PROCEEDINGS
XIth ICPhS

The Eleventh International Congress of Phonetic Sciences
August 1—7, 1987
Tallinn
Estonia, U.S.S.R.

VOLUME
2

Tallinn 1987
INTEGRATION AND SEGREGATION IN SPEECH PERCEPTION

BRUNO H. REPP

Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695

In this paper I present an overview of some recent research on speech perception. To reduce my task to manageable size, I have chosen to focus on the topics of perceptual integration and segregation, which have guided, more or less explicitly, a considerable amount of speech perception research and theorizing in recent years. This will be a selective review, therefore, but I hope it will nevertheless convey some of the flavor of contemporary ideas and findings, even though that flavor will be tinged with my own favorite spices.

CONCEPTUAL FOUNDATIONS

Integration and segregation are hypothetical perceptual functions (or processes) that link physical structures in the world with mental structures in the brain. An integrative function maps multiphysical units (trivially, a single physical unit) onto a single mental unit, whereas a segregative function maps multiple physical units (sometimes, paradoxically, a single physical unit) onto different mental units. Though mutually exclusive for any particular physical structure at any given time, these two processes nevertheless cooperate in sorting a complex stream of sensory inputs into an orderly sequence of perceived objects and events.

These definitions seem rather straightforward, but they rest on four important assumptions: (1) The physical and mental worlds are not isomorphic. (2) There are objectively definable units in the physical world. (3) There are units in the mental world that are different from the physical units. (4) There are perceptual functions or processes that accomplish the mapping between the two types of units. I will briefly defend each of these assumptions; at the end of this presentation I will consider the consequences of abandoning some or all of them.

The first assumption, that the mental world is not isomorphic with the physical world, reflects the facts that physical variables are filtered and transformed by sensory systems, that perception is a function not only of the current sensory input but also of the past history of the organism, and that there is often an element of choice in perception which allows alternative perceptual organizations for the same sensory input. Without this assumption, it would be difficult to say anything meaningful about perception, except that it happens.

The second assumption, concerning the existence of physical units, is necessary in order to be able to talk about perceptual integration. These units or dimensions are what is being integrated. Perceptual segregation, too, ordinarily implies that certain objective lines of division can be found in the sensory input. It is always possible to find a physical description that is more finely grained than our description of the perceptual end product. The fact that the machines we use to assess physical characteristics of speech are mere transducers (or, at best, model only peripheral auditory processes) generally assumes a mismatch between physical and perceptual descriptions even when the grain size is comparable (and even though our visual perception is engaged in interpreting the machine outputs). Although there are different ways of characterizing the physical energy pattern, they are all equally valid for descriptive purposes. It is an empirical question whether or not perceivers are sensitive to any observed physical divisions, i.e., whether these divisions can serve as the basis for perceptual segregation or whether they are bridged by integrative processes. Research of this kind may enable us to find a physical description with a simpler mapping onto perceptual units.

The third assumption concerns the existence and nature of perceptual (mental) units. There is no theory of speech perception that does not assume mental units, usually the ones supplied by linguistic theory. The argument has been over the "perceptual reality" of syllables, phonemes, and features, and over their relative primacy in perceptual processing (see, e.g., [69, 83, 95, 102, 146]). However, which level of the linguistic hierarchy is perceptually and behaviorally salient depends very much on the task and the situation a perceiver is in. As McNeill and Lindig ([102], p. 430) have aptly put it, "what is 'perceptually real' is what one pays attention to." The validity of the basic linguistic categories, questions of detail aside, is guaranteed by the success of linguistic analysis. Linguistic units provide us with a vocabulary in which to describe the time course of accumulation and perceptual processing of linguistic information. Even though the perceptual processes themselves may be of an analog nature, we need discrete concepts to theorize and communicate about these processes. From this perspective, it is not an empirical issue but a fact that perceivers process features, phonemes, syllables, words, etc., since they are what speech is made of. Their awareness of these categories is another
matter that shall not concern us here. (See [90, 99, 106].) Clearly, speech perception generally proceeds without awareness of all but the highest levels of description (i.e., the meaning of the message).

The fourth assumption is that there are perceptual processes in the brain that map sensory inputs onto internal structures. While such processes have been traditionally assumed in psychology since the demise of radical behaviorism, a new challenge (to the other assumptions as well) comes from the so-called direct realist school of perception, which claims that perceptual systems merely "pick up" the information delivered by the senses [54, 60]. I will return to this issue later. Here I merely note that the same input is not always perceived in the same way. Contextual factors, past experience, expectations, and strategies may alter the perceptual outcome, and this seems to require the assumption of perceptual processes that mediate between the input and the perceiver's interpretation of it. Whether these processes (and indeed, integration and segregation as such) are thought of as neural events with actual time and space coordinates or as abstract functional relationships between physical and mental descriptions is irrelevant to most of the research I will discuss here.

Having attempted to justify the four principal assumptions, it remains for me to mention two issues that are important in much research on perceptual integration and segregation. One is the question of whether the processes inferred are specific to the perception of speech or whether they represent general capacities of the auditory or cognitive system. By a speech-specific function, I mean one that operates on properties that are unique to speech. There is no question that general capacities to integrate and segregate are common to all perceptual and cognitive systems. Speech perception presumably results from a combination of general and speech-specific perceptual functions (see, e.g., [39]). Just as speech resembles other sounds in some respects and differs in others. One frequent research strategy, therefore, is to determine whether or not particular instances of integration or segregation can be observed in both speech and nonspeech perception. This question can be asked only if the physical characteristics of speech and nonspeech stimuli are comparable--a condition that is notoriously difficult to satisfy (see, e.g., [112]). The mental descriptions of speech and nonspeech are, by definition, different at some higher level; thus the empirical question is whether that level is engaged in a particular integrative or segregative process.

The other issue is whether a particular integrative or segregative function is obligatory or optional. This question is sometimes linked with that of speech-specificity in that a higher-level, speech-specific function might seem easier to disengage than a lower-level auditory one. This is true in so far as adopting the deliberate strategy of listening to speech as if it were nonspeech (which is often difficult to achieve) may have the effect of eliminating certain forms of integration or segregation. It seems to be difficult or impossible to disengage phonetic processes through conscious strategies within the speech mode (e.g., by linguistic parsing [135, 136]). Moreover, it has been suggested [86] that some speech-specific functions do not really represent a "higher" level of perception but rather a mode of operation that, because of its biological significance, takes precedence over nonspeech perception, and if so, these functions may indeed be difficult to manipulate. On the other hand, in the auditory (nonspeech) mode listeners often have a variety of perceptual strategies available, especially when there are few ecological constraints on the stimulation, even though certain functions of peripheral auditory processing are surely obligatory. Thus, although it is useful to gather information about the relative flexibility of a process, this may not bear directly on the question of speech-specificity, as both speech and nonspeech perception are likely to involve levels of varying rigidity.

One final prefatory remark: Although one may legitimately talk about the integration of syllables into words and of words into sentences, or about the segregation of syntactic constituents from each other, I am not going to consider such higher linguistic processes in the present review. By speech perception, I mean primarily the perception of phonetic structure without regard to lexical status or meaning, and my review is restricted accordingly.

**INTEGRATION**

The function of integrative processes is to provide coherence among parts of the input that "belong together" according to some perceptual rule or criterion. Auditory integration occurs within the physical dimensions of time, (spectral) frequency, and even space (in the case of artificially split sources); thus it creates temporal, spectral, and spatial coherence of sound sources. In part this is due to the limited resolution of the auditory system along each of these dimensions, but auditory events will often cohere even when there are discriminable changes within them. The larger these changes are, the more noteworthy the integrative process will seem to us. The perception of phonetic structure involves, in addition, integration of relevant information across all physical dimensions of the speech signal--a function requiring higher-level perceptual or cognitive mechanisms.

**Temporal integration**

Basic processes of sensory integration and auditory organization ensure the temporal coherence of any relatively homogeneous auditory input, including components of speech. This form of integration is so obvious as to hardly deserve comment. Thus, for example, successive pitch periods of a vowel are perceived as belonging together (i.e., as a single vowel, not two or many) even though their duration and spectral composition may change as a function of intonation,
diphthongization, and coarticulation. While there
to
c
be
may be a physical basis for subdividing a sound
into smaller units such as individual pitch pulses
or transition versus steady state, the rate and
extent of change from one unit to the next are too
small to disrupt sensory integration.

Nevertheless, changes occurring within such units
(e.g., transitions in a vowel or fricative noise)
may have perceptual effects. That is, perception
of temporal coherence does not imply insensitivity
to changes over time, only that these changes are
not large enough to cause perceptual segregation.

Growth of loudness. Temporal integration at
this most elementary level has the consequence
that, as the duration of a relatively homogeneous
sound increases, its perceived loudness or
perceptual prominence will also increase, up to a
 certain limit. In psychoacoustic research, the
lowering of the detection threshold and the growth
of loudness with increasing stimulus duration are
well established phenomena (see, e.g., [26, 192]). The
time constant of the (exponential) integration
function is about 200 ms, which encompasses the
durations of virtually all relatively homogeneous
speech events. While loudness judgments or
explicit threshold measurements are uncommon in
speech perception research, the effect of an
increase in the duration of a signal portion can be
shown to be phonetically equivalent to that of an
increase in its intensity, especially when the
relevant signal portion is brief.

One example is provided by studies in which the
duration and relative intensity of aspiration
noise were varied orthogonally as cues to the
voicing distinction in synthetic syllable-initial
English stop consonants [31, 126]. Although the
trading function obtained was much steeper than the
traditional auditory temporal integration function, it
bore some similarity to integration functions
obtained in an auditory backward masking situation
[189], which is not unreasonable in view of the
time-intensity reciprocity reflects basic
properties of the auditory system, rather than
speech-specific processes. Indirect support for
this hypothesis comes from a study showing that the
 trading relation between aspiration duration and
intensity holds regardless of whether or not
listeners can rely on phonemic distinctions in
discriminating speech stimuli [131]. In another
recent study, stop consonant release burst duration
and intensity were varied in separate experiments
as cues to stop consonant manner in /b/-/p/
clusters [134]. Since both parameters proved to be
perceptually relevant, a trading relation between
them was implied. An analogous conclusion may
be drawn from an older informal study [88], in which
the duration and intensity of stop closure voicing
were varied as cues to the perceived voicing status
of an intervocalic stop consonant.

Auditory short-term adaptation. An effect
closely related to temporal integration is that the
auditory nerve fibers responsive to a continuous
sound become increasingly adapted. Auditory
adaptation is well known to great interest to
psychoacousticians and auditory physiologists, who
have identified at least three different time
constants or adaptations in animals" (see, e.g.,
[45]). So-called auditory short-term adaptation,
with a time constant of about 60 ms, seems the most
relevant to phonetic perception. Although ongoing
adaptation seems to have no direct perceptual
consequences, the recovery of auditory nerve fibers
following the offset of a relatively homogeneous
stimulus results in reduced sensitivity to other,
spectrally similar inputs for a short time period.
Consequently, the auditory representation of a
speech component whose spectrum overlaps that of a
preceding segment will be modified. A striking
demonstration of such an interaction was provided
in recordings from cats' auditory nerve fibers
responding to synthetic /ba/ and /ma/ syllables [34, 35].
Even though the two syllables were identical except
for the nasal murmur in /ma/, the auditory response
at vowel onset was very different. The murmur,
having strong spectral components in the
low-frequency range, effectively acted as a
high-pass filter, reducing the neural response in the
low-frequency region at vowel onset. Recent
experiments suggest, however, that this particular
auditory interaction has no important consequences
for perception of nasal consonants under normal
listening conditions [136]. In a more artificial
situation, Summerfield and colleagues [160, 162]
have demonstrated an auditory aftereffect
attributed to short-term adaptation: A sound with
a uniform spectrum was perceived as a vowel when
preceded by a sound whose spectrum was the
complement of the perceived vowel's spectrum.
Generalizing to natural speech, these authors
point out that auditory adaptation effectively
enhances spectral change and thus may aid phonetic
perception in adverse listening conditions.

One general lesson to be learned from
psychoacoustic research on temporal integration,
adaptation, and other auditory interactions is that
adjacent portions of the speech signal should not
be thought of as mutually independent in the
auditory system. Whenever a particular component
is singled out for attention in careful analytic
listening (to the extent that this is possible),
influences of surrounding context on the perceived
sound must be reckoned with. It is important to
keep in mind, however, that listeners normally
do not listen analytically but rather attend to the
continuous pattern of speech. All peripheral
auditory transformations are a natural part of the
pattern and, because of past learning, are also
represented in a listener's long-term memory of
phonemic norms, which provide the criteria for
phonemic classification in a language. Since
auditory input and central reference both
incorporate the distortions imposed by the
peripheral auditory system, these distortions
cannot be said to either help or hinder speech
perception [136]. Only a change in auditory
transformations, as might be caused by simulated
or real hearing impairment, would prove disturbing to
listeners; in normal speech perception, peripheral
auditory processes probably do not play a very
important role.

PI 2.2.3
Spectral integration

Most speech sounds have complex spectra determined by the resonance frequencies of the vocal tract formants. These formants are usually visible as prominent energy bands in a spectrogram or as peaks in a spectral cross-section. Why are these bands perceived as a single sound with a complex timbre and not as separate sounds with simpler qualities? Why, indeed, are the individual harmonics of periodic "speech sounds" not heard as so many simultaneous tones? Even though these questions are provoked by our instrumental and visual method of spectral analysis, they are not unreasonable, since the ear operates essentially as a frequency analyzer. One answer to these questions is that we do process these spectral components, only we are not conscious of them and find it difficult to focus selectively on them when asked to do so. Multidimensional statistical analyses of vowel similarity judgments have confirmed that the lower formants function as perceptually relevant dimensions, even though they seem to blend into a complex auditory quality [56, 115, 119], and psychoacoustic pitch matching tasks have revealed that listeners can detect a number of lower harmonics in a complex periodic sound (e.g., [100, 114]). Some central integrative function must be responsible for the perceptual coherence and unity of all these spectral components.

Critical bands: Some spectral integration does take place in the peripheral auditory system. A large amount of psychoacoustic research has established the concept of critical bands, i.e., frequency regions over which spectral energy is integrated, and whose width increases with frequency in a roughly logarithmic fashion [105, 190]. It is now quite common to represent speech spectra on a critical-band frequency scale (the Bark scale) to better take account of the resolving power of the auditory system. However, critical bands cannot account for the fact that formants are integrated into a unitary percept, because the lower formants of speech are usually several critical bands apart, and thus potentially separable. Even the lower harmonics, especially of female and child speech, are spaced more than 1 Bark apart. Critical bands may explain why higher harmonics and higher formants are not well resolved auditorily, but these spectral components do not contribute much phonetic information.

It is difficult, therefore, to point to any direct consequences of critical band limitations for speech perception, except in hearing-impaired listeners, whose critical bandwidths are abnormally large. A recent study by Celsinger and Bienville [7] may serve as an example. These investigators digitized speech materials, degraded their spectra by simulating critical band integration ranging from one-half to seven times the normal widths, then they manipulated the spectra back into sound, and presented the degenerated spectra to normal listeners and to hearing-impaired listeners believed to have abnormally wide critical bandwidths according to independent psychoacoustic tests. The results showed that the degree of critical bandwidth filtering required to cause an intelligibility decrement was directly related to the subjects' measured critical bandwidth. Thus, normal subjects were sensitive to filtering at twice the normal bandwidths, while hearing-impaired subjects, though their intelligibility scores were lower to begin with, tolerated up to five times the normal bandwidths before any decrement of intelligibility occurred. Many other studies, too numerous to review here, have examined correlations between measures of critical bandwidth (or frequency resolution) and measures of speech perception in hearing-impaired individuals, with mixed results (see, e.g., [44, 152]). The looseness of the correlation may be accounted for by the facts that speech perception engages higher-level functions that help overcome peripheral limitations, often requires only relatively coarse spectral resolution, and relies on other physical parameters beside spectral structure.

Integration of harmonics. Given that the lower harmonics of a periodic speech sound are not automatically integrated by the peripheral auditory system, not to mention the lower formants themselves, the question of why they are grouped together in perception still needs to be answered. The most general answer is that they share a "common fate": They usually start and end at the same time; they are at integral multiples of the fundamental frequency; they have similar amplitude envelopes; and there is no alternative grouping that suggests itself. Below I will have more to say about the factors that may cause segregation of harmonics. Principles of auditory organization have received much attention in recent years (see, e.g., [10, 28, 184]), and one interesting conclusion from this research is that, even at such a relatively early stage in auditory processing, speech-specific criteria begin to play a role. They are speech-specific in the sense that a listener's tacit knowledge of what makes a good speech pattern influences the perceptual grouping of auditory components, as presumably does knowledge of other familiar auditory patterns. Yet another answer to the question of why harmonics (and formants) are grouped together is, therefore: They make a speech sound—that is, a complex sound that could possibly have emanated from a human vocal tract.

If it is the case that formant frequencies are salient parameters of speech perception (an assumption that is not made by some researchers who favor a whole-spectrum approach; e.g., [7, 154]), then it is of interest to ask how listeners estimate the actual resonance frequencies of the vocal tract from the energy distribution in the relevant spectral region. This question is especially pertinent with respect to the first formant (F1) in periodic speech sounds, for which critical bands are narrow and frequency difference limits are small. This means that the actual F1 frequency often falls between auditorily resolvable harmonics. Early work by Mushnikov and Chistovich [107] suggested that the brain takes the frequency of the single most intense harmonic as the estimate of F1. Later studies [1, 18], however, have...
indicated that the subjective F1 frequency corresponds to a weighted average of the two most intense harmonics, and one experiment [30] has shown that the perceptual boundary between /ɪ/ and /e/ can be affected by the intensity of as many as five harmonics between 250 and 750 Hz, spaced 125 at apart. This indicates that the weighting function applied by the speech perception system in estimating formant frequencies extends over several critical bands (which are 100 Hz or less in this frequency region). The function is also asymmetric, giving more weight to higher than to lower harmonics, which may reflect a speech-specific constraint related to the fact that changes in actual F1 frequency affect primarily the amplitudes of the higher harmonics in the vicinity of the spectral peak [1]. Listeners thus seem to have tacit knowledge of the physical constraints on the shape of the vocal tract transfer function [29].

Integration of formants. This leads us to the more general question of whether the speech perception system integrates over adjacent formants (or any two peaks in the spectrum) when they are close in frequency but not within a critical band. It has been known for a long time that reasonable approximations to virtually all vowels can be achieved in synthesis with just two formants, and even with a single formant in the case of back vowels [33]. Delattre et al. [33] noted that the approximations were best when the two formants replaced by a single formant were close in frequency (F1 and F2 in high back vowels; F2 and F3 in high front vowels), and that the best single-formant substitute tended to be intermediate in frequency, suggesting that closely adjacent vowel formants form a perceptual composite or average. This idea was later elaborated by the Stockholm research group [18, 19] into the concept of F2', a hypothetical effective formant intermediate in frequency between F2 and F3 (except for /ɪ/, where it falls between F3 and F4). These authors developed a formula for calculating F2' from F1, F2, F3, and F4, which gave good approximations to the results of perceptual matching experiments.

More recently, Chistovich and her collaborators have conducted a number of experiments on the "center of gravity" effect—the demonstrable phonetic equivalence of a single formant to two adjacent formants of varying frequency and/or intensity (see [22] for a review). One important question concerned the critical frequency separation of the two formants beyond which no satisfactory single-formant match could be achieved; it turned out to be about 3.5 Bark, i.e., 3.5 critical bands [23]. This finding has received considerable attention. For example, the 3.5 Bark limit has been related to the separation and boundaries between English vowel categories in acoustic space [166], and it has been used, together with the center of gravity concept, to explain perceived shifts in the height of nasalized vowels, which often have two spectral prominences in the F1 region [4].

It is noteworthy, however, that already Delattre et al. [33] were unable to achieve satisfactory single-formant matches to arbitrary two-formant patterns that did not correspond to familiar vowel categories. This finding, which was replicated by Traunmüller [172, 174] suggests that spectral integration over 3.5 Bark is tied to the perception of phonetic (or phonemic) categories. Specifically, it may reflect the resolution of the auditory long-term memory in which phonetic reference patterns are stored [174]. Indeed, it is an open question whether the 3.5 Bark limit explains the acoustic spacing of vowel categories [166], or whether it is the other way around. A recent study by Schwartz and Esumuler [151], however, provides evidence that the 3.5 Bark limit is not the consequence of phonemic categorization. Their data suggest that there is indeed a higher level of auditory representation that serves phonetic classification and includes wide-band spectral integration. The cause of this integration is unknown at present.

Redintegration of artificially separated spectral components. Ultimately, it must be a higher-level process that decides whether a spectral array constitutes a single event or several. Integration over the whole spectrum is the natural state of affairs, since most natural sounds have complex spectra and could not easily be recognized if integration were not the default operation. Even an unrelated set of pure tones is perceived as a single complex structure when sounded simultaneously, as long as no alternative organizations suggest themselves [63, 77]. Such integration is disrupted by temporal or spatial separation of signal components, however; for example, the "auditory profiles" studied by Green and his coworkers are not well perceived when the sinusoidal components are divided between the two earphone channels [64]. With familiar natural events such as speech, perceptual coherence of spectral components may be centrally guided and hence greater and more resistant to disruption. One possible example of this is the phenomenon called spectral-temporal fusion [27] or duplex perception [84], which has been studied extensively in recent years.

Precursors of this research are found in experiments where the formants of synthetic syllables were separated and presented to opposite ears (e.g., F1 to one ear and F2 and F3 to the other). It was found early on that this presentation gave rise to an intact speech percept, with little or no awareness of separate stimuli in the two ears [14]. Similar fusion of dichotic stimuli into a single perceived sound is observed with complete synthetic syllables in the two ears [122] and even with harmonically related tones [33]. More surprising is the finding that perceptual integration continues to occur even when listeners are aware of separate stimuli in the two ears. Thus, Cutting [27] presented the dichotically separated formants at different fundamental frequencies and observed that subjects still reported the percept corresponding to the combination of the formants. (For similar effects with diotic presentation, see [28]). In what is
How can our brain perform integrative feats in speech perception that exceed the capabilities of the auditory system? One possibility is that there exists a biological specialization in humans, a "speech module," which performs this task [49, 86]. Alternatively, the answer may be mental precompilation, as a consequence of the perceptual learning [75]--an assembled module, as it were. What distinguishes speech perception from the auditory perception of arbitrary tones and noises (but not necessarily from the perception of other ecologically significant auditory events) is that the input can be mapped onto meaningful units of various sizes. The integration of the auditory components relating to each unit represented in the perceiver's long-term memory has taken place long ago during the process of speech and language acquisition; it may be instantiated neurally as a flexible (context-sensitive) system of interconnections [46, 75]. These precompiled units then enable a perceiver to immediately relate a number of functionally independent auditory features to a common phonetic percept. Some interesting (and arduous) attempts to simulate this process of perceptual learning and unit formation in nonspeech auditory perception have been reviewed by Watson and Foyle [183], who stress the importance of central processes in the identification and discrimination of complex stimuli. Experienced Morse code operators exhibit similar skills of "integrating" the acoustic dots and dashes into larger units [17], and so probably perceivers of other meaningful acoustic events in our environment [70, 180], although in none of these instances does the auditory stimulus structure recede as much from awareness as it does in speech perception. From this perspective, speech is unique not so much because it requires specialized perceptual and cognitive functions but because it is structurally different, having originated in the articulatory motor system. Our biological specialization may simply lie in the fact that we can mentally represent a system that is complex.

"Integrated" auditory properties. The ability to integrate over dynamically changing sound patterns has occasionally been attributed to the auditory system. Thus, Stevens and Blumstein [8, 153, 154] hypothesized that the onset spectrum following the release of stop consonants provides invariant acoustic correlates of place of articulation. Since there are often rapid spectral changes immediately following the release, and since a spectrum cannot be computed instantaneously, the hypothetical auditory onset spectrum must derive from an integrative process. Stevens and Blumstein hypothesized that the human auditory system integrates over about 25 ms and thus extracts the acoustic property relevant to place of articulation.

The work of Stevens and Blumstein has come under criticism in recent years. Kewley-Port [73] has argued that, for all we know, the auditory system could be explained on the basis of short-term memory processing. Future experiments involving shorter than 25 ms and presumably delivering information about these changes to phonetic decision mechanisms. Perceptual studies [8, 74]
have suggested that listeners are indeed sensitive to 'spectral changes' immediately following the release of stop consonants. The onset spectra themselves do not appear to be as invariant as was originally claimed [61, 64]. Recent students meanwhile have abandoned the search for invariant properties in onset spectra and have instead gone on to define integrated properties based on the relationship between spectra or intensity measures obtained some interval apart [71, 79, 81]. Even though some of these properties are quite complex, their derivation is still attributed to the auditory system by these researchers. However, since it seems highly implausible that there are general auditory functions which yield so specialized a result, the epithet "auditory" should perhaps be understood as referring merely to the input modality. Clearly, out of the infinity of possibilities, particular relational properties are selected on the basis of phonetic relevance. The integrative computational process thus is specific to speech perception.

Integration of silence and other signal components. Even though it seems unlikely that the auditory system integrates over spectral variation in the speech signal lasting tens of milliseconds, this hypothesis has some measure of plausibility, given the basic continuity of the signal changes. There are many more abrupt changes in the speech signal, however, such as changes in source (from voiced to voiceless, or vice versa), in spectrum (such as /z/ followed by /u/), and in intensity (into and out of closures filled with nasal murmur, voicing, or silence), usually in several of these dimensions simultaneously. It would seem absurd to attribute to the auditory system the capability to integrate across such dramatic signal changes, since the task of auditory perception is to detect changes, not to conceal them. Nevertheless, there is ample evidence from perceptual experiments that listeners can integrate phonetic information across such discontinuities in the signal. Clearly, this integration must be a higher-level function in the service of speech perception.

Perhaps the most striking instance is the perception of silence in speech. (I have in mind brief silent intervals of up to 200 ms duration, not longer pauses.) From an auditory perspective, silence is the absence of energy, a gap, an interruption that separates the signal portions to be perceived. In speech perception, however, silence is bridged by, and participates in, integrative processes. Rather than being the neutral backdrop for the theater of auditory events, silence is informationally equivalent to energy-carrying signal portions. Relative duration of silence has been shown to be a cue for the perception of stop consonant voicing [76, 87, 116], manner [3, 134, 141], and place of articulation [3, 116, 133]. Why does silence function in this way in speech? The answer must be that it is an integral part of the acoustic patterns that a human listener has learned to recognize. Being an acoustic consequence of the oral closure connected with (voiceless) stop consonants, it has become a defining characteristic of that manner class. Lawful variations in its duration as a function of voicing status or place of articulation also have assumed the function of perceptual "cues." A listener's long-term representation of the acoustic pattern corresponding to a stop consonant thus includes the "spectro-temporal" properties of the signals preceding and following the closure as well as the closure itself. (The precise nature of that mental representation, or rather of our description of it, need not concern us here; it suffices to note that listeners behave as if they knew what acoustic pattern to expect.) The silence thus is not really "actively" integrated with the surrounding signal portions; rather, the integration has already taken place during past perceptual learning and is embodied in the perceiver's long-term knowledge of speech patterns to which the input is referred during perception.

Not only is silence integrated (in the sense just discussed) with surrounding signal portions in phonetic perception, but acoustically rather different components of the signal are integrated with each other. Thus, for example, the spectrum of a fricative noise and the adjacent vocalic formant transitions both contribute to perception of a prevocalic fricative consonant [91, 185], the formant transitions in and out of a closure contribute to stop consonant perception [166], etc. Just as articulation distributes acoustic information about individual phonemes over time, perceptual integrative functions collect that information and relate it to internal criteria for linguistic category membership. An especially interesting demonstration of this was provided quite recently by Tomlak et al. [171]. Using a well-known technique [59] for testing listeners' ability to selectively attend to stimulus dimensions, they showed that the "fricative noise" and "vowel" portions of noise-tone analogs to fricative-vowel syllables were processed separately by subjects who perceived the stimuli as nonspeech sounds, but were processed integrally by subjects who had been told that the stimuli represented syllables. These latter subjects were unable to selectively attend to either of the two stimulus portions, even though coarticulatory interactions were not present in the noise-tone stimuli. Listeners in the "speech mode" thus seem to process auditory components of speech in an integrative manner even some of the information to be integrated is not actually there; they are scanning for it, as it were.

Independent aspects of the speech signal that contribute to the same phonemic decision combine according to a simple decision rule, as demonstrated in many experiments by Massaro (e.g., [36, 98]). It is possible to trade various of these cues, changing the physical parameters of one while changing those of another in the opposite direction, without altering the phonemic percept. This phenomenon, often referred to as "phonetic trading relations," has been demonstrated in a large number of studies (reviewed in [129]). Pitts et al. [47] showed that listeners were great difficulty discriminating two phonemically equivalent stimuli created by playing off two cues against each other, and they argued that this reflects the operation of a special phonetic
process that makes audition differences unavailable to perception. Whether the process of phonetic information integration is speech-specific or debatable [138], even though it is agreed that the information being integrated is speech-specific, listeners' difficulty in discriminating phonemically equivalent stimuli is familiar from classical categorical perception research (reviewed in [122]). Experiments on phonetic trading relations that include identification and discrimination tests [6, 47] are generalized categorical perception tasks, in which several physical parameters are varied simultaneously. If each parameter variation by itself is difficult to discriminate except when it cues a category distinction, then joint variations in these parameters will be almost as difficult to discriminate unless a phonemic contrast is perceived. This does not mean, however, that auditory discrimination of such variations is impossible. Appropriate training and use of low-uncertainty discrimination paradigms has been shown to reduce or eliminate categorical perception of single dimensions [20, 128], and it is likely that similar training would enable subjects to discriminate simultaneous variations in several cues, thus demonstrating that their integration does not take place in the auditory system (see also [65]). There is also evidence that certain phonetic trading relations occur only when listeners can make phonemic distinctions, but not within phonemic categories [131].

In summary, the various forms of phonetic cue integration seem to represent, for the most part, speech-specific functions in so far as the articular processes and the corresponding linguistic categories that cause the integration are specific to speech. This idea is embodied in Massaro's "fuzzy logical model" of phonetic decision making [98], which assumes that, for each phonemic category, listeners have internal criteria for the degree of presence of various acoustic features in the speech signal. Diehl and his colleagues have recently argued that many trading relations may have a general auditory basis [39, 109]. While their research may show that some trading relations (especially those within a physical dimension) indeed rest on auditory interactions, this is unlikely to be true of the many trading relations that cut across physical dimensions. Although phonetic perception is certainly immune to auditory interactions, cue integration appears to be mainly a function of speech-specific classification criteria.

Phonetic context effects. Perceivers not only integrate cues directly pertaining to a particular phoneme or complex of articular gestures, but they adapt their perceptual framework to the surrounding phonetic context. Examples of such phonetic context effects are the shift in the /æ/-/i/ category boundary depending on the following vowel [78, 91] and the shift in the /br/-/p/ voice-onset-time category boundary depending on the speaking rate or duration of the surrounding segments [65, 103, 157]. For reviews, see [103, 125, 140]. As in the case of phonetic trading relations, some of these effects may have general auditory processing explanations; thus, for example, the effect of vowel duration on perception of the /æ/-/i/ distinction [104] probably is not speech-specific, as that comparable effect also has been obtained with non-speech stimuli [113]. Many other effects, however, seem to reflect listeners' tacit knowledge of coarticulatory dependencies in speech production. For example, the different /s/-/ʃ// boundaries in the context of rounded and unrounded vowels may be related to the occurrence of anticipatory lip-rounding during the constriction phase in utterances such as "soup" but not in "saw." In a series of experiments using cross-spliced fricative noises and vowels, Whalen [186, 188] has shown that even when the fricative noise itself is quite unambiguous, subjects' reaction time in a fricative identification task is influenced by the following vocalic context, being slower when the fricative noise spectrum is not exactly what would be expected in that context (cf. the study by Tomlak et al. [171] reviewed above). In an unpublished series of experiments, Repp [123] demonstrated an effect of coarticulation perception, which consisted of slower reaction times to decide that the two consonants are the same in the stimulus pair /æba/-/aba/ than in the pairs /æba/-/aba/, even though the pre-closure (VC) portions of these synthetic CV stimuli were identical in both cases. That is, even though subjects could have made their decisions by hearing /æba/ in the second member of a stimulus pair, they somehow had to take the CV portions of the stimulus into account and then were slowed down by the inequality of the vowels. All these studies show that perceivers integrate all information that possibly could bear on phonetic decisions, and this integration often seems obligatory in nature. It requires special instructions, special (non-phonetic) tasks, and usually some amount of training to disengage phonetic integration mechanisms in the laboratory [6, 127, 128, 136].

Cross-modal integration. In natural speech communication, humans make use not only of auditory but also of visual information, if available. Audiovisual integration at the level of the perceiver has been a research topic of considerable interest since the discovery by McGurk and MacDonald [101] that subjects presented with conflicting auditory and visual speech stimuli report that they "hear" what they see. Their findings have been replicated and extended in a number of studies [89, 97, 157] and others. Massaro [96, 97] has shown that a general rule of information integration based on the degree to which signal features match expected feature values can explain audiovisual integration, auditory cue integration, as well as many other forms of perceptual integration outside the domain of speech. This suggests that we may be dealing with a general function following basic laws of decision theory. Liberman and his collaborators [85, 141], on the other hand, have argued that integration of speech cues, within or across modalities, occurs because they represent the multiple-distributed consequences of articular act or gestures. Some internal reference to processes of speech production is thus implied, as in the "motor theory" of speech perception [86]. However, this
account is complementary rather than antithetic to Massaro's model. It is a theory of why integration occurs, whereas Massaro is concerned with how integration works. The phonemes of a language are articulatory events which have characteristic acoustic and optic consequences, and perceivers presumably have tacit knowledge incorporating both of these aspects. If a portion of the speech input satisfies certain auditory and visual criteria for phonemic category membership (as in Massaro's model) this also implies that the gestures characterizing a particular phoneme have been recovered (as in the motor theory). Whether the sensory or the articulatory aspect is stressed in a particular theory is largely a matter of philosophy and perhaps of economy. A complete theory must include both.

Audiovisual integration at the more global level of word, sentence, and discourse comprehension has, of course, been of interest for a long time in connection with hearing impairment and communication in noisy environments. Research on this topic has received a boost in recent years with the advent of modern signal processing technology and of cochlear implants. (See [158] for a review.) The information provided by residual hearing or by electrical stimulation of the auditory nerves supplements that obtained from lipreading to yield enhanced comprehension. In many respects, these two sources of information are complementary, with the auditory channel providing information that is difficult to see, and vice versa. What is of special interest in the present context is that audiovisual comprehension performance often seems to exceed what might be expected from a mere combination of independent sources of information. Thus, Rosen et al. [143] demonstrated that speech intelligibility is improved substantially when lipreading in hearing subjects is supplemented with the audible fundamental frequency contour, or even just with a constant buzz representing the occurrence of voicing. (See also [9, 62].) Since this auditory component by itself provides virtually no information about phonetic structure, it must be the temporal relationships between the auditory and visual channels that contribute to intelligibility [100]. Thus audiovisual speech perception is often more than the sum of its parts; in terms of Massaro's [96] model, the separate sources are integrated before central evaluation. The close integration of inputs from the two modalities is witnessed by anecdotal reports that voicing-triggered buzz accompanying lipreading may assume phonetic qualities [159].

The theoretical issues raised by audiovisual integration have been discussed thoroughly by Summerfield [159]. He, too, concludes that auditory and visual cues to linguistic structure are integrated before any categorical decisions are made. There are four ways of conceptualizing how this integration occurs: (1) The two channels make independent contributions to linguistic decisions, but temporal relationships provide a third source of information. (2) The visual information is translated into an auditory metric of vocal tract area functions. (3) The auditory information is translated into a visual metric of articulatory kinematics. (4) Both are translated into an abstract representation of dynamic control parameters of articulation. This last-mentioned approach [15, 72] may ultimately provide the most economic description of speech information in both modalities, and thus may yield the most appropriate vocabulary in which to describe intermodal integration.

Higher-level integration. Human listeners not only integrate auditory and visual information about a speaker's articulations, but they also bring phonotactic, lexical, syntactic, semantic, and pragmatic expectations to bear on their linguistic decisions, provided the auditory and/or visual input is sufficiently ambiguous to give room to effects of such expectations. Some well-known demonstrations of effects in this category are the "phoneme restoration" phenomenon discovered by Warren [181] and studied more recently by Samuel [145], in which lexical expectations fill in missing acoustic information, as it were; the lexical bias effects of gating [98] and replicated by Fox [57], which causes a relative shift in the category boundaries on acoustic word-word (e.g., DASH-TASH versus DASH-TASK) continua in favor of word percepts; and the "fluent restorations" in rapid shadowing of semantically anomalous passages [98]. These phenomena, and a host of related ones often referred to as "top-down" effects, may be considered general forms of cognitive information integration in speech perception. Indeed, Massaro [96] has argued that the rules by which such higher-level information is integrated with the "bottom-up" information delivered by the senses are the same by which acoustic (and optic) speech cues are integrated. Others argue that top-down influences should be strictly separated from bottom-up processes—that they represent general cognitive functions that operate outside the autonomous speech module [49, 96]. According to this second view, integration of bottom-up cues to phoneme identity is a fundamentally different process from the integration of bottom-up and top-down information. My own view in this matter is that speech perception at every level requires domain-specific knowledge stored in a perceiver's long-term memory. The processes by which this knowledge is brought to bear upon the sensory input are part of our metaphorical representation of brain function and thus are bound to be general [138]. In the absence of a radically different vocabulary in which to characterize the processes within a module (though such a vocabulary will perhaps emerge from the study of articulatory dynamics and coordination), the postulate of a speech module harks back to the "black box" of behaviorism. It is quite likely, of course, that phonetic perception is modular in the sense that integration of phonetic cues precedes, and is not directly influenced by, higher-level factors. This issue can be addressed empirically [49, 58, 145, 165]. My point here is that integration, whether it occurs inside a module or outside it, is conceptually the same thing: a many-to-one mapping. Indeed, Massaro's (e.g., [96]) extensive research suggests that the rules of integration are independent of

PL 2.2.9
modularity.

SEGREGATION

The preceding section has illustrated the pervasiveness of integrative processes in speech perception. Much of perceptual and cognitive processing is convergent, with multiple sources of information contributing to single decisions, be they explicit or implicit. Nevertheless, we also need hypothetical mechanisms to prevent all information from converging onto every decision "node." Even though a perceiver's internal criteria for linguistic category membership will automatically reject irrelevant information, information that does not belong is nevertheless often potentially relevant. Thus, in the oft-cited cocktail party situation, the voices of several speakers must be kept apart to avoid semantic and phonetic confusions. Various environmental sounds could simulate phonetic events and need to be segregated from the true speech stream. In the speech signal itself, information pertaining to speaker identity, emotion, room acoustics, etc., needs to be distinguished from the phonetic structure, and the overlapping consequences of segmental articulation need to be sorted out. These segregative processes have an important complementary role to play in speech perception: They ensure that integration is restricted to those pieces of information that belong together. Logically, segregation precedes integration, even though functionally they may be just the two sides of one coin. The more physically similar and intertwined the aspects to be segregated are, the more remarkable the segregative process will seem to us.

Temporal and spatial segregation

Without any doubt, there are several factors that enable perceivers to distinguish different sound sources or events, regardless of whether they are speech or not. One of these is temporal separation. Sounds occurring a long time apart will usually not be considered as belonging to the same event, although they may come from the same source. In speech, a few seconds are usually enough to segregate phrases or utterances, and a few hundreds of milliseconds of separation usually prevent integration of acoustic cues into a single phonemic decision. One demonstration of this fact may be found in studies of the distinction between single and geminate stop consonants. In a classic experiment, Pickert and Decker [111] asked English-speaking subjects to distinguish between utterances such as "topic" and "top pick," varying only the duration of the silent /p/ closure. Between 150 and 300 msec were needed to obtain judgments of two /p/s (and two words) rather than just one; the precise duration depended on the overall speaking rate. (See also [108, 124, 125].)

If two different stop consonants follow each other, as in the nonsense word /aba/da/, about 100 ms of silent closure are needed to prevent integration of the two sets of formant transitions into a single stop consonant percept [43, 124]. Dorman et al. [43] cued the perception of /p/ in "split" solely by inserting a silent interval between an /s/ noise and the syllable "lit" (a percept that may be said to be a pure temporal integration illusion), and subsequently investigated how much silence was needed before subjects reported hearing an "a" followed by "lit." This duration turned out to be as long as 600 msec. A subsequent replication [136] obtained a shorter but still surprisingly long interval of 300-400 msec. To cite a final example, Tillmann et al. [170] investigated how much temporal offset of one noisy and acoustically presented syllables was needed to destroy the audiovisual integration effect discovered by McGurk and MacDonald [101]. It turned out to be 250-300 msec. These various situations have little in common, which explains the different results. The precise duration of the critical interval for segregation surely depends on many factors and does not reflect any general limits of temporal integration. Rather, within the auditory modality it may be related to the closure durations normally encountered in natural speech [111, 130].

Temporal asynchrony is a helpful cue in distinguishing speech from other environmental sounds. This was elegantly demonstrated in a series of studies by Darwin [29, 32, 119], who investigated under what conditions a pure tone added to one of the (pure tone) harmonics of a synthetic vowel was treated by listeners as part of the vowel spectrum or as a separate nonvowel event. Darwin showed that, when the tone coincided with the vowel, it affected the perceived vowel quality. However, when the onset of the tone preceded that of the vowel or, to a lesser extent, when its offset lagged behind that of the vowel, listeners excluded it from the phonetic information. Similar principles of segregation or "auditory stream formation" have been demonstrated in the perception of nonvowel sounds [12].

Another factor that may cause segregation is spatial separation. In real life, the separation of several simultaneous voices or of speech from background noises is often possible because they are perceived as coming from different locations. In the laboratory, presentation over the two channels of earphones has been used to induce segregation. One interesting case in which this form of spatial separation does not seem to prevent integration is split-formant or duplex perception, discussed above. Note, however, that in duplex perception one component of the speech signal (the "chirp") is segregated and heard as a separate auditory event; the paradox is that this event is still, at the same time, integrated with the speech in the other ear. (See [111].) There are many other instances, however, particularly those in which there is no temporal overlap between the two signals, where spatial separation is sufficient to disrupt perceptual integration. For example, informal observations suggest that, if the artificial "split" created by concatenating "s" and "lit" with some intervening silence is divided between the two ears, so that "s" occurs in one ear and "lit" in the other, listeners report hearing that is, there is no /p/ percept any more. Similarly, when
nasal-consonant-vowel syllables such as /ni/ or /ni/ are divided between the two ears, so that the nasal murmur occurs in one and the vocalic portion containing the formant transitions in the other, listeners have great difficulty identifying the consonant, even in the case in which it is more prominent than if the two components were presented by themselves [137]. Of course, it is always possible to integrate independent sources of information at a cognitive level. These two examples illustrate the role of spatial separation as a segregating factor. Unfortunately, in real life both temporal and spatial separation are often unavailable as segregating agents, and listeners need additional means of sorting out the incoming stream of auditory information.

Spectral segregation

When irrelevant (speech or non-speech) sounds are superimposed on speech, listeners have basically two means of segregation at their disposal: Segregation according to local spectral disparity, and according to spectro-temporal (and, in part, speech-specific) criteria of pattern coherence. There are, of course, many sounds in the environment, including those produced by most musical instruments, that are sufficiently different from speech to be perceived immediately as different sources. Local spectral segregation is not always effective, however, and for good reason: First, some non-speech events (e.g., the pops of bottles or the hisses of steam valves) are spectrally similar to speech sounds and thus are difficult to separate from them locally. Second, and more importantly, speech itself is composed of acoustic segments of diverse spectral composition, and it would be counterproductive if listeners were prone to segregate them, because these segments are more often not map onto the same linguistic unit. Indeed, perceptual segregation of spectrally dissimilar natural speech components can usually be demonstrated only under special conditions, which rarely occur outside the laboratory. Thus, Cole and Snow [24] reported it, and Frost et al. [25] reported that listeners sometimes reported two streams of events: a train of fricative noises, and a train of vowels, especially when the vocalic formant transitions were removed. A similar phenomenon was obtained with the repeated syllable /ka/ by Diehl et al. [40] who then used their findings to explain the different effects of /spa/ or /sk/ stimuli as adaptors (or precursors) in selective adaptation and pairwise contrast paradigms [147, 148]. The selective adaptation task requires cyclic repetition of a single stimulus, the adaptor, and thus may produce "streaming" of signal components, so that /spa/ is heard as /sa/ and /ba/, with the phonological status of the stop consonant altered. Repp [128] was able to induce listeners through some training to segregate a fricative noise from a following vowel and "hear out" the spectral quality of the noise. Even the individual formants of vowels may segregate under certain conditions. Following earlier studies showing that it was difficult to perceive the correct temporal order of four rapidly cycling steady-state vowels [169, 182], Dorman et al. [42] found that this was because in such artificial sequences individual formants tend to group together and form separate auditory streams. There are anecdotal reports of phoneticians being able to "hear out" individual formants of vowels (e.g., [66, 150]), but this ability has remained rare. Still, these reports underscore the fact that spectrally diverse components of the speech signal are potentially segregable; fortunately, however, they are perceptually integrated under normal circumstances.

When two different speech streams co-occur, differences in fundamental frequency, intonation pattern, or voice quality may provide cues for separation, in addition to higher-level factors such as syntactic and semantic continuity. Effects of this kind have been found in classical work on selective attention (reviewed in [175]). More recently, Broks and Hooteboom [16] obtained a beneficial effect of differences in fundamental frequency and intonation on the identification of meaningless sentences presented against a background of a read story. In the much more artificial situation of two simultaneous steady-state vowels, Scheffers [149] and Zwicker [191] found an improvement in recognition performance when a fundamental frequency difference was introduced. Since the magnitude of the difference beyond one semitone did not seem to play a role, the function of F0 differences in this case seems to be to prevent fusion of the two sounds. Similar, though small, effects of F0 on identification scores have also been obtained in dichotic listening studies using synthetic syllables [67, 121, 167] or vowels [191].

The potential of fundamental frequency (F0) and voice quality cues to segregate successive portions of speech has also been demonstrated in the laboratory. The mechanisms studied here must be involved in separating different speakers from each other. Several relevant studies have used stimuli in which perception of a stop consonant rested on the duration of a silent closure interval. Dorman et al. [35] found that in the speech on each side of the silence was produced by different voices, the silence lost its perceptual effectiveness; that is, listeners did not integrate across it. On the other hand, it has been shown [118, 177] that silence retains its effectiveness between syllables produced by male and female voices if the general articulatory and intonational pattern is continuous across the two speakers (achieved by cross-splicing two intact utterances). When the second syllable was spliced onto a first syllable originally produced in utterance-final position, however, the phonetic effect of the silence was disrupted. Thus it seems that dynamic spectro-temporal information about articulatory continuity can override differences in F0 or voice quality. A disruptive effect of discontinuities in intonation on stop consonant perception has also been reported [117], but such an effect was absent in a recent study [135] in which a constant fricative noise preceded the listeners. The above studies suggest that the breaks in the F0 contour are effective only when voiced signal portions immediately abut the silent closure interval.
Segregation of linguistic and paralinguistic information

So far I have discussed segregation of two kinds: One separates speech from other, irrelevant sounds (including competing speech streams), and the other dissociates consecutive parts of the same speech stream—a laboratory-induced phenomenon to be avoided in natural speech communication. These segregative processes are "literal" in the sense that they result in the perception of separate sound sources. Segregative processes are also essential, however, when listening to a single speech source, and for two reasons. First, the speech signal conveys information about phonetic composition, speaker characteristics (vocal tract size, sex, age, identity, emotion), and room or transmission characteristics (reverberation, distortion, filtering). A listener needs to separate these three kinds of information, which Chistovich [22] has termed "phonetic quality," "personal quality," and "transmission quality," respectively (see also [175]). Second, the acoustic information for adjacent phonemes is overlapped and merged, a phenomenon commonly referred to as coarticulation or "encoding." If phonemic units are to be recovered, the information pertaining to one phoneme needs to be separated from that for another—or so it seems. Both these kinds of segregation are not literal in the sense that they make a speech stream disintegrate perceptually; rather, they separate different aspects of a coherent perceptual event by relating these aspects to different conceptual categories or dimensions represented in long-term memory. They operate on the information in the signal, not on the signal itself.

Of the various types of information the first kind, that of separating vocal tract size information from phonetic information, has received the most attention under the heading of speaker normalization. An explicit solution to this problem is that of vital importance to automatic speech recognition as well as to any theory of speech perception. In fact, the focus has been so exclusively on the speaker-independent recovery of phonetic information that it is sometimes forgotten that listeners extract several kinds of information in parallel. Rather than "normalizing" their internal representation of the speech wave and discarding information in the process, they presumably use all available kinds of information to mutual advantage.

Studies of speaker normalization have, for the most part, been concerned with vowels rather than consonants, and with acoustic analysis and automatic recognition rather than with human perception. Older normalization algorithms often required knowledge of a speaker's whole vowel space or average formant frequencies (see [41]), whereas more recent work has focused on perceptually more relevant transformations based on parameters that are immediately available in the incoming speech signal (e.g., [165, 166, 173]). There have been relatively few perceptual studies on this topic; the general assumption has been that it is sufficient to obtain acoustic properties that are relatively speaker-invariant and also plausible in the light of what is known about the auditory system. Demonstrations of "perceptual normalization" usually show a performance decrement in a listening situation where speaker characteristics are varied repeatedly and unpredictably, compared to one in which the speaker remains constant [80, 161, 178]. Although emphasis is sometimes placed on the perceptual "advantage" resulting from effective normalization, the negative consequences of presenting contrived and misleading stimuli are perhaps the more salient outcome of this research (which is why no means unique in this respect).

Analogous experiments have been conducted on normalization in the temporal domain—this is, on the perceptual separation of speaking rate from phonetic length (reviewed in [103]). An especially interesting question arises in research on tone languages, where the listener must segment lexical tones from the overall intonation contour [25] and from speaker-dependent information [82]. In that connection, it is noteworthy that there is mounting evidence (reviewed in [144]) that tone and intonation perception (and production) are controlled by opposite hemispheres of the brain. At least some forms of linguistic/paralinguistic segregation may thus have a basis in neuropsychological compartmentalization. A general explanation can be drawn from research on perceptual normalization that the auditory parameters underlying phonetic classification are not absolute quantities but relationships in the spectral and/or temporal domain, computed over a relatively restricted temporal interval, whereas properties signalling speaker sex or identity, emotion, speaking rate, etc., accumulate over longer stretches of speech and/or are based on more nearly absolute quantities.

Segregation of intertwined linguistic information

The emphasis on linguistic information in the vast majority of speech perception studies makes it difficult to find good examples of research on perceptual segregation of linguistic and (rather than from) nonlinguistic information. Examples of segregation of equivalent information are easier to find when only linguistic information is involved. This leads me to the final topic, one that has been of enormous significance in speech perception research—the problem of segmentation, that is, the perceptual separation of the overlapped acoustic correlates of adjacent phonemic units, particularly of vowels and consonants.

One traditional view of the listener's task has been that it is one of phoneme (or feature) extraction, including "compensation" for contextual influences on a segment's acoustic correlates (see [54]). Numerous studies have shown that listeners perceive segments as if they knew all the contextual modifications their acoustic representations undergo [129, 140]. Thus, for example, a fricative noise ambiguous between /s/ and /f/ in isolation is perceived as /s/ when followed by /w/ but as /f/ when followed by /a/.
Current efforts at Haskins Laboratories to model articulation as a sequence of overlapping segmental gestures (e.g., [15, 72]) may ultimately provide ways of recovering these gestures from the acoustic signal and thus provide a machine implementation of Fowler's vector-analytic concept. A promising mathematical technique for achieving this goal, based on 'principal' components analysis of vocal tract area function parameters, has been proposed by Atal [2] and is currently being explored [92, 93]. The recovery of articulatory parameters from the acoustic signal remains a central problem in speech research because phonemes and alphabets surely represent an articulatory, not an acoustic classification. However, while a solution of this problem would bring us a great step forward, processes of integration and segregation would still be needed to translate the articulatory "score" into a sequence of discrete segments.

**Speech Perception Without Integration and Segregation?**

In the introduction, I discussed four basic assumptions: the separation of the physical and mental worlds, the existence of physical units, the existence of mental units, and the existence of processes relating the two kinds of units. Can a theory of speech perception do without them? The assumptions are not independent, of course: If the physical and mental worlds are distinct, they must receive different descriptions; to be easily communicable in the scientific world, these descriptions must be in terms of discrete concepts or units; and this results in certain functions or relationships between the two descriptive domains. If the physical and mental worlds were isomorphic, there would be no need for a theory of perception. If one or the other description were without units (more likely an error of omission than a deliberate theoretical choice), then perception would seem either entirely integrative or entirely segregative—not an attractive state of affairs. Denial of functions, however abstract, linking the two domains would merely impoverish perceptual theory. Certainly we need these functions in theories of auditory processing and organization. As to the perception of phonetic information, however, an alternative approach has been proposed.

This approach, stated most eloquently by Studdert-Kennedy [155] and Fowler [54], follows the "direct-realist" perspective of ecological psychology [61, 179]. Although it affirms the existence of linguistic units as articulatory events, it essentially abandons the distinction between the physical and mental domains. The segmental structure of speech (as characterized by the linguist or phonetician) is assumed to be ever-present on its way from the speaker to the listener's brain. There is assumed to be a direct isomorphism between physical and mental descriptions of speech events (such as phonemes), though it is acknowledged that the appropriate physical and motor-dynamic descriptions have not been fully worked out. Thus this school of thought rejects the idea that the input is divided into
parts that need to be integrated or segregated by the listener; rather, the input units are taken to be identical with the perceptual units—that is, they are already integrated or segregated with respect to more primitive acoustic or auditory units. The deliberate strategy of this philosophy is to eliminate classical problems in perceptual research (such as segmentation and invariance) by redefining and redescribing physical events. Rather than being attributed to the perceiver's brain, the burdens of information integration and segregation thus fall upon the investigator trying to find an "integral" description of "separate" speech events. However, this effort is equivalent to that of finding a principled explanation of perceptual integration and segregation: If we can show that certain pieces of input are always integrated, we might as well call them integral and treat them as a single piece in our descriptions—if we only had names for them. Behind the rhetoric and the different terminologies of mentalistic and realistic approaches lies a common goal: to arrive at the most economic characterization of linguistic structure in all its physical incarnations. Clearly, even speech research propelled by a mentalistic philosophy (still predominant in the field) must strive to minimize the work attributed to a speaker-listener's mind. But will we be able to relieve it of its entire burden to integrate and segregate? What we take away (in theory) is likely to re-emerge as logical conjunctions, disjunctions, and relational terms in our physical characterization of speech events. As long as we scientists communicate in conventional language, integration and segregation at some stage in our theories will be difficult to avoid.

ACKNOWLEDGMENTS

Preparation of this paper was supported by NICHD Grant HD-01994 to Haskins Laboratories. I am grateful to Ignatius Hattingly, Susan Nittouer, and Michael Studdert-Kennedy for helpful comments on an earlier version.

REFERENCES


and unstressed vowels. Journal of Speech and Hearing Research, 46, 127-139.


[78a] University of Tokyo.


on stop closure duration as a cue for voicing and place. *Journal of Phonetics, 7*, 45-56.


