IN SPEECH PERCEPTION, TIME IS NOT WHAT IT SEEMS

Alvin M. Liberman

Haskins Laboratories

New Haven, Connecticut

Ann. NY Academy of Science, 682, 264-271 (1993).

The subject of this symposium and the titles of some of its papers imply a belief that general principles of temporal processing can enlighten us about language behavior and the ills that attend it, whether in speech or in writing-reading. To determine just how well founded that belief is, we must I think, resolve two issues. One concerns the relation between the two kinds of language behavior we are trying to understand. The other looks in a different direction, at the relation of speech, the more basic of these behaviors, to the nonlinguistic modalities where the roots of that understanding are presumed to lie.

The salient fact about the relation of speech to writing-reading is the vast difference in the underlying biology. Speech is a product of biological evolution, having emerged with us as the most important of our species-specific characteristics. Writing systems, on the other hand, are recently developed artifacts, part discovery, part invention. In the case of an alphabet, the momentous discovery was that human beings had, for untold thousands of years, been speaking phonologically. Quite without knowing it, they had all been using a marvelously generative scheme for producing an indefinitely large number of meaningful words by variously combining and permuting a small number of meaningless segments. Once that was understood, it remained only to invent the idea that if each of the meaningless segments were to be represented, however arbitrarily, by a distinctive optical shape, then all could write and read, provided only that they knew the language and could manage to become consciously aware of the internal phonological structure of its words.

Seen this way, an alphabetic transcription is an accurate account of species-specific behavior, hence an early achievement of ethological science. Accordingly, one might characterize the relation of speech to writing-reading by a simple equation: the speaker-listener is to the writer-reader as an ordinary human being is to an ethologist.

What, then, is biologically distinct about the species-specific behavior that an alphabetic transcription is an ethological account of? In short, what evolved? Not why, or when, or by what progression from earlier-existing states. Only what.

I mean, in due course, to suggest that what evolved were processes that are specific to speech, hence part of the larger specialization for language and not likely, therefore, to be properly understood by reference to putatively comparable processes in nonlinguistic modalities. Touching more particularly on the point of this symposium, I will say that the evolution of these speech-specific processes was itself guided largely by requirements of temporal processing, which is our subject, then quickly add that these requirements were special, having been imposed by the special nature of phonologic communication.

But first I must take account of the opposite view, which is that the processes we are here concerned with are not specific to speech, that language simply appropriated, for phonological purposes, motor and auditory mechanisms of a very general sort. (For representative examples of this view, see Crowder and Morton, 1969; Fujisaki and Kawashima, 1970; Oden and Massaro, 1978; Samuel, 1977; Miller, 1977; Stevens, 1975; Cutting and Rosner, 1974; Hillenbrand, 1984; Diehl and Kluender, 1989; Lindblom, 1991). Borrowing a word that Fodor (1983) used to characterize a much

broader version of this view, I will call it 'horizontal', in contrast to my view, which is appropriately called 'vertical'. By any name, however, the horizontal view is the more conventional and also the more congenial to the purpose of this symposium. It rests on four assumptions that compare neatly with four facts about writing-reading.

The first assumption is that the ultimate constituents of language are sounds. Obvious though this may seem, it is, nevertheless, only an assumption, and, as I will argue, quite wrong. However, its counterpart in writing-reading is a hard fact: the constituents of an alphabet are optical shapes and nothing else.

The second assumption is that the consituent sounds are produced by articulatory maneuvers and motor control processes that are not specific to speech, but are, rather, of some quite general sort. As for its counterpart in writing, we know that the movements the writer makes cannot be specific biological adaptations to language, if only because writing was not part of linguistic evolution.

The third assumption is companion to the second. Just as the production of speech sounds is managed by processes of a general nonlinguistic sort, so, too, is their perception. According to the horizontalists, perception of speech is no different from perception of other sounds. All are governed by the same general processes of hearing, processes that evoke in a common perceptual register a common set of auditory primitives: pitch, loudnesss, timbre, and the like. The representations evoked by a stop consonant and a squeaking door are made of the same perceptual stuff; they differ only in the relative values assigned to the primitives they share. This assumption that perception of speech sounds is generally auditory corresponds to the fact that perception of alphabetic characters is generally visual.

We come, then, to the fourth assumption, which is made necessary by the second and third. For if the motor and perceptual representations are not themselves distinctly linguistic, they must be made so, and that can be done only by a cognitive translation. Accordingly, horizontalists say explicitly that after experiencing the purely auditory percept that was evoked by the acoustic signal, the listener connects it to language by giving it a phonologic name or otherwise associating it with a phonologic unit. Thus, perceiving speech is, to the horizontalist, a matter of experiencing something auditory and calling it something linguistic. For speech perception, this proves to be a very troubling, if necessary, assumption, but for reading it is an indisputable fact; the purely visual percepts evoked by the letters of the alphabet do require to be translated into units of the language. Indeed, as I mean to emphasize later, it is just the need for this translation that distinguishes reading from speech perception.

Thus, the first shortcoming of the horizontal view is that everything it assumes about speech is a fact about writing-reading, so the horizontal view cannot rationalize the profound biological difference between the two kinds of behavior that are the objects of our concern. But if a theory of speech cannot cope with a fact about language as basic and obvious as the difference in naturalness between its spoken and written forms, then there must be something profoundly wrong with it.

Other shortcomings go deeper. Perhaps the deepest has to do with the most basic requirement that a communication system must meet, which is simply that sender and receiver be bound by a common understanding about what counts: what counts for the sender must count for the receiver. Though this requirement does not commonly figure in the evalution of theories of language, Ignatius Mattingly and I have thought it important enough to deserve a name, so we have called it the 'requirement for parity', and we have challenged theories to explain how it was established as language developed in the history of our species and as it develops anew in each child (Mattingly and Liberman, 1990; Liberman and Mattingly, 1989; Liberman, in press).

To see the point most clearly, consider parity in the case of writing-reading. If someone draws a 'B' and a squiggle, all who are literate in the Roman alphabet understand that the one has linguistic significance and the other does not. Moreover, all know what the particular linguistic significance of the 'B' is. Thus, parity exists, so all can use this signal for communication. As for the origin of parity in this case, it is to be found in an agreement, arrived at by those who presided over the development of the Roman alphabet, that invested a select set of optical shapes with linguistic significance by arbitrarily assigning each one to a phonological segment.

But, surely, we cannot make a similar statement about speech. People did not simply agree that 'da' would count but a snapping of the fingers would not. Neither was it an agreement that determined what 'da' would count for. Yet, as we have seen, the horizontal view allows no alternative -- since it assumes that the motor and perceptual representations of speech are connected to language, and to each other, only because speaker and listener choose to call them by the same phonologic names or somehow connect them to the same phonologic units. Indeed, the harder one looks at the parity question from the horizontal point of view, the more one is forced to the unacceptable conclusion that phonology must have been an invention, a new and better mode of communication devised by human beings who were smart enough to appreciate the generative advantages of the phonological principle and creative enough to have seen how to exploit it, given the resources of the vocal tract and the ear.

Having said that speech as seen on the horizontal view would not plausibly meet the parity requirement, which is imposed on all forms of communication, I turn now to requirements that are specific to phonology. There are two. The first, which must be met if phonology is to serve its generative function, is that the segments of the phonological structure be commutable -- that is, that they be discrete, invariant, and categorical. The second, which is only slightly less obvious, has to do with rate and, accordingly, with the subject of this symposium: temporal processing. The point is that if all utterances are to be formed by variously stringing together an exiguous set of segmemts. then, inevitably, the strings must run to considerable lengths. Moreover, these segments must be organized into words, the words into phrases, and the phrases into sentences. There is, then, a need for rapid production and perception of the segments.

But discrete, invariant, and categorical sounds would require correspondingly discrete, invariant, and categorical gestures, so, on the horizontal view, communication could be managed only at unacceptably slow rates. To know how slow, one has only to consider writing, even cursive writing, where it is similarly necessary that the optical shapes be discrete, invariant and categorical. So, if the ultimate constituents of speech were sounds, speaking would be like writing: it would not be possibale to say 'bag', but only 'b' 'a' 'g'. And to say 'b' 'a' 'g' is not to speak but to spell.

I should add that the problem is fundamentally insoluble on the horizontal strategy, for if Nature had tried to avoid the limitations on rate by endowing her human creatures with acoustic devices specifically adapted to producing a rapid-fire string of sounds, she would have defeated the ear. The point is that, as I speak to you now, I am producing phonological segmetns at a rate that averages about 10 per second and, for short stretches, reaches 20 or more. If each of those were a discrete sound, rates that high would seriously strain the temporal resolving power of the ear and its ability to place the segments in their proper temporal order. Thus, phonological communication would be impossible.

The foregoing arguments of plausibility are about just those shortcomings of the horizontal view that are most relevant to the purpose of this symposium. They reduce to the consequences of the most general assumption of this view, which is that there is, at the level of action and perception, no such thing as a distinctly and specifically linguistic mode. Thus, the question, "What evolved?", that I earlier asked gets from the horizontal view the simple answer, "No more in speech than in writing-reading". Since that answer flies in the face of the most obvious fact about the relation between these two behaviors, I find the view that supplies it a poor guide to the understanding we seek.

I turn, then, to the view that I think more likely to be helpful, the view that I earlier referred to as 'vertical'. (For general accounts, see Liberman and Mattingly, 1985; Mattingly and Liberman, 1988; Mattingly and Liberman, 1990; Liberman and Mattingly, 1989). The first assumption of this view is that nature managed the rate problem by defining the constituents of language, not as sounds, but as gestures. Thus, the constituent we write as 'b' is a closing of the vocal tract at the lips; 'm' is a closing at the lips and an opening at the velum; and so on. Putting aside admittedly difficult questions about exactly how these gestures are to be characterized, we see nonetheless clearly the great advantage of the gestural strategy, which is that it permits coarticulation. Given phonologic segments that are instantiated as more or less abstract motor units, and given successive segments that are realized at the periphery by independent articulators such as the lips for 'b' and the tongue for 'a', the speaker says , not 'b' 'a', but 'ba'. At all events, coarticulation is characteristic of all languages, and it is, without question, the necessary condition for the rapid rates of phonologic communication that are, in fact, achieved.

The second assumption of the vertical view is that the articulatory movements and their controls were not lying conveniently to hand, just waiting to be used by language. They are, rather, the products of

evolution. Consider, for example, that the movements we make when we speak are a distinct set, different from those we make with the same organs when we chew, swallow, move food around in the mouth, lick our lips, or worry a sore tooth with the tip of the tongue. Phonologic gestures serve a linguistic function and no other. Presumably, they were selected in evolution largely because of the ease with which they lent themselves to being coarticulated. As for their controls, they must be specialized, too. Is there another motor system in which the kinematics are managed by, and perfectly preserve information about, strings of temporally ordered motor units that are discrete, invariant, and categorical?

Of course, the organs that are used in speech are also controlled by motor systems that have nothing to do with speech, and these must certainly affect speech as well. Indeed, they will often take precedence over speech, as in breathing, coughing, or gagging. And, surely, the phonological specialization has much in common with other motor systems -- for example, the need to control degrees of freedom -- but, as I said earlier, it also has to manage the rapid production of strings of discrete, invariant, and categorical motor units, and that task seems peculiar to phonological communication. So, somewhere upstream from the final common paths, there must be a specialization for that special phonological task. Just where that 'somewhere' is can only be determined empirically.

The third assumption of the vertical view takes account of the perceptual consequences of coarticulation. These are happy consequences from the standpoint of rate of communication, since coarticulation folds into a single piece of sound information about several phonologic segments, and so, by achieving parallel transmisstion, considerably relaxes the constraint on rate imposed by the temporal resolving power of the ear. Therefore, listeners can perceive phonologic structures as fast as the coarticulating speaker can produce them. But this comes at the cost of a considerable and specifically linguistic complication in the relation between acoustic signal and phonologic message. Thus, as all speech researchers know, the acoustic signal for each particular phonologic segment is different, often grossly, depending on the segments with which it is coarticulated. For some segments, the signal also varies as a function of rate of articulation and condition of linguistic stress. What remains constant is only some representation of the articulatory gesture. One is led, then, to suppose that the phonologic component of the language specialization has two complementary processes: one for computing the articulatory movements from the more abstract specification of the gestures, as in the second assumption above, the other for automatically analyzing the acoustic signal in such as way as to recover the coarticulated gestures that caused it.

As in speech production, so, too, in speech perception, we recognize that the most peripheral structures and functions are shared with processes that are not linguistic, and it must, again, be an empirical matter to determine where the division into phonologic and nonphonologic occurs. What is clear, I think, is that it does not occur at a point where a preliminary auditory representation is cognitively translated into something phonologic, for there is no such preliminary auditory representation.

There is no need, then, on the vertical view, as there was on its horizontal opposite, for a fourth assumption about a cognitive translation into language, because the motor and perceptual representations of speech are, by their nature, already linguistic, hence perfectly appropriate for further processing by the other components of the language system. Indeed, it is exactly this that is the primary biological difference between speech and writing-reading.

As for parity, we see, on the vertical view, that it is the very essence of the system, for what evolved was nothing less than a communicative modality, including the specifically phonological gestures that are integral parts of it. Accordingly, these gestures form a natural class, so their representations are set apart, biologically, from all others by their membership in that class, not by having phonologic names assigned to them. Speaker and listener can communicate in a perfectly natural way, as writer and reader cannot, because the speaker sends and the listener receives exactly the same specifically phonologic gestures; there is no need to connect a generally motor representation that underlies production to a generally auditory representation that underlies perception by means of an agreement that arbitrarily makes them comparable.

The vertical view also allows us to understand exactly what it is that the would-be reader must learn that experience with speech will not have taught him. Consider that a speaker does not have to know how to spell a word in order to produce it. Indeed, he does not even hav e to know that it has a spelling. He has simply to think of the word; the phonological specialization spells it for him, automatically selecting and coordinating the relevant gestures. The listener is in similar case, for to perceive a word, he need not puzzle out the complex relation between the acoustic signal and the string of segments it conveys, or even be in any way aware of how very complex that relation is. Rather, he need only listen, relying on the the automatic processes of phonology to parse the signal into its segments and arrange them in the proper temporal order. Speaker and listener lack awareness of these processes of production and perception, because the governing mechanisms are modular, hence insulated from consciousness. Of course, these modular processes do make their representations available to consciousness; indeed, if they did not, then alphabetic writing and reading would be impossible. But, as we've seen, these representations, being immediately phonologic, do not require the translation that would put them at the focus of attention. Taking all this into account, we understand why experience with speech, no matter how extensive, is not likely to produce the awareness of phonological structure that the would-be reader needs if he is to understand the phonologic principle, and so be able to connect the alphabetic transcription to the language it indexes (Liberman, 1973).

Like writing-reading, speech must, of course, be learned, but, unlike writing-reading, it need not be taught. The language module does, of course, depend on the phonologic environment for its proper development and calibration, and in this respect is no different, in principle, from such modules as those that are, for example, responsible for stereoscopic vision and sound localization (Liberman, 1992). But this kind of learning is precognitive, which is to say that it requires little more than the appropriate stimulation from the environment. Given the phonologic

environment in which it finds itself, the language module will adapt without cognitive effort on the part of the child. Reading an alphabetic script, on the other hand, requires a cognitive, conscious understanding of the facts of phonological structure. Thus, we see how reading and writing are intellectual achievements in a way that the development of speech is not.

Now, at last, I redeem the promise of my title by describing the results of two kinds of experiments that show how very special are the perceptual consequences of the flow of time in the speech signal. The first has to do with the fact that speech perception requires the listener to respond to resonances that move rapidly up or down in center frequency. In the case of stop consonants, for example, the critically important resonances complete excursions as large as 500 Hz in about 50 msec. On the horizontal view, one might suppose that tracking these rapid changes is an auditory process that produces a correspondingly auditory representation, and that this representation can be used as a base for obtaining psychophysical data relevant to the processes of speech perception. That matters are not so straightforward is shown by the results of experiments on a phenomenon, called 'duplex perception', in which a speech pattern is divided into two discordant or discontinuous parts, with the result that one of the parts is simultaneously perceived as two distinct events, one phonetic, the other not (Rand, 1974; Mattingly and Liberman, 1990; Bentin and Mann, 1990; Liberman and Mattingly, 1989; Whalen and Liberman, 1987). For example, a rapid movement up or down in center frequency that can be made responsible for the perceived difference between 'da' and 'ga' will be perceived as a nonspeech chirp (which is what it sounds like in isolation) and, at the same time, be responsible for the phonologic distinction. A duplex percept of this kind provides a unique opportunity to measure the listener's ability to discriminate tokens of the same rapidly changing resonance (in the same acoustic context) when, on the one side of the duplex percept, the resonance is producing the nonspeech chirp, and when, on the other, it is, at the same time, evoking the phonologic segment. The finding is that the discrimination functions obtained with these simultaneously available percepts are very different, both in shape and in level. Apparently, responding to rapidly changing resonances when they cue a phonologic segment does not depend on the same processes that underlie perception of these same resonances when, failing to engage the language module, they are perceived as acoustic events (Mann and Liberman, 1983; Hence, generalizing psychophysical results from the auditory to the phonologic modalities is, at best, a chancy thing.

The second relevant observation is more directly about the processing of temporal order. It pertains to the fact that, as Gunnar Fant pointed out many years ago, there is no direct correspondence in segmentat;ion between the acoustic signal and the perceived phonetic message (Fant, 1962). Since then, dozens of experiments have justified two important generalizations about temporal order in speech: (1) the articulation of a phonological segment typically has acoustic consequences that cover a wide span of the signal and overlap quite thoroughly with the acoustic consequences of the articulation of other segments in the string; and (2) the speech-perceiving system is sensitive to all these consequences, no matter how widely distributed, overlapped, or acoustically heterogeneous. In the contrast between the words 'slit' and 'split, for example, the

Bentin, S. & V. A. Mann. (1990). Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes. Journal of the Acoustical Society of America, 88(1): 64-74.

Whalen, D. H., & A. M. Liberman. 1987. Speech perception takes precedence over nonspeech perception. Science, Vol. 23: 169-171.

Mann, V. A. & Liberman, A. M. 1983. Some differences between phonetic and auditory modes of perception. Cognition, 14: 211-235.

Fant, C. G. M. 1962. Descriptive analysis of the acoustic aspects of speech. Logos, 5: 3-17.

Repp, B. H. 1985. Perceptual coherence of speech: Stability of silence-cued stop consonants. Journal of Experimental Psychology: Human Perception and Performance, 11, No. 6: 799-813.