Taking speech in the narrow sense—as the production and perception of the sounds that convey phonetic structure—one finds two very different views of its relation to language. The more conventional holds that speech is merely a vehicle, bearing no organic relationship to the linguistic baggage it carries. On that view, speech is produced and perceived by processes that are not specialized for language but rather serve horizontally the broadest possible variety of behaviors, linguistic and non-linguistic alike. The outcomes of those primary processes are then presumably sent on to language proper, a separate domain where they find the mental machinery capable of the heavy lifting required by phonology, morphology and syntax. Preferring a name that reflects the nature of a theory rather than its currency, we will call the conventional view ‘horizontal’ (as in Ref. 1). The other, less conventional view is that the biological roots of language run deep, penetrating even to the level of speech and to the primary motor and perceptual processes that are engaged there. Seen from that perspective, speech is a constituent of a vertically organized system, specialized from top to bottom for linguistic communication. Such a view may be appropriately called ‘vertical’.

To evaluate these strongly contrasting views, we propose here to determine which of the two provides the more coherent and plausible account when challenged by several simple and seemingly obvious biological considerations. As we will try to show, those considerations provide decisive tests of any theory of speech, yet they do not normally figure in the calculations of the theorists, nor have they been permitted to influence the implicit assumptions that guide (or misguide) almost all applied language research, including that which is aimed at determining how to convert fluency in speech to fluency in the use of its alphabetic transcription. Indeed, bringing those considerations to notice is one purpose of this paper. But first, we will give a brief description of the theories.

The horizontal viewpoint

The first assumption of the horizontal view is that the elements of speech are sounds. That is not merely to say the obvious, which is that speech exploits an acoustic medium, but rather to identify sounds as the primitives that are exchanged when linguistic communication occurs. The invariant acoustic patterns that might justify such an assumption have, indeed, been claimed for a variety of phonetic segments, including, for example, stops2, nasals3, and the voicing contrast of fricatives4. However, most students of language, and virtually all of those interested in its many applications, simply take for granted that sounds are the elements of speech. It is not usual, therefore, to find that premise developed or defended explicitly as a basis of the horizontal theory, though it is that. And if few bother to dismiss imaginable alternatives, it is, apparently, because alternatives do not spring readily to mind.

As for what should be the second assumption, the one that would concern the articulatory gestures that produce the elemental sounds, there is, again, little in the way of explicit

†Alvin M. Liberman and Doug H. Whalen

There are two widely divergent theories about the relation of speech to language. The more conventional view holds that the elements of speech are sounds that rely for their production and perception on two wholly separate processes, neither of which is distinctly linguistic. Accordingly, the primary motor and perceptual representations are inappropriate for linguistic purposes until a cognitive process of some sort has connected them to language and to each other. The less conventional theory takes the speech elements to be articulatory gestures that are the primary objects of both production and perception. Those gestures form a natural class that serves a linguistic function and no other. Therefore, their representations are immediately linguistic, requiring no cognitive intervention to make them appropriate for use by the other components of the language system. The unconventional view provides the more plausible answers to three important questions: (1) How was the necessary parity between speaker and listener established in evolution, and how maintained? (2) How does speech meet the special requirements that underlie its ability, unique among natural communication systems, to encode an indefinitely large number of meanings? (3) What biological properties of speech make it easier than the reading and writing of its alphabetic transcription?
concepual development or consideration of reasonable alternates. Consequently absent, at all events, is the idea – important for the purposes of this review – that those gestures might be specialized for language. One is left to infer, then, that they are not, that their biology is no more distinctively linguistic than that which underlies the motor processes that caused the toss or twiddle the thumbs. Perhaps the most explicit statement of this view is that of the linguist Bjorn Lindblom who has written:

Spoken adaptively tone phonetic gestures to the various needs of speaking situations, and languages make their selection of phonetic gestures inventories under the strong influence of motor and perceptual constraints that are language independent and in no way special to speech. (Ref. 5, p. 7)

In further dismissing the possibility that speech production is a specialized adaptation, Lindblom asks why, if this is so, do "inventories of vowels and consonants nevertheless show evidence of being optimized with respect to motoric and perceptual limitations that must be regarded as biologically general and not at all special to speaking and listening." (p. 7)

Unlike Lindblom, most horizontalists give speech production short shrift, assigning to it a secondary role in which its chief function is simply to articulate those sounds that fit comfortably the language-independent properties of the ear6–9. Consistent with that view is the particularly close attention to the perceptual limitations that must be regarded as biologically general and not at all special to speaking and listening. (Ref. 5, p. 7)

The first assumption of the vertical view is that the phonetic elements of speech, the true primitives that underlie linguistic communication, are not sounds but rather the articulatory gestures that generate those sounds21–25. Those units, which serve as the primary representations of speaker and listener alike, are presumed to be something like the coordinate structures that Turvey and others have assumed26–29. As expressed in the peripheral structures, the gestures are changes in the cavities of the vocal tract – openings or closings, widening or narrowing, lengthenings or shortenings. They are supposed on the vertical view that the processes underlying the gestures evolved with language, specifically in the service of a linguistic function; hence, the gestures are phonetic af initio, requiring no cognitive translation to make them so30–33. Coincident with the evolution of the phonetic mechanisms was what Studdert-Kennedy has called the 'particulation' of the vocal tract30 – that is, the division of vocal-tract action into independently controllable components (for example, the categorical perception of phonetic elements) in the service of phonetic privileges and so made usable by the special processes of language. That translation is variously accomplished by matching the primary auditory percept to phonetic templates34–37, by associating it with the abstract 'distinctive features' so familiar to linguists38, or, as we are left to infer by some treatments, simply by giving it a phonetic name39–41. For the purposes of this review, these are distinctions without a difference; in assuming the need for translation to linguistic status, however done, the horizontalist accepts that language is a specialization, while at the same time denying that speech is part of it.

The vertical alternative

The first assumption of the vertical view is that the phonetic elements of speech, the true primitives that underlie linguistic communication, are not sounds but rather the articulatory gestures that generate those sounds21–25. Those units, which serve as the primary representations of speaker and listener alike, are presumed to be something like the coordinate structures that Turvey and others have assumed26–29. As expressed in the peripheral structures, the gestures are changes in the cavities of the vocal tract – openings or closings, widening or narrowings, lengthenings or shortenings. They are supposed on the vertical view that the processes underlying the gestures evolved with language, specifically in the service of a linguistic function; hence, the gestures are phonetic af initio, requiring no cognitive translation to make them so30–33. Coincident with the evolution of the phonetic mechanisms was what Studdert-Kennedy has called the 'particulation' of the vocal tract30 – that is, the division of vocal-tract action into independently controllable components (for example, the categorical perception of phonetic elements) in the service of phonetic privileges and so made usable by the special processes of language. That translation is variously accomplished by matching the primary auditory percept to phonetic templates34–37, by associating it with the abstract 'distinctive features' so familiar to linguists38, or, as we are left to infer by some treatments, simply by giving it a phonetic name39–41. For the purposes of this review, these are distinctions without a difference; in assuming the need for translation to linguistic status, however done, the horizontalist accepts that language is a specialization, while at the same time denying that speech is part of it.

The vertical alternative

The first assumption of the vertical view is that the phonetic elements of speech, the true primitives that underlie linguistic communication, are not sounds but rather the articulatory gestures that generate those sounds21–25. Those units, which serve as the primary representations of speaker and listener alike, are presumed to be something like the coordinate structures that Turvey and others have assumed26–29. As expressed in the peripheral structures, the gestures are changes in the cavities of the vocal tract – openings or closings, widening or narrowings, lengthenings or shortenings. They are supposed on the vertical view that the processes underlying the gestures evolved with language, specifically in the service of a linguistic function; hence, the gestures are phonetic af initio, requiring no cognitive translation to make them so30–33. Coincident with the evolution of the phonetic mechanisms was what Studdert-Kennedy has called the 'particulation' of the vocal tract30 – that is, the division of vocal-tract action into independently controllable components (for example, the categorical perception of phonetic elements) in the service of phonetic privileges and so made usable by the special processes of language. That translation is variously accomplished by matching the primary auditory percept to phonetic templates34–37, by associating it with the abstract 'distinctive features' so familiar to linguists38, or, as we are left to infer by some treatments, simply by giving it a phonetic name39–41. For the purposes of this review, these are distinctions without a difference; in assuming the need for translation to linguistic status, however done, the horizontalist accepts that language is a specialization, while at the same time denying that speech is part of it.

The vertical alternative

The first assumption of the vertical view is that the phonetic elements of speech, the true primitives that underlie linguistic communication, are not sounds but rather the articulatory gestures that generate those sounds21–25. Those units, which serve as the primary representations of speaker and listener alike, are presumed to be something like the coordinate structures that Turvey and others have assumed26–29. As expressed in the peripheral structures, the gestures are changes in the cavities of the vocal tract – openings or closings, widening or narrowings, lengthenings or shortenings. They are supposed on the vertical view that the processes underlying the gestures evolved with language, specifically in the service of a linguistic function; hence, the gestures are phonetic af initio, requiring no cognitive translation to make them so30–33. Coincident with the evolution of the phonetic mechanisms was what Studdert-Kennedy has called the 'particulation' of the vocal tract30 – that is, the division of vocal-tract action into independently controllable components (for example, the categorical perception of phonetic elements) in the service of phonetic privileges and so made usable by the special processes of language. That translation is variously accomplished by matching the primary auditory percept to phonetic templates34–37, by associating it with the abstract 'distinctive features' so familiar to linguists38, or, as we are left to infer by some treatments, simply by giving it a phonetic name39–41. For the purposes of this review, these are distinctions without a difference; in assuming the need for translation to linguistic status, however done, the horizontalist accepts that language is a specialization, while at the same time denying that speech is part of it.

The vertical alternative

The first assumption of the vertical view is that the phonetic elements of speech, the true primitives that underlie linguistic communication, are not sounds but rather the articulatory gestures that generate those sounds21–25. Those units, which serve as the primary representations of speaker and listener alike, are presumed to be something like the coordinate structures that Turvey and others have assumed26–29. As expressed in the peripheral structures, the gestures are changes in the cavities of the vocal tract – openings or closings, widening or narrowings, lengthenings or shortenings. They are supposed on the vertical view that the processes underlying the gestures evolved with language, specifically in the service of a linguistic function; hence, the gestures are phonetic af initio, requiring no cognitive translation to make them so30–33. Coincident with the evolution of the phonetic mechanisms was what Studdert-Kennedy has called the 'particulation' of the vocal tract30 – that is, the division of vocal-tract action into independently controllable components (for example, the categorical perception of phonetic elements) in the service of phonetic privileges and so made usable by the special processes of language. That translation is variously accomplished by matching the primary auditory percept to phonetic templates34–37, by associating it with the abstract 'distinctive features' so familiar to linguists38, or, as we are left to infer by some treatments, simply by giving it a phonetic name39–41. For the purposes of this review, these are distinctions without a difference; in assuming the need for translation to linguistic status, however done, the horizontalist accepts that language is a specialization, while at the same time denying that speech is part of it.

The vertical alternative

The first assumption of the vertical view is that the phonetic elements of speech, the true primitives that underlie linguistic communication, are not sounds but rather the articulatory gestures that generate those sounds21–25. Those units, which serve as the primary representations of speaker and listener alike, are presumed to be something like the coordinate structures that Turvey and others have assumed26–29. As expressed in the peripheral structures, the gestures are changes in the cavities of the vocal tract – openings or closings, widening or narrowings, lengthenings or shortenings. They are supposed on the vertical view that the processes underlying the gestures evolved with language, specifically in the service of a linguistic function; hence, the gestures are phonetic af initio, requiring no cognitive translation to make them so30–33. Coincident with the evolution of the phonetic mechanisms was what Studdert-Kennedy has called the 'particulation' of the vocal tract30 – that is, the division of vocal-tract action into independently controllable components (for example, the categorical perception of phonetic elements) in the service of phonetic privileges and so made usable by the special processes of language. That translation is variously accomplished by matching the primary auditory percept to phonetic templates34–37, by associating it with the abstract 'distinctive features' so familiar to linguists38, or, as we are left to infer by some treatments, simply by giving it a phonetic name39–41. For the purposes of this review, these are distinctions without a difference; in assuming the need for translation to linguistic status, however done, the horizontalist accepts that language is a specialization, while at the same time denying that speech is part of it.

The vertical alternative

The first assumption of the vertical view is that the phonetic elements of speech, the true primitives that underlie linguistic communication, are not sounds but rather the articulatory gestures that generate those sounds21–25. Those units, which serve as the primary representations of speaker and listener alike, are presumed to be something like the coordinate structures that Turvey and others have assumed26–29. As expressed in the peripheral structures, the gestures are changes in the cavities of the vocal tract – openings or closings, widening or narrowings, lengthenings or shortenings. They are supposed on the vertical view that the processes underlying the gestures evolved with language, specifically in the service of a linguistic function; hence, the gestures are phonetic af initio, requiring no cognitive translation to make them so30–33. Coincident with the evolution of the phonetic mechanisms was what Studdert-Kennedy has called the 'particulation' of the vocal tract30 – that is, the division of vocal-tract action into independently controllable components (for example, the categorical perception of phonetic elements) in the service of phonetic privileges and so made usable by the special processes of language. That translation is variously accomplished by matching the primary auditory percept to phonetic templates34–37, by associating it with the abstract 'distinctive features' so familiar to linguists38, or, as we are left to infer by some treatments, simply by giving it a phonetic name39–41. For the purposes of this review, these are distinctions without a difference; in assuming the need for translation to linguistic status, however done, the horizontalist accepts that language is a specialization, while at the same time denying that speech is part of it.
important, however, because they increase the possibilities for co-articulation beyond what could be done with discreet segments. For the particular purposes of this paper, the difference between segment and gesture is of no great importance, so for convenience, in what follows we will not distinguish them. At any event, we will not try to explain what we do not yet understand, which is how the one is gathered up so as to form the other.

The second assumption is that the articulation and co-articulation of the gestures is controlled by a species-specific specialization for language – a phonetic module. The inventory of specialized phonetic gestures it controls is severely limited, both in the style of movement of the gesture (manner of production) and the surface of the vocal tract that is the target (place of articulation). As already implied, those restricted gestures form a natural class, specifically phonetic in nature, that stands apart from the non-phonetic movements (e.g., chewing, swallowing, moving food around in the mouth) made by the same organs. A critical function of the phonetic module is to take advantage of the particular vocal tract so as to produce a great deal of overlapping, interleaving and merging of the gestures, while precluding those temporal relationships among the gestures that would cause the acoustic consequences of the one or the other to be obscured.

The third assumption, which is about perception of the elemental gestures, takes account of the fact that co-articulation creates a complex relationship between the acoustic signal and the phonetic structure it conveys. As is well known, there is a lot of context-conditioned variation in the acoustic information for any phonetic segment, as well as a lack of correspondence in segmentation between acoustic signal and phonetic structure. Unraveling that complex relationship between signal and message is the business of the same phonetic module that produced it, for that module incorporates the constraints necessary to process the signal so as to recover the very gestures that were, by their co-articulation, responsible for its apparent complications. (As indicated at the beginning of this section, Fowler and Liberman have called it the requirement for “pauci”.

As should be said again, the initial representation is already abstract phonetic unit to which it is somehow linked, because, as should be said again, the initial representation is already phonetic. To see the phonetic module even more clearly as a component of the larger language module, a theorist does well to note how far its processes resemble those of syntax. Consider, then, how the phonetic module presumably goes about deciding that some signal does or does not contain information about phonetic structures. The important point is that it can hardly do that by reference to the surface properties of the acoustic signal, as might be seen most readily in the case of sine-wave speech. There, the formants and the band-limited noises (of the fricatives) are replaced by sinusoids that follow the centers of the formants and the noises. Those sinusoids have no common fundamental, no common movement, nor indeed, any acoustic property that would make them coherent from an auditory standpoint. Yet, if they describe phonetically telling trajectories of the articulators, listeners hear phonetic structures. Thus, the only basis for perceptual coherence is at the same time gestural and phonetic. That being so, the phonetic module presumably goes about deciding that some signal does or does not contain information about phonetic structures. The important point is that it can hardly do that by reference to the surface properties of the acoustic signal, as might be seen most readily in the case of sine-wave speech. There, the formants and the band-limited noises (of the fricatives) are replaced by sinusoids that follow the centers of the formants and the noises. Those sinusoids have no common fundamental, no common movement, nor indeed, any acoustic property that would make them coherent from an auditory standpoint. Yet, if they describe phonetically telling trajectories of the articulators, listeners hear phonetic structures. Thus, the only basis for perceptual coherence is at the same time gestural and phonetic. That being so, we should not expect to identify surface acoustic characteristics that will reliably distinguish three sine waves that cause the listener to perceive phonetic structures from those acoustically similar sine waves that will be heard as unrelated tones. In that respect, phonetic
The relation of speech to language

Opinion  
Liberman and Whalen – The relation of speech to language

The signal is speech if and only if the pattern of phonetically significant articulatory gestures that must have produced it can be reconstructed. We call this property “genera
duction”, having in mind the analogous situation in the domain of sentence processing. There, superficial features cannot distinguish grammatical sentences from ungrammatical ones. The only way to determine the
grammaticality of a sentence is to parse it – that is, to try to regenerate the syntactic structure intended by the speaker.

(Ref. 38, p. 785)

Surely it is appropriate, if speech is a component of the
larger specialization for language, that its perception should rely on the same kind of synthetic processing that character-
izes syntax. It is equally appropriate that the gestures should have evolved with language in the service of an exclusively phonetic function, and that they are, therefore, like syntactic structures, linguistic by their very nature.

The second facet of the parity requirement applies to
two-way communication – that is, communication in which speaker and listener exchange roles. In this case, it is essential
that the phonetic representation in the brain of the one party be replicated in the brain of the other; otherwise they cannot communicate in the same code. But on the horizontal view, the primary representation of the one party is, at any given moment, purely motor, whereas the representation of the
other is purely auditory. Those representations are in no way alike, except that neither has anything to do with language. Where, then, do the parties find the common ground on which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.

Where, then, do the parties find the common ground on
which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar
motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.

Where, then, do the parties find the common ground on
which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar
motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.

Where, then, do the parties find the common ground on
which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar
motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.

Where, then, do the parties find the common ground on
which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar
motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.

Where, then, do the parties find the common ground on
which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar
motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.

Where, then, do the parties find the common ground on
which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar
motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.

Where, then, do the parties find the common ground on
which they must stand, and what reserves that ground exclu-
sively for events that have communicative significance? We
suppose the horizontalist would say that the very dissimilar
motor and perceptual representations are connected to the
same phonetic unit. But that runs foul of the troubling ques-
tions, raised earlier, about the nature and origin of such
a phonetic unit and the process by which certain motor and
perceptual representations are connected to it.

The vertical view, alternatively, permits us to see that the
parties conduct their business in the common currency of
phonetic gesture: the gesture that is the unit of production
for the speaker is the unit of perception for the listener. Those representations are, in no way alike, with both
what they must have to do with language.
discover, as Cyril Harris did many years ago, that a tape would be similarly limited, as it would require listening to the alphabet. But speech cannot be an acoustic alphabet, for if it were, we could speak only as fast as we can spell. Perception would be similarly limited, as it would require listening to spelled speech. No one should have been surprised, then, to discover, as Cyril Harris did many years ago, that a tape recording of speech cannot be broken up into phoneme-size segments and then rearranged to yield an intelligible permutation. Neither should anyone have been surprised to learn from the very early research on speech that it is, in fact, not an acoustic alphabet but something more like what the vertical view takes it to be. On the vertical view, the particular (hence alphabetic) nature of speech is represented by the elemental gestures, which are, as they must be, discrete, invariant and categorical. They are produced and perceived as rapidly as language requires because they are co-articulated, which preserves their commutability at the level of the gestures but not in the acoustic signal.

How is perceptual form fit to phonetic function?

That speech is normally co-articulated is established beyond dispute, as are many of the consequences for the acoustic signal. Indeed, that part of the vertical view is so widely accepted that it can hardly be regarded as unconventional. What is just as widely rejected, however, is the essential vertical premise that the elements of speech are phonetically significant gestures, not sounds. The corollary of that premise, also explicitly rejected, is that a specialization for language processes the gestures that are appropriate for language without integrating into a single perceptual unit (/p/) information that is spread across the adjoining units. What could possibly have been more useful? And would they not have been dysfunctional for non-speech perception, causing continuously varying acoustic signals that reflect continuously varying events to be divided, percep- tually, into discrete and disconnected units that would in no way reflect the physical reality? Or, in other cases, to integrate into a unitary percept information relevant to distinct and successive physical events?

On the vertical view, however, it is possible to see that all of the aforementioned effects of phonetic perception are what one would expect if it is the gesture that is produced and perceived. For then, in the case of /spi/, the broadly spread acoustic cues are the common products of the same gestural segments: in the case of /fag/ there are three discrete but overlapped gestures; and in the case of the ‘silent center’ vowels, information about the vowel gesture is spread throughout the syllable, just as it is in /bag/. Thus, the phonetic specialization processes the gestures that are appropriate for language without in any way interfering with the ability of the auditory system to offer a veridical representation of the acoustic world.

It is also widely accepted that the phonetic segments – most obviously the consonants – tend to be perceived categorically, and, further, that the categorical perception of each segment is invariant across all the conditions that cause the vocal tract and the ear, the particles must be laid down in strings, and given the limitations of the vocal tract, the number of particles must be small. A consequence is that the strings will typically run to considerable lengths. It is essen- tial, therefore, that the particles be produced and perceived rather rapidly if they are to be organized into the larger units of the language hierarchy. In fact, the consonants and vowels that are formed by the particles are delivered in speech at about 10 or 12 per second on average, and for short stretches the rate rises even higher.

How do discrete, invariant and categorical elements meet the rate requirements? Consider now that if the elements of speech were sounds, as the horizontal view has it, then the particular nature of speech would necessarily be manifest at the acoustic surface. In that case, speech would be an acoustic alphabet. But speech cannot be an acoustic alphabet, for if it were, we could speak only as fast as we can spell. Perception would be similarly limited, as it would require listening to spelled speech. No one should have been surprised, then, to discover, as Cyril Harris did many years ago, that a tape recording of speech cannot be broken up into phoneme-size segments and then rearranged to yield an intelligible permutation. Neither should anyone have been surprised to learn from the very early research on speech that it is, in fact, not an acoustic alphabet but something more like what the vertical view takes it to be. On the vertical view, the particular (hence alphabetic) nature of speech is represented by the elemental gestures, which are, as they must be, discrete, invariant and categorical. They are produced and perceived as rapidly as language requires because they are co-articulated, which preserves their commutability at the level of the gestures but not in the acoustic signal.

How is perceptual form fit to phonetic function?

That speech is normally co-articulated is established beyond dispute, as are many of the consequences for the acoustic signal. Indeed, that part of the vertical view is so widely accepted that it can hardly be regarded as unconventional. What is just as widely rejected, however, is the essential vertical premise that the elements of speech are phonetically significant gestures, not sounds. The corollary of that premise, also explicitly rejected, is that a specialization for language processes the gestures that are appropriate for language without integrating into a single perceptual unit (/p/) information that is spread across the adjoining units. What could possibly have been more useful? And would they not have been dysfunctional for non-speech perception, causing continuously varying acoustic signals that reflect continuously varying events to be divided, perceptually, into discrete and disconnected units that would in no way reflect the physical reality? Or, in other cases, to integrate into a unitary percept information relevant to distinct and successive physical events?

On the vertical view, however, it is possible to see that all of the aforementioned effects of phonetic perception are what one would expect if it is the gesture that is produced and perceived. For then, in the case of /spi/, the broadly spread acoustic cues are the common products of the same gestural segments: in the case of /fag/ there are three discrete but overlapped gestures; and in the case of the ‘silent center’ vowels, information about the vowel gesture is spread throughout the syllable, just as it is in /bag/. Thus, the phonetic specialization processes the gestures that are appropriate for language without in any way interfering with the ability of the auditory system to offer a veridical representation of the acoustic world.

It is also widely accepted that the phonetic segments – most obviously the consonants – tend to be perceived categorically, and, further, that the categorical perception of each segment is invariant across all the conditions that cause the
acoustic signal to vary. The consistent finding about categorical perception is that, given stimuli that are adjacent on some acoustic continuum between two consonants, there is an increase in discriminability at the phonetic boundary55. That increase establishes two categories, one on either side of the perceptual discontinuity that the increased discriminability reflects. (The increase in discriminability is normally not so great as to indicate that listeners can discriminate only as well as they can identify, which would be the case if perception were perfectly categorical. But the shortfall may be a consequence of the limitations of acoustic synthesis. As it is used in experiments, such synthesis typically changes only one of the relevant acoustic cues, causing the percept to depart progressively from the phonetic form and to be perceived therefore as having more and more of a non-phonetic cast; a result of this is that discrimination of a purely auditory kind crops in. A remedy does not come easily, not only because there is presently available no articulatory synthesizer that would change all the acoustic cues appropriately, but, more importantly, because the articulatory routes from one consonant to another are not continuous. That is to say that production is no less categorical than perception, which is, of course, precisely the point.)

On the horizontal view, there have been two very different explanations of categorical perception. One is that ‘categoricalism’ is a consequence of the cognitive component of the two-stage process that the horizontalists must assume if their primary non-linguistic auditory percepts are to be given phonetic status. Thus, several investigators have assumed that the listener attaches a phonetic label before the echo of the primary auditory percepts has faded26.56.57. It is, therefore, the label, not the primary percept, that is categorical. However, as in all other attempts to reconcile a horizontal view with the requirement of parity, there is no basis for understanding, even in the most general terms, why the phonetic labels come from, or how it is decided which labels go with which auditory percepts.

The more common explanation of categorical perception is that the discontinuity is a property of the auditory system. To evaluate that explanation, it is relevant first to reflect on the widely recognized fact that the acoustic boundaries between segments are not fixed. Even if we accept that some aspects of the acoustic signal are invariant for consonants, there remain the formant transitions—sufficient cues for most stop-consonant perception—which invariably change with phonetic context55. The transitions also move as a function of position in the syllable, most dramatically when, as cues for initial and final consonants, they are mirror images59. Lexical stress and rate of articulation affect the phonetic boundaries for the voicing distinction55,56. Further, it is also surely relevant that, although the acoustically defined dimensions might be the same across many languages, the placement of the phonetic boundaries varies60,61. Given these variations in the salient acoustic cues, the number of category-determining discontinuities in the auditory system is beyond enumerating. One wonders, then, what would have selected for those numerous auditory discriminations? Again, it was surely not in anticipation of their usefulness for a behavior—speech—which had yet to appear on the biological scene. In any case, would not those auditory categories be seriously maladaptive, grossly distorting physical reality by causing perception of continuous non-speech acoustic events to appear discontinuous? And is it not implausible to suppose that speakers are able to manage their limited articulatory possibilities so as to make the acoustic result conform to the many categories defined by those highly variable discontinuities? On the vertical view, by contrast, one sees that the perceptual boundaries are exactly where the conveniences of articulation and co-articulation put them, not where the properties of the auditory system would have set them down.

The horizontal view also appears the more implausible when one considers that phonetic boundaries are typically marked not by one acoustic cue but by several. For example, the differences among the voiced stops are signaled by the second-formant transition, but also by the third, and there is no way a speaker can control these independently64. If the speaker has managed her articulations so as to produce signals appropriate to the second-formant boundary, what is the probability that the acoustic result would happen to be appropriate also for the third-formant boundary, and, moreover, that this happy but highly improbable coincidence would occur for all phonetic segments, for all contexts, and for all positions in the syllable?

Still another difficulty for the horizontal view arises from the fact that there is, within limits, perceptual equivalence among the several acoustic cues for each category, no matter how acoustically diverse they may be. Thus, the difference between /f/ and /v/ can be caused by varying only the second-formant transition, but also by the third, and there is no way a speaker can control these independently64. If the speaker has managed her articulations so as to produce signals appropriate to the second-formant boundary, what is the probability that the acoustic result would happen to be appropriate also for the third-formant boundary, and, moreover, that this happy but highly improbable coincidence would occur for all phonetic segments, for all contexts, and for all positions in the syllable?

Speech versus reading and writing.

How do we account for the biological gulf that separates speech from the reading and writing of its alphabetic transcription? The prehistoric child is a prodigy of phonetic development. Commanding thousands of words, he readily produces their phonetic structures when speaking, and just as readily parses them when listening. Thus, he exploits the particulate principle quite naturally, without its ever having been taught to him, and without his having to be aware of the principle or of the remarkable ability it makes available to him. For the skillful use of that principle in speech, it is enough to be a normal member of the human race and to
have been exposed to a mother tongue. By contrast, applying the particular principle to the task of reading and writing is not an automatic outgrowth of the natural capacity for language but an achievement of a distinctly intellectual kind. To understand the reading process, one would think it critical, therefore, to know exactly where the biological difference lies.

Yet, in all the vast literature on reading, that question is never answered or even asked (but see Refs 69,70, certainly not by that overwhelming majority of researchers who explicitly or tacitly accept a horizontal view. A likely reason for this serious omission is that, given the horizontal assumption, no reasonable explanation is possible. Consider that, on any view, the relationship between the alphabet and speech is entirely arbitrary: the visual percepts that are then cognitively translated into the linguistic units they symbolize. Accordingly, reading is always a translation, and translation is, by its nature, effortful and deliberate. But on the horizontal view, the same requirement is imposed on speech. As we saw earlier, on that view, the sounds of speech are thought to evoke auditory percepts which, like the visual percepts of the reader, become linguistically useful only after translation into the phonetic segments they happen to convey. Indeed, one might expect from the horizontal view that reading and writing would be easier than speech. After all, the alphabetic characters are clearer signals than the sounds of speech; the hand and fingers are more versatile effectors than the tongue; and the eyes are more accommodating receptors than the ears.

Even so, parting aside the notion that speech should be harder if one accepts the horizontal view, one must wonder why reading it, at least, not equally easy. Why does mastery of speech not fully prepare the speaker or listener for the seemingly trivial task of substituting the letters of the alphabet for the sounds of speech? An important and proper answer was provided by I.Y. Liberman and Shankweiler71, who brought to notice that, contrary to what the horizontalist puts humans at odds with the biology of communication as it is evident in all other species. Which way, then, dare assume about a non-human creature that it perceives communicative signals exactly as it perceives all others, recognizing their special significance only by some secondary assumption attributes to human speech perception? Presumably not. So, in assuming a cognitive translation for speech, the horizontalist puts humans at odds with the biology of communication, not just to the one we humans enjoy. Would we, then, dare assume about a non-human creature that it perceives communicative signals exactly as it perceives all others, recognizing their special significance only by some secondary cognitive process similar to that which the horizontalist assumption attributes to human speech perception? Presumably not.

How special is speech?

The claim from the vertical standpoint that speech in the narrow sense is a specialization has seemed to some to call for the application of Occam’s razor. Why have a special system when something more general might do79? The answer is that a more general process will not do. At the very least, there must be parity, which requires that signals with communicative significance belong to a special mode, as it were, where they are clearly marked for their distinct communicative function. That requirement applies to all animal communication, not just to the one we humans enjoy. Would we, then, dare assume about a non-human creature that it perceives communicative signals exactly as it perceives all others, recognizing their special significance only by some secondary cognitive process similar to that which the horizontalist assumption attributes to human speech perception? Presumably not. So, in assuming a cognitive translation for speech, the horizontalist puts humans at odds with the biology of communication as it is evident in all other species. Which way, then, do we want Occam’s razor to cut?

To see how speech perception can be put in a class of perceptual specializations, and so made to appear less exceptional from a biological point of view, we should first take note of one of its most apparent but least remarked characteristics, which is that phonetic units do not have an end organ of their own. Accepting Berlekamp’s explicit treatment of the matter in his ‘New Theory of Vision70’, many psychologists assume that a primary percept is evoked only by an appropriate end organ. It follows that a phonetic representation must be a translation from the ordinary auditory primitives that own the ear as their end organ, and presumably stand, therefore, as the only primary percepts that can be evoked by the acoustic signals to which the ear responds. However, there are other percepts that are primary, yet, like speech, have no end organ, and hence no labeled line to peripheral equipment that is dedicated to their needs. The most relevant, perhaps, is sound localization. There, the acoustic cues are internal differences of time and intensity, but nobody takes those differences to be the primary percepts that are then cognitively translated into location. Rather, it is understood that there is a system
specialized to process the intraural cues and represent them immediately as location. Visual perception of depth presents a similar case in that information about binocular disparity is processed by a system specialized to represent it immediately as depth.

These systems all depend on what Komihi has called a ‘central synthesis’32. As Hering and Lashley have put it37, such systems are ‘heteromorphic’, in that the percept is incommensurate with the stimuli, but only when the stimuli conform reasonably to the ecological circumstances for which the perceptual system is specifically adapted; otherwise, they assume a ‘homomorphic’ form. Thus, except when the intraural time and intensity differences far exceed what is ecologically possible, the stimuli for sound localization are not heard homomorphically as sounds that arrive at the ears at different times or with different intensities, but heteromorphically, as location in azimuth. Stereopsis presents a similar case: except when the binocular disparity is far beyond normal limits, the stimuli for visual depth are not perceived homomorphically as disparate images but heteromorphically as phenomenal depth. For speech, the stimulus is an ensemble of several resonances that change their spectral positions more or less continuously, yet, except when those changes do not reflect the trajectories of phonetically significant gestures, they are not perceived homomorphically as continuously changing timbres, but heteromorphically as a string of discrete and categorical segments.

The heteromorphic specializations have several characteristics in common with phonetic perception. One is a plasticity that allows them to be calibrated by the environment. Thus, the specialization for stereopsis, depending as it does on binocular disparity, must be recalibrated as the child’s head grows and the distance between the eyes increases. Much the same must happen in the case of sound localization, as growth causes a change in the distance between the ears and therefore in the interaural time difference. Such calibration represents a kind of learning. It is not the kind of learning that psychologists commonly study, which is unfortunate for our purposes, because it is the kind that occurs also in speech. For in speech, as in the other cases, the specialization is calibrated by the environment. The necessary and sufficient condition for appropriate phonetic calibration is simply exposure to the right environment; the required perceptual ‘learning’ is effortless and, for neurologically normal children, inevitable.

Speech perception also shares with the heteromorphic specializations an elasticity that allows them to respond in their usual way to stimuli that are, within limits, ecologically impossible. Thus, viewers will perceive depth even when the binocular disparity is made to correspond to a greater interocular difference than a real head could ever provide. Beyond a certain disparity, however, the limit of elasticity is reached, and the viewer perceives not only homomorphically depth but also homomorphically double vision (diplopia). With further departures from what is ecologically possible, depth perception ceases entirely, leaving only the homomorphic (diplopic) representation33. Phonetic perception has been shown experimentally to behave in a strikingly similar way34,35. Ecologically impossible speech-like stimuli were created by dividing the signal into two parts that could not have come from the same source. One part presented as sinusoids the cues critical for the distinction between two consonants. These were connected to an acoustic base that conformed to the normal resonances of speech. In isolation, the sinusoids sounded like non-speech whistles that differed in pitch. The base was perceived as a consonant–vowel syllable, but in the absence of the critical cues, the consonant was ambiguous. To control the evidence for two independent sources, and thus for ecological plausibility, the experimenters varied the intensity of the sinusoids, while holding the intensity of the base constant36. When the intensity of the sinusoids was very low, the elasticity of the module enabled it to accommodate them, with the result that the signals engaged the phonetic system and caused listeners to perceive both consonant and vowel correctly. But at those levels, the whistles that the sinusoids produced in isolation were identified at a chance level; indeed, they were not even heard.

It is noteworthy that an analogous result has been obtained by Eimas and Miller37 with two- to four-month-old infants. One supposes that the phonetic percepts could hardly have been a translation from auditory percepts, because the auditory percepts had not yet become identifiable. Further increases in the intensity of the sinusoids in the experiment by Xia et al.38 strained the elasticity of the phonetic module, causing the listeners to hear simultaneously (and correctly) both the consonants and the whistles. That is, exactly the same acoustic signal produced in the same brain, at the same time, two very different percepts, one distinctly phonetic, the other not. Finally, as still higher intensities, where the elasticity of the module could no longer accommodate the strong evidence for two sources, the sinusoids no longer engaged the phonetic system, producing only the homomorphic whistles and the ambiguous syllabic base.

That kind of experiment not only relates phonetic perception to stereopsis, another perceptual specialization that lacks an end organ and produces a heteromorphic percept, but also shows quite directly that the primary perceptual response to speech is phonetic and independent of its auditory counterpart. The fact that the consonants were correctly identified at levels of intensity where the whistles were not has been called ‘phonetic precedence’39. It nicely illustrates the exquisite sensitivity of the phonetic module when it does what it was specifically adapted to do. As we have seen, that adaptation was to several requirements that language meets. Had the auditory system been bent to accommodate those requirements, it would have become useless for the purpose of rendering accurately the sounds of the non-linguistic world.

The biological solution was the evolution of a distinct phonetic mode as part of the larger language mode, not in the higher reaches of the cognitive machinery, but down among the nuts and bolts of action and perception.

Conclusion
We have examined two starkly contrasting theories about the production and perception of the sounds that convey phonetic structure: the more conventional, horizontal theory, which holds that speech production is linked to ordinary (non-linguistic) motor and auditory representations that are then constrained by purely cognitive means to language proper; and the less conventional, vertical theory, according to which the
primary representations are immediately phonetic gestures of the articulatory apparatus, having been produced in a specialized phonetic mode that serves as the basis of the larger specialization for language. The aim of this essay was to show that the verbal theory provides the more plausible answers to important questions of a biological kind, questions that are, for some reason, rarely asked. Thus, it is possible to see how, by creating distinctly phonetic motor structures to serve as the ultimate constituents of language, the phonetic specialization enables speech to meet the requirements for serving as the basis of the larger articulatory apparatus, having been produced in a specialization that serves as the basis of the larger specialization for language. The aim of this essay was to show that the verbal theory provides the more plausible answers to important questions of a biological kind, questions that are, for some reason, rarely asked. Thus, it is possible to see how, by creating distinctly phonetic motor structures to serve as the ultimate constituents of language, the phonetic specialization enables speech to meet the requirements for serving as the basis of the larger articulatory apparatus, having been produced in a specialization that serves as the basis of the larger specialization for language.