How Abstract Must a Motor Theory of Speech Perception Be?*

A. M. Liberman+

Your kind invitation suggested that I talk about the motor theory of speech perception. Though I was reluctant to speak on that subject, I nevertheless accepted. The reason for my reluctance was that I then had nothing new to say about a motor theory, and I did not wish to rehearse the old and tired arguments. Therefore, I took the liberty of submitting an abstract that, as you may have noticed, was not exactly responsive to the invitation. Since submitting that abstract, however, my colleagues—Michael Dorman, Lawrence Raphael, Bruno Kepp—and I have collected some data that are at least relevant to a motor theory. Not critical, I hasten to emphasize, only relevant; and illustrative, perhaps, of the ways one might do research on the question. These new data may also be interesting because, as we will see, they suggest that such a theory must be carefully hedged about. At all events, I decided that I could be comfortable talking once again about a motor theory if only because we do have some new data, and because these data will enable me to make explicit a restriction on the theory that has not been much discussed. Hence, I am about to take my second liberty, which is to base my talk on your invitation, after all, and not on the abstract I submitted. Even so, the abstract will be relevant, if incomplete.

Throughout this talk I will, then, be concerned with the question you wanted me to ask: Is the perception of speech linked to its production? In the first part I will say why I think that is a proper question. For that purpose I will consider what we know most generally about speech that motivates us even to wonder about a motor theory. In the second part I will describe the specific experiments by my colleagues and me that illustrate how one might go about getting an answer.

Each part will itself be organized to take into account that we should not adopt a motor theory, or any other similarly special view, until we have reason to reject an auditory theory, which is the most ordinary view. An auditory theory assumes that perceiving the phonetic message is merely an overlaid function, carried out by processes no different from those underlying the perception of music, the noises of a busy highway, or the rustle of leaves in the wind. Only if that most parsimonious theory should prove inadequate are we justified in considering some apparently less parsimonious one. It is appropriate, then, to divide the question. We should ask first: Do we have reason to suppose that

---

*This paper was delivered as a plenary address to the 8th International Congress of Phonetic Sciences, Leeds, England, 21 August 1975.

+Also University of Connecticut, Storrs, and Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]
something other than auditory processes must be invoked? If the answer is yes, we may raise the second question: Can we guess what those other-than-auditory processes might be? I think that division is a particularly important one; hence I have organized what I will say so as to respect it.

Now to the general part of our discussion: What do we know about speech that motivates us to consider the questions I have just raised? Since those questions imply that the perception of speech may require its own distinctive processes, we should therefore look in speech for its most distinctive characteristic. We find that, I think, in the very peculiar nature of the relation between sound and phonetic message. The peculiar relation I speak of is a species of grammar—a speech grammar, if you will—and it is there that we should expect to find the need for special processes.

Because I will view speech as a kind of grammar, I should say a few words about grammatical codes in general so we can see where speech fits. And instead of asking about their form, which is what students of language most commonly do, I will ask rather about their function. For the moment, then, our concern is not with what these grammatical codes are, but with what they do.

To appreciate the function of grammatical codes, we need only consider the nature and shortcomings of agrammatic communication. In an agrammatic mode, which is common among animals and in man's nonlinguistic communication, the relation of message to signal is straightforward. Each message is directly linked to a signal, and every signal differs holistically from all other signals. There is no grammatical structure, only a list of all possible messages and their corresponding signals. Now if all communication were of that kind, we would not have to wonder about distinctive linguistic processes. At the one end, the signals would have to be discriminated and identified, but that is what auditory perception is all about. At the other end, the messages to which those signals are so directly connected would have to be comprehended and stored, but that is the business of processes that lie squarely in the cognitive domain. So if we knew all about auditory perception and all about cognition, we should understand the perception of agrammatic communication. No special theory would be required.

Of course, communication would be very restricted in that agrammatic world. That would be so because agrammatic communication would work as well as it needed to only if there were reasonable agreement in number between the messages people want to send and the holistically different signals they can produce and perceive. But there is the rub. The number of messages our cognitive apparatus can generate and comprehend is uncountably large, while, in contrast, our vocal tracts and ears can cope efficiently with only a small number of signals.

From a biological point of view, the existence of that mismatch is hardly to be wondered at. After all, the mismatched organs—a cognitive apparatus at the one end, a vocal tract and ear at the other—evolved separately, which is to say in connection with wholly different activities. It is tempting, then, to suppose that grammatical codes developed in evolution as a kind of interface between organs that were not made for each other. On that assumption, the function of grammar is to restructure the to-be-communicated messages so as to match the potentialities of the intellect to the limitations of the vocal tract and the ear. To the extent that the codes work well, communication becomes vastly
more efficient and various than it would otherwise be. But that gain is not to be had cheaply, since there is now a peculiar complication in the relation between signal and message, and a need for equally peculiar processes to deal with it. It is, of course, precisely those peculiar processes that would distinguish language from other psychologically interesting activities.

Given, now, that we have reason to look for special grammatical processes, we ask: What characteristics of language suggest the shape that such processes might take? Considering the problem from the standpoint of the perceiver, who is our chief concern today, I will remark the obvious: all the grammatical complications he must cope with are just those that were introduced by the speaker. It is only slightly less obvious that the same is not true for most other forms of perception. In vision, for example, the complications of shape perception are, in a very important sense, external to the perceiver. Therefore, it is reasonable, and not especially novel, to suppose about language that producing and perceiving are only different sides of the same coin. If so, the special linguistic processes that are necessary to perceive language might be expected to have something in common with those that produce it.

So much for grammatical codes in general. What about speech? One might suppose that speech is functionally different from language in that the need for a special grammatical interface has ended with the production of the phonetic message. In that case, the segments of that message would be connected to the sound in a most straightforward, agrammatical way. The perception of speech would then be no different from the perception of other sounds, and its connection to language would be only incidental. But the interfacing does not end with the phonetic message, which is only a stage in the grammatical restructuring that efficiently links meaningful message to acoustic signal. Further, and still quite drastic, changes are necessary, because the requirements of phonetic communication are not well matched to the characteristics of the vocal tract and the ear.

That mismatch has been much discussed (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman, Mattingly, and Turvey, 1973; Liberman, 1974), so I will only recall that it is possible to be quite explicit about it and thus to see clearly what the rest of the grammar—the grammar of speech—must do. For example, it is plain that if the phonetic segments were represented agrammatically in the sound, each phonetic segment by a unit sound, then we could neither speak nor listen as fast as we do. Having in mind that the phonetic message is, in fact, transmitted at rates that reach 25 or more segments per second, at least for short stretches, we see that the speeds we achieve with speech would be impossible if, as in agrammatic communication, the articulators had to change their states in step with the holistically different segments. And even if the articulators could do that, the listener's ear could not possibly resolve the unit sounds so produced: at 25 discrete acoustic segments per second, speech would become an incomprehensible buzz. Moreover, we know about auditory perception of nonspeech sounds that even at rates low enough to avoid the merging of discrete sound segments, the listener nevertheless has difficulty identifying the order in which the segments occurred. Clearly, then, there is a need for a further recoding of the information if the phonetic message is to be efficiently transmitted. And we know that such recoding does in fact occur.

It is no news to this audience that the phonetic message is not transmitted by a discrete set of articulatory gestures, one for every phonetic segment and
each in its proper turn. Rather, the segments are broken down into features; these are assigned to gestures, reorganized into units longer than a segment, and then coarticulated. That arrangement permits us to speak more rapidly than we otherwise could, since it makes for the production of phonetic segments at rates faster than we must change the states of our muscles.

But coarticulation has a perceptual function, too. Consider two of its effects on the acoustic signal and the relation of these effects to the limitations of the ear: first, information about successive segments of the message is transmitted simultaneously, so the number of acoustic segments the ear must resolve is now less than the number of phonetic segments transmitted; and, second, acoustic cues for any particular phonetic segment are different in different contexts, so the order of transmitted segments is marked, not so much by temporal order, which the ear has trouble keeping straight, but by differences in the shape of the acoustic signal. Thus we see how, by grammatical restructuring of the message, the requirements of phonetic communication are matched to the properties of the auditory system. But the fit of grammatical form to perceptual function is achieved at the cost of a complex peculiarity in the grammatical form: there is no correspondence in segmentation between message and signal, and there are odd kinds of context-conditioned variations in the acoustic cues. If listeners nevertheless cope with such peculiarities, as they do quite easily, it must be that they have access to equally peculiar processes.

Having now seen why special processes might be required for speech perception, we can ask what those processes might be. The answer would seem to be the same for speech as it was for the other grammatical codes: the complications the listener must deal with are just those that were introduced by the speaker; thus the key to the code is in the manner of its production. To adopt a motor theory is, in these terms, only to guess that the processes of perception might in some way make use of such a key.

Now I will turn, as I said I would, from general background considerations to some recent experiments bearing on the questions we have raised. First I should say that these experiments are not mine alone but are, as I indicated earlier, the results of a collaborative effort by Michael Dorman, Lawrence Raphael, Bruno Repp, and me (Dorman et al., 1975). I should add, however, my colleagues are not to be blamed for the faults of the particular interpretations I will offer today.

The experiments all have to do with a fact that is in general familiar to you namely, that the "sound of silence" is a phenomenon of speech perception and not merely a poetic image. We all hear that sound whenever we perceive stop consonants, because silence is a manner cue for those segments. Our experiments were designed to deal with that fact, and to answer four questions about it: (1) how large is the effect of the silence cue? (2) is the effect to be accounted for by the properties of the ear? (3) if not, then should we look to the vocal tract? and (4) if so, whose vocal tract?

Now to the first question: How large is the effect of silence as a cue for stops? In the first experiment the stop was in syllable-initial position after a fricative. Figure 1 gives a schematic representation of a syllable consisting of a fricative noise [f], formant transitions of a type appropriate for the stop consonant [p], followed by the steady-state formants that continue the vowel [ε].
There was another pattern like the one in the figure, except that the stop was [k] instead of [p]. In the experiment, we varied the length of the silent interval between the noise patch and the beginning of the formant transitions. Those silent intervals ranged from 0 to 100 msec. The stimuli were real speech, not synthetic; they were presented in random order with instructions to identify each one as [ʃpɛ], [ʃkɛ], or [ʃɛ].

Figure 1: Schematic representation of fricative-stop-vowel syllable, illustrating the kind of pattern used to determine the effect of silence on the perception of syllable-initial stops.

Figure 2 shows the results. We see that at silent intervals of less than 20 msec our listeners reported hearing [ʃɛ], not [ʃpɛ] or [ʃkɛ]. That is, at short silent intervals the stop consonant was, for all practical purposes, not heard. The effect of the silence cue is very large indeed.

The second experiment was intended, in similar fashion, to assess the role of silence as a cue for stops, but now in syllable-final position. Figure 3 shows one of the schematic synthetic patterns that was used. There you see a two-formant disyllable [bɛdɛ]. (The other disyllable [bɛdɛ] is identical except that the second-formant transition at the end of the first syllable is rising instead of falling.) In these cases, the silence we are interested in is the period between the end of the first syllable and the beginning of the second. To assess its importance, we varied its duration from 0 to 120 msec and presented the stimuli for judgment as [bɛbɛdɛ], [bɛgɛdɛ], or [bɛdɛdɛ].

Figure 4 shows the results. We see that when the intersyllable interval is less than about 50 msec, the listeners hear [bɛdɛ], not [bɛbɛdɛ] or [bɛgɛdɛ]. That is, with short intervals of silence, our listeners again do not hear the stop. Plainly, the effect of the silence cue is so large as to be total.
Figure 2: Percent of responses reporting the presence and absence of the stop as a function of the interval between the end of the fricative noise and the beginning of the formant transitions.

Figure 3: Schematic representation of the patterns used to determine the effect of silence on the perception of syllable-final stops.
Figure 4: Percent of responses reporting the presence and absence of the stop as a function of the interval between the end of the first syllable and the beginning of the second.

But the data of those experiments only provide a background against which to ask the next question, which is more relevant to our concerns today: Can we account for the effect of the silence cue in purely auditory terms or must we look to some other-than-auditory processes? Having observed that a silent period is necessary if the stop consonant is to be heard, can we suppose that the importance of silence is owing to some general characteristic of the ear—some characteristic that has no more to do with speech than with any other sounds? Perhaps it is. In the case of [\textipa{ʃpɛ}] and [\textipa{ʃkɛ}], for example, we note that we have conformed to the paradigm for auditory forward masking. It is possible, therefore, that the noise of the fricative masks the transition cues for the stop, rendering them ineffective at an auditory level. On that account, the period of silence between fricative noise and transitions is necessary if the latter is to escape the masking effect of the former. In the case of [\textipa{bɛb ɻɛ}] and [\textipa{bɛg ɻɛ}], we have the paradigm for backward masking. Conceivably, the second syllable "backward masks" the stops at the end of the first syllable, in which case the period of silence would presumably permit the syllable-final stop to evade the masking effect.

When we examine the experimental literature, we find reasons for rejecting an auditory interpretation, especially in the syllable-initial case. Thus, a review of what is known about auditory forward masking reveals that it is typically not a large effect; there appears to be no precedent for the total masking
that we should have to assume in order to account for the disappearance of the stop consonant when the fricative noise is placed in front of it (Elliott, 1971; Leshowitz and Cudahy, 1973). More affirmatively, we find in the literature on speech perception that when the transition cues appear immediately after the fricative noise they do nevertheless have an effect—that is, they are not masked—even though they do not lead to the perception of a stop consonant. Thus, Darwin (1969) found quite incidentally that fricative-vowel syllables were more intelligible with the appropriate consonant-vowel transitions than without them. More recently, Ganong (1975) has found in an adaptation-shift experiment that the boundary between [de] and [be] was moved just as much by adaptation with [se] (that is, fricative noise followed by transitions appropriate for [se] and for [de]) as by adaptation with [de] itself. Given that the [be-de] boundary did not shift nearly so much when the transitions were removed from the adapting [se] stimulus, we may conclude that the transition cues were "getting through" effectively even though they followed close on the fricative noise.

We have aimed to get at the matter more directly. To do that in the case of the syllable-initial stops [fÈ] and [kÈ], we varied the silent interval after the fricative noise, exactly as we had before, but instead of the whole syllables [pÈ] or [kÈ], we presented instead only the isolated second- and third-formant transitions, which are the distinguishing cues. As you know, these isolated transitions do not sound like speech but like chirps. Fortunately for our purposes, listeners can readily learn to identify them differentially as "high" and "low." The fricative-chirp patterns were presented for judgment as "high," "low," and "no chirp."

Figure 5 shows the results. First, it reproduces in the more nearly solid line, the results of the earlier experiment. That solid curve represents the frequency with which the subjects heard [fÈ]—that is, the frequency with which they heard no stop. We see again that at short intervals of silence they failed to hear a stop. Now we see in the bottom line how often they failed to hear the chirp, which was never. That is, for those cases in which the listeners did not hear the stop, they nevertheless heard the essential but isolated transition cues; moreover, they heard them loud and clear.

Figure 6 shows percent correct in the perception of the stops, on the one hand, and the chirps on the other. We see, as we had before, that [pÈ] and [kÈ] are perceived correctly only when there is an interval of silence of 20 to 40 msec, but the corresponding chirps are heard correctly at all silent intervals.

A similar experiment was carried out on the stops in syllable-final position—that is, in [bÈ dÈ] and [bÈ gÈ]. The [b] and [g] transitions were isolated and placed before the syllable [dÈ] with the same intervals of silence that had been used before. Figure 7 shows the results. The more nearly solid line is from the earlier experiment and shows that the listeners heard [bÈ dÈ] at short intervals of silence. That is, at short intervals they did not hear the syllable-final stop. The dashed line at the bottom shows that the subjects never reported not hearing the chirps or, to put it affirmatively, that they always heard the chirps. Figure 8 shows how often the listeners perceived the stops and the chirps correctly. The lower, more nearly solid line reproduces the earlier result and shows that at short intervals of silence the listeners did not correctly perceive the syllable-final stops [b] and [g]. The upper, dashed line shows that at these same short intervals of silence they did perceive the chirps—that is, the transition cues—correctly, though there is perhaps a small effect of masking on the accuracy of that perception.
Figure 5: Percent of responses reporting the absence of the syllable-initial stop and the chirp (isolated stop cues) as a function of the interval between the end of the fricative noise and the beginning of the formant transitions.

Figure 6: Percent of responses correctly identifying the syllable-initial stops and the corresponding chirps (isolated stop cues) as a function of the interval between the end of the fricative noise and the beginning of the transitions.
Both the experiments I just described agree with the already available evidence to which I alluded and support the conclusion that there is little or no forward or backward auditory masking; more generally, they indicate that the essential stop-consonant cues are fully effective as purely auditory events, even at the very shortest intervals of silence. We should suppose, then, that the effect of the silence cue is at some other-than-auditory level, and is to be accounted for by some other-than-auditory process.

Where, now, do we look for a proper account? You students of phonetics will have been wondering to yourselves all this time why I don't take into account that a speaker cannot produce a stop without closing his vocal tract, and that the resulting silence provides information, not time to evade masking. I do want to take account of that now and to make explicit that the essential information provided to the listener is information about what the speaker's vocal tract is doing. But I would also emphasize that in this case the listener cannot perceive what the vocal tract cannot do. To see how interesting that is, contrast it with what happens in visual perception. Imagine that I show you a picture of a horse standing on its tail and ask you to tell me what you see. You would say that you see a horse standing on its tail, and you might then add...
that I had contrived the picture, since you know that horses cannot stand on their tails. But you would nevertheless have seen the horse standing on its tail. Consider how different were the results of our experiments. Had you been one of our listeners, what might you have heard and what might you have thought? Would you have heard the stop consonant and then supposed that I had synthesized the sound because you know that a vocal tract could not have responded that fast? No. You would not have heard the stop, period. Which brings us, then, to the third question relevant to our experiments: Are we here dealing with cases in which the constraint on perception is not so much by the ear as by the vocal tract?

The relevant experiments are like the earlier ones on the perception of stops in syllable-final position, except that now we use all three stops at the end of the first syllable, followed in all possible combinations by all three stops at the beginning of the next. That is, we now have [bab], [bud], and [beg], followed in all pairings by [ba], [da], and [ga] and, as before, we vary the intersyllable interval and ask our listeners to identify the stop at the end of the first syllable.

Figure 9 shows the results. I invite your attention to the fact that the nine curves form three clusters. Let us look first at the cluster that rises most slowly, the one described by short dashes connecting solid circles. Those
curves have in common that they represent data obtained with the geminates: [bab ba], [bad da], and [bag ga]. The production of such geminates requires a long interval of silence between the syllables; it is of interest, surely, that their perception requires a long interval too. In any event, those perceptual results group themselves most obviously according to an articulatory criterion, not an acoustic one. The same appears to be true of the other two clusters. Thus, the middle cluster—the three curves formed by the solid line connecting stars—represents those cases in which the place-of-closure for the syllable-initial stop—that is, those cases in which the place-of-closure moved from front to back: [bab da], [bab ga], and [bad ga]. The cluster of curves that rise most rapidly—those formed by dashed lines connecting open circles—have in common that the place of closure moves, as in a hinge, from back to front: [bag ba], [bag da], and [bad ba]. I am not prepared to say what vocal tracts do differently regarding shifts from front to back and back to front. And I will have more faith in the generality of our results—that is, more faith that there is no simple auditory basis for the clustering—if, with a variety of vowels and hence a variety of consonant cues, we nevertheless get the same result. But, taking the data as we so far find them, I believe they do suggest that perception was here constrained by properties that belong not so much to ears as to vocal tracts.

Figure 9: Percent of responses correctly identifying the syllable-final stops for all following syllable-initial stops, as a function of the interval between the end of the first syllable and the beginning of the second.
So we come now to the last question: Whose vocal tract? I should suppose that it could hardly be that of the listener or the speaker or, indeed, of any particular person, but that it must rather be an abstract conception. In this regard our situation may be similar, in its own small way, to the one in which the philosopher, Bishop Berkeley, found himself. Believing as he did that things exist only if they are perceived, he had to account for the tree that is for some time not in the mind of the gardener or of any other mortal being. He asserted that the tree did exist nevertheless because it was at all times perceived by the infinite mind of God (Berkeley, [1713] 1954). Lacking Berkeley’s ecclesiastical credentials, I hesitate to make the vocal tract we are concerned with abstract in the same way. Moreover, I am not a philosopher but an experimentalist, so I must test an abstract assumption by concrete means.

We can see by experiment how abstract the theoretically relevant vocal tract must be by exploiting two facts about the ecology of speech. The first, which we have already noted, is that a speaker cannot articulate a disyllable like [bob da] without closing his vocal tract so as to produce a silent interval between the syllables; the second is that there need be no such silent interval if the syllables are articulated by two different speakers, the first syllable by one speaker, the second by the other. Taking advantage of those facts, my colleagues and I first replicated one of our earlier experiments and found, as we had before, that when the two syllables are produced by a single speaker, the syllable-final stop can be heard only when there is a silent interval of some length between the syllables. Figure 10 shows the result. There, in the solid line connecting the solid circles, we see that, as in the earlier experiment, the listeners correctly perceived a syllable-final stop only when there was a decent interval following it. But that, I would emphasize, is what happened when the two syllables were produced by the same voice, in this case a male.

What happens when the first syllable is produced by that male voice and the second syllable by a different-sounding female voice? The answer given by eight of our first ten subjects is shown in the curve that runs almost straight across the top of the graph, the dashed line connecting open circles. That curve says that in the different-voice condition the syllable-final stop was perceived almost perfectly at all values of intersyllable intervals including even the zero value. The remaining two subjects gave results shown by the broken line connecting the open squares. As you see, they performed in the different-voice condition much the way they (and everyone else) did in the same-voice condition. It is of interest that one of those two remarked spontaneously after the experiment that she thought the voices were not different. I should add that further research on this problem has yielded results like those produced by the eight subjects; they reinforce the conclusion that in the different-voice condition the syllable-final stop can be heard at all intersyllable intervals. Thus, when the two syllables are spoken by a single voice, the syllable-final stops cannot be perceived at the short intersyllable intervals that the single voice cannot produce; but when the syllables are spoken by two different voices, production and perception are possible at all intervals between the syllables, even at no interval at all.

To develop the implications of this last finding will require many more experiments, some of which we do not yet know how to do. But one implication is fairly clear, and it happens to be one that is most relevant to our concerns in this paper. It is, as I have already said, that a motor theory must be quite
Figure 10: Percent of responses correctly identifying the syllable-final stops when both syllables are produced by a male speaker and when one is produced by a male, the other by a female.

abstract. But if our results require that the theory be so qualified, they do not yet suggest that it be wholly abandoned. To take account of these results I should suppose that the articulatory model the listener has access to acts as if it knew about the capabilities of vocal tracts in general. Computations carried out in terms of that model might yield something like the results we have observed.

I should summarize.

By examining the function of grammatical codes, we see why specialized linguistic processes might be necessary and why perception and production might be linked. In the case of speech we can be more explicit about the need for grammatical recoding and especially about the fit of grammatical form to perceptual function. We see more clearly, then, that specialized perceptual processes might be required if the listener is to cope with the code, and we see just as clearly that the key to the code is in the manner of its production. Because we suspect that such a key may be a part of the specialized processes, we look with favor on a motor theory of speech perception.

Several recent experiments illustrate that data relevant to such a theory can be obtained. In these experiments it was found that (1) silence is an important manner cue for the perception of stop consonants; (2) this cue is not constrained primarily by the properties of the auditory system; (3) the constraint
appears rather to be related to what the vocal tract can and cannot do; and (4) it is not any particular vocal tract that imposes the constraint but some conception of vocal tracts in general. I should conclude that a motor theory is still a reasonable way to make sense of some of the phenomena of speech perception, but only if we assume that the implied reference to production is highly abstract.

REFERENCES


