.

Also relevant to the argument is the fact that phonetic perception integrates into a coherent phonetic segment a numerous variety of cues that are, because of coarticulation, widely dispersed through the signal and used simultaneously to provide information for other segments in the string, including not only their position, as earlier noted, but also their phonetic identity. The sim-
plest examples of such dispersal were among the very earliest findings of our speech research, as described in the first half of this essay. Since then, the examples have been multiplied to include cases in which the spread of the cues for a single segment is found to be much broader than originally supposed, extending, in some utterances, from one end of a complex syllable to the other; yet, even in these cases, the phonetic system integrates the information appropriately for each of the constituent segments. I have great difficulty imagining what function such integration would serve in a system that is adapted to the perception of nonspeech events. Indeed, I should suppose that it would sometimes distort the representation of events that were discrete and closely sequenced.

Again, the Revised Motor Theory has a ready, if by now expected, explanation: the widely dispersed cues are brought together, as it were, into a single and perceptually coherent segment because they are, again, the common products of the relevant articulatory gesture.

**Integrating acoustic and optical information.** It is by now well known that, as Harry McGurk demonstrated some years ago, observers form phonetic percepts under conditions in which some of the information is acoustic and some optical, provided the optical information is about articulatory gestures. Thus, when observers are presented with acoustic [ba], but see a face saying [de], they will, under many conditions of intensity and clarity of the signal, perceive [da], having taken the consonant from what they saw and the vowel from what they heard. Though the perceptual effect is normally quite compelling, the result is typically experienced as slightly imperfect by comparison with the normal case in which acoustic and optical stimuli are in agreement. But the observers can’t tell what the nature of the imperfection is. That is, they can’t say that it is to be attributed to the fact that they heard one thing but saw another. Left standing, therefore, is the conclusion that the McGurk effect provides strong evidence for the equivalence in phonetic perception of two very different kinds of physical information, acoustic and optical.

For those who believe that speech perception is auditory, the explanation of the McGurk effect must be that the unitary percept is the result of a learned association between hearing a phonetic structure and seeing it produced. As an explanation of the phenomenon, however, such an account seems manifestly improbable, since it requires us to believe, contrary to all experience, that a convincing auditory percept can be elicited by an optical stimulus, or that an auditory percept and a visual percept become indistinguishable as a consequence of frequent association. Indeed, we are required to believe, even more implausibly, that the seemingly auditory percept elicited by the optical stimulus is so strong as to prevail over the normal (and different) auditory response to a suprathreshold acoustic stimulus that is presented concurrently. If there were such drastic perceptual consequences of association in the general case, then the world would sometimes be misrepresented to observers as they gained experience with percepts in different modalities that happened often to be contiguous in time. Fortunately for our relation to the world, there is no reason to suppose that such modality shifts, and the consequent distortions of reality, ever occur. As for the implications of the horizontal account for the McGurk effect specifically, we should expect that the phenomenon would be obtained between the sounds of speech and print, given the vast experience that literate adults have had in associating the one with the other. Yet the effect does not occur with print. It also weighs against the same account that prelinguistic infants have been shown to be sensitive to the correspondence between speech sounds and seen articulatory movements, which is, of course, the basis of the McGurk effect.

On the vertical view, the McGurk phenomenon is exactly what one would expect, since the acoustic and optical stimuli are providing information about the same phonetic gesture, and it is, as I have said so relentlessly, precisely the gesture that is perceived.

**Just how ‘special’ is speech perception?**

The claim that speech perception is special has been criticized most broadly, perhaps, on the ground that it is manifestly unparsimonious and lacking in generality. Unparsimonious, because a “special” mechanism is necessarily an additional mechanism; and lacking in generality, because that which is special is, by definition, not general.

As for parsimony, I have already suggested that the shoe is on the other foot. For the price of denying a distinctly phonetic mode at the level of perception is having to make the still less parsimonious assumption that such a mode begins at a higher cognitive level, or wherever it is that the auditory percepts of the horizontal view are converted to the postperceptual phonetic shapes they must assume if they are to serve as the vehicles of linguistic communication. But generality is another matter. Here, the horizontal view might appear to have the advantage,
since it sees the perception of speech as a wholly unexceptional example of the workings of an auditory modality that deals with speech just as it does with all the other sounds to which the ear is sensitive. In so doing, however, this view sacrifices what is, I think, a more important kind of generality, since it makes speech perception a mere adjunct to language, having a connection to it no less arbitrary than that which characterizes the relation of language to the visually perceived shapes of an alphabet. The vertical view, on the other hand, shows the connection to language to be truly organic, permitting us to see speech perception as special in much the same way that other components of language perception are special. I have already pointed out in this connection that the output of the specialized speech module is a representation that is, by its nature, specifically appropriate for further processing by the syntactic component. Now I would add that the processes of phonetic and syntactic perception have in common that the distinctly linguistic representations they produce are not given directly by the superficial properties of the signal. Consider, in this connection, how a perceiving system might go about deciding whether or not an acoustic signal contains phonetic information. Though there are, to be sure, certain general acoustic characteristics of natural speech, experience with synthetic speech has shown that none of them necessarily signals the presence of phonetic structure. Having already noted that this was one of the theoretically interesting conclusions of the earliest work with the the highly schematized drawings used on the Pattern Playback, I add now that more convincing evidence of the same kind has come from later research that carried the schematization of the synthetic patterns to an extreme by reducing them to three sine waves that merely follow the center frequencies of the first three formants. These bare bones have nevertheless proved sufficient to evoke phonetic percepts, even though they have no common fundamental, no common fate, nor, indeed, any other kind of acoustic commonality that might provide auditory coherence and mark the sinusoids acoustically as speech. What the sinusoids do offer the listener—indeed, all they offer—is information about the trajectories of the formants, which is to say movements of the articulators. If those movements can be seen by the phonetic module as appropriate to linguistically significant gestures, then the module, being properly engaged, integrates them into a coherent phonetic structure; otherwise, not. There are, then, no purely acoustic properties, no acoustic stigmata, on the basis of which the presence of phonetic structure can, under all circumstances, be reliably apprehended. But is it not so with syntax, too? If a perceiving system is to determine whether or not a string of words is a sentence, it surely cannot rely on some list of surface properties; rather it must determine if the string can be parsed—that is, if a grammatical derivation can be found. Thus, the specializations for phonetic and syntactic perception have in common that their products are deeply linguistic, and are arrived at by procedures that are similarly synthetic.

As for specializations that are adapted for functions other than communication, Mattingly and I have claimed for speech perception that, as I earlier hinted, it bears significant resemblances to a number of biologically coherent adaptations. Being specialized for different functions, each of these is necessarily different from every other one, but they nevertheless have certain properties in common. Thus, they all qualify as modules, in Fodor's sense of that term, and therefore share something like the properties he assigns to such devices. I choose not to review those here, but rather to identify, though only briefly, several other common properties that such modules, including the phonetic, seem to have.

To see what some of those properties might be, Mattingly and I have found it useful to distinguish broadly between two classes of modules. One comprises, in the auditory modality, the specializations for pitch, loudness, timbre, and the like. These underlie the common auditory dimensions that, in their various combinations, form the indefinitely numerous percepts by which people identify a correspondingly numerous variety of acoustic events, including, of course, many that are produced by artifacts of one sort or another, and that are, therefore, less than perfectly natural. We have thought it fitting to call this class 'open.' It is appropriate to the all-purpose character of these modules that its representations be commensurate with the relevant dimensions of the physical stimulus. Thus, pitch maps onto frequency, loudness onto amplitude, and timbre onto spectral shape; hence, we have called these representations 'homomorphic.' It is also appropriate to their all-purpose function that these homomorphic representations not be permanently changed as a result of long experience with some acoustic event. Otherwise, acquired skill in using the sound of automobile engines for diagnostic purposes, for example, would render the relevant modules maladapted for every one of the many other events for which they must be used.
Members of the other class—the one that includes speech—are more narrowly specialized for particular acoustic events or stimulus relationships that are, as particular events or relationships, of particular biological importance to the animal. We have, therefore, called this class 'closed'. It includes specializations like sound localization, stereopsis, and echolocation (in the bat) that I mentioned earlier. Unlike the representations of the open class, those produced by the closed modules are incommensurate with the dimensions of the stimulus; we have therefore called them 'heteromorphic'. Thus, the sound-localizing module represents interaural differences of time and intensity, not homomorphically as time or loudness, but heteromorphically as location; the module for stereopsis represents binocular disparities, not homomorphically as double images, but heteromorphically as depth. The echo-locating module of the bat presumably represents echo time, not homomorphically as an echoing (bat) cry, but heteromorphically as distance. In a similar way, the phonetic module represents the continuously changing formants, not homomorphically as smoothly varying timbres, but heteromorphically as a sequence of discrete and categorical phonetic segments.

Unlike the open modules, those of the closed class depend on very particular kinds of environmental stimulation, not only for their development, but for their proper calibration. Moreover, they remain plastic—that is, open to calibration—for some considerable time. Consider, in this connection, how the sound-localizing module must be continuously recalibrated for its response to interaural differences as the distance between the ears increases with the growth of the child's head. The similarly plastic phonetic module is calibrated over a period of years by the particular phonetic environment to which it is exposed. Significantly, the calibration of these modules in no way affects any of the specializations of the open class, even though their representations figure importantly in the final percept, as in the paraphonetic aspects of speech, for example. This is to say that the closed modules must learn by experience, as the phonetic module most surely does, but the learning is of an entirely precognitive sort, requiring only neurological normality and exposure (at the right time) to the particular kinds of stimuli in which the module is exclusively interested.

As implied above, the two classes have their own characteristically different ways of representing the same dimension of the stimulus. Why, then, does the listener not get both representations—heteromorphic and homomorphic—at the same time? Given binocularly disparate stimuli, why does the viewer not see double images in addition to depth? Or, given two syllables [da] and [ga] that are distinguished only by the direction of the third-formant transition, why does the listener not hear, in addition to the discrete consonant and vowel, the continuously changing timbre that the most nearly equivalent nonspeech pattern would produce, and to which the two transitions would presumably make their distinctively different, but equally nonphonetic, contributions.

Mattingly and I have proposed that the competition between the modules is normally resolved in favor of the members of the closed class by virtue of their ability to preempt the information that is ecologically appropriate to computing the heteromorphic percept, and thus, in effect, to remove that information from the flow. As for what is ecologically appropriate, the closed modules have an elasticity that permits them to take a rather broad view. Thus, the module for stereopsis represents depth for binocular disparities considerably greater than would ever be produced by even the most widely separated eyes. The phonetic module will, in its turn, tolerate rather large departures from what is ecologically plausible. Imagine, for example, two synthetic syllables, [da] and [ga], distinguished only by the direction of a third-formant transition that, as I indicated earlier, sounds in isolation like a nonspeech chirp. If the syllables are now divided into two parts—one, the critical transition cue; the other, the remainder of the pattern (the 'base') that, by itself, is ambiguously [da] or [ga]—then, the phonetic system will integrate them into a coherent [da] or [ga] syllable even when the two parts have, by various means, been made to come from perceptibly different sources. (Mattingly has a different interpretation of this particular phenomenon, so I must take full responsibility for the one I offer here.) In what is, perhaps, the most dramatic demonstration of this kind of integration, the isolated transition cue is presented at a location opposite one ear, the ambiguous base at a location opposite the other. Under these ecologically implausible circumstances, the listener nevertheless perceives a perfectly coherent [da] or [ga], and, more to the point, confidently localizes it to the side where only the ambiguous base was presented. (This happens, indeed, even when both the base and the critical transition cue are made of frequency-modulated sinusoids, which is a most severe test, since the differently located sinusoids would, as I pointed out earlier, seem to lack any kind of acoustic co-
In both speech and stereopsis, providing further evidence of ecological implausibility causes the heteromorphic percept (phonetic structure or depth) to weaken as its homomorphic counterpart (nonphonetic chirp or double images) strengthens, until, finally, the closed module fails utterly, and only the homomorphic percept of the open modules is represented. Thus, the information in the stimulus can seemingly be variously divided between the two kinds of representation, and, since either gains at the expense of the other, it is as if there were a kind of conservation of information.

Putting the matter most generally, then, I should say that speech is special, to be sure, but neither more nor less so than many other biologically coherent adaptations, including, of course, language itself.

How do speaking and listening differ from writing and reading?

Among the most obvious, and obviously consequential, facts about language is the immense difference in biological status, hence naturalness, between speech, on the one hand, and writing/reading, on the other. The phonetic units of speech are the vehicles of every language on earth, and are commanded by every neurologically normal human being. On the other hand, many, perhaps most, languages do not even have a written form, and, among those that do, some competent speakers find it all but impossible to master. Having been thus reminded once again that speech is the biologically primary form of the phonological behavior that typifies our species, we readily appreciate that alphabetic writing is not really the primary behavior itself, but only a fair description of it. Since what is being described is species typical, alphabetic writing is a piece of ethological science, in which case a writer/reader is fairly regarded as an accomplished ethologist. It weighs heavily against the horizontal view, therefore, that, as I have already said and will say again below, it cannot comprehend the difference between speaking/listening and writing/reading, for in that respect it is like a theory of bird song that does not distinguish the behavior of the bird from that of the ethologist who describes it.

To see the problem created by the horizontal view, we need first to appreciate, once again, that writing-reading did not evolve as part of the language faculty, so the relevant acts and percepts cannot be specifically linguistic. The important consequence, of course, is that they require to be made so, and, as I have said so many times, that can be done only by some kind of cognitive translation. Now I would emphasize that it is primarily in respect of this requirement that writing and reading differ biologically from speech. Indeed, it is precisely in the need to meet this requirement that writing-reading are intellectual achievements in a way that speech is not. But the horizontal view of speech does not permit us to see that essential difference. Rather, it misleads us into the belief that the primary processes of the two modes of linguistic communication are equally general, hence equally nonphonetic. That being so, we must suppose that the relevant representations are equally in need of a cognitive connection to the language, and so have the same status from a biological point of view. We are, of course, permitted to see the obvious and superficial differences, but, for each one of those, the horizontal view would
Some Assumptions about Speech and How They Changed

seem, paradoxically, to give the advantage to writing-reading, leading us to expect that writing-reading, not speaking-listening, would be the easier and more natural. For, surely, the printed characters offer a much better signal-to-noise ratio than the phonetically relevant parts of the speech sound; the fingers and the hand are vastly more versatile than the tongue; the eye provides far greater possibilities for the transmission of information than the ear; and, for all the vagaries of some spelling systems, the alphabetic characters bear a more nearly transparent relation to the phonological units of the language than the context-variable and elaborately overlapped cues of the acoustic signal.

The vertical view, on the other hand, is appropriately revealing. Given the phonetic module, speakers do not have to know how to spell a word in order to produce it. Indeed, they do not even have to know that it has a spelling. Speakers have only to access the word, however that is done; the module then spells it for them, automatically selecting and coordinating the appropriate gestures. Listeners are in similar case. To perceive the word, they need not puzzle out the complex and peculiarly phonetic relation between signal and the phonological message it conveys; they need only listen, again leaving all the hard work to the phonetic module. Being modular, these processes of production and perception are not available to conscious analysis, so the speakers and listeners cannot be aware of how they do what they do. Though the representations themselves are available to consciousness—indeed, if they were not, use of an alphabetic script would be impossible—they are already phonological in nature, hence appropriate for further linguistic processing, so the reader need not even notice them, as he would have to if, as in the case of alphabetic characters, some arbitrary connection to language had to be formed. Hence, the processes of speech, whether in production or perception, are not calculated to put the speaker’s attention on the phonological units that those processes are specialized to manage.

On the basis of considerations very like those, Isabelle Liberman, Donald Shankweiler, and Ignatius Mattingly saw, more than twenty years ago, that, while awareness of phonological structure is obviously necessary for anyone who would make proper use of an alphabetic script, such awareness would not normally be a consequence of having learned to speak. Isabelle and Donald proceeded, then, to test this hypothesis with preliterate children, finding that such children do, indeed, not know how to break a word into its constituent phonemes. That finding has now been replicated many times. Moreover, researchers at the Laboratories and elsewhere have found that the degree to which would-be readers are phonologically aware may be the best single predictor of their success in learning to read, and that training in phonological awareness has generally happy consequences for the progress in reading of those children who receive it. There is also reason to believe that, other things equal, an important cause of reading disability may be a weakness in the phonetic module. That weakness would make the phonological representations less clear than they would otherwise be, hence that much harder to bring to awareness. Indeed, there is at least a little evidence that reading-disabled children do have, in addition to their problems with phonological awareness, some of the other characteristics that a weak phonetic module might be expected to produce. Thus, by comparison with normals, they seem to be poorer in several kinds of phonologically related performances: short-term-memory for phonological structures, but not for items of a nonlinguistic kind; perception of speech, but not of nonspeech sounds, in noise; naming of objects—that is, retrieving the appropriate phonological structures—even when they know what the objects are and what they do; and, finally, production of tongue-twisters.

The vertical view was not developed to explain the writing-reading process or the ills that so frequently attend it, but rather for all the reasons given in earlier sections of this paper. That it nevertheless offers a plausible account, while the horizontal view does not, is surely to be counted strongly in its favor.

Are there acoustic substitutes for speech?

When Frank Cooper and I set out to build a reading machine for the blind, we accepted that the answer to that question was not just ‘yes’, but ‘yes, of course’. As I see it now, the reason for our blithe confidence was that, being unable to imagine an alternative, we could only think in what I have here described as horizontal terms. We therefore thought it obvious that speech sounds evoked normal auditory percepts, and that these were then named in honor of the various consonants and vowels so they could be used for linguistic purposes.

On the basis of our early experience with the nonspeech sounds of our reading machines, we learned the hard way that things were different from what we had thought. But it was not until
we were well into the research on speech that we began to see just how different and why. Now, drawing on all that research, we would say that the answer to the question about acoustic substitutes is, 'no', or, in the more emphatic modern form, 'no way'. The sounds produced by a reading machine for the blind will serve as well as speech only if they are of a kind that might have been made by the organs of articulation as they figure in gestures of a specifically phonetic sort. If the sounds meet that requirement, then they will engage the specialization for speech, and so qualify as proper vehicles for linguistic structures; otherwise, they will encounter all the difficulties that fatally afflicted the nonspeech sounds we worked with so many years ago.

**FOOTNOTE**

*This essay will appear as the introduction to a collection of papers to be published by MIT Press.*