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 multiple physical 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 permits 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 assures 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 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, tne 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 typical 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 following vowel. It seems likely that the observed 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 /s/-stop 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 a topic of great interest to psychoacousticians and auditory physiologists, who have identified at least three different time constants of adaptation 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 nerves 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 [138]. 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 pointed 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 phonetic 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 [138]. 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.

Spectral integration

Most speech sounds have complex spectra determined by the resonance frequencies of the vocal tract. 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 methods 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., [110, 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 Celmer and Bienvenue [21] 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, converted the manipulated spectra back into sound, and presented them to groups of 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 in 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 besides 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 that 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 limens 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 /I/ and /e/ can be affected by the intensity of as many as five harmonics between 250 and 750 Hz, spaced 125 Hz 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 /i/, 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 Escudier [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 [37]. 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 now called the duplex perception paradigm, Rand [120] presented the formant transitions distinguishing two synthetic consonant-vowel syllables (such as /da/ and /ga/) to one ear and the remainder common to the two syllables (the "base") to the opposite ear. In this situation, listeners continue to report one or the other syllable depending on which formant transition is presented, even though that transition is also heard simultaneously as a lateralized nonspeech "chirp." The intact syllable (not the base) is heard in the ear receiving the base. Thus, subjectively at least, auditory fusion takes place despite the auditory segregation of the chirp--a paradoxical situation. This fusion continues to operate when the two signal components are presented at different fundamental frequencies [27] or with slight temporal offsets [139]. A very similar phenomenon can be produced diotically by making the critical formant transition audible through temporal offset [139], amplification [187], or different fundamental frequencies (informal observations). None of these manipulations, within certain limits, destroys the fused speech percept.

One interpretation of these findings [86] is that a specialized speech "module" is responsible for the perceptual integration and apparent fusion, whereas the general auditory system is responsible for the separate chirp percept. Bregman [11], on the other hand, has proposed that the paradoxical co-occurrence of fusion and nonfusion arises from conflicting cues for integration and segregation in the general process of "auditory scene analysis." He and other students of auditory organization have stressed the relative independence of What and Where decisions in auditory perception [13, 28, 38, 184]. It seems that auditory components that have been segregated can nevertheless be recombined in the perception and classification of familiar sound structures. That this recombination in the duplex perception paradigm is genuinely perceptual and not cognitive is indicated not only by the subjective impression of an intact syllable but by the fact that the components (chirp and base) presented by themselves generally do not suggest the "correct" phonetic percept [142].

Integration of phonetic information

Speech consists of a sequence of diverse sound segments which, as everyone knows, do not correspond directly to linguistic units. Changes in spectral structure are often very rapid and lead to great spectral heterogeneity over time. Equally striking is the alternation of qualitatively different sound types (periodic vs. aperiodic, as well as silence). Nevertheless, listeners perceive a coherent event, and thus believe speech to be a coherent stream of sounds. Since there is absolutely no reason to assume that very disparate sound structures are automatically integrated by the auditory system, the subjective impression of auditory continuity must be due to higher-level articulatory and linguistic properties of cohesiveness that capture the listener's attention--a kind of categorical perception (see [132]).

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 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 do 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 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 tracks spectral changes over time intervals shorter than 25 ms and presumably delivers 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 [81, 164]. Blumstein and her 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 acoustic 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 [168], 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 Tomiak 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]). Fitch et al. [47] showed that listeners have 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 auditory differences unavailable to perception. Whether the process of phonetic information integration is speech-specific is 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 [132]). 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 [6]). 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 articulatory 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 for the many trading relations that cut across physical dimensions. Although phonetic perception is certainly not 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 articulatory gestures, but they adapt their perceptual criteria to the surrounding phonetic context. Examples of such phonetic context effects are the shift in the /s/-/ʃ/ category boundary depending on the following vowel [78, 91] and the shift in the /b/-/p/ voice-onset-time category boundary depending on the speaking rate or duration of the surrounding segments [65, 103, 157]. For reviews, see [103, 129, 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 /ba/-/wa/ distinction [104] probably is not speech-specific, as a comparable effect has also been obtained with nonspeech 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 liprounding during the constriction phase in utterances such as "soup" but not in "sap." 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 Tomiak et al. [171] reviewed above). In an unpublished series of experiments, Repp [123] demonstrated an effect he dubbed "coperception," which consisted of slower reaction times to decide that the two consonants are the same in the stimulus pair /aba/-/aba/ than in the pair /aba/-/abi/, even though the pre-closure (VC) portions of these synthetic VCV stimuli were identical in both cases. That is, even though subjects could have made their decisions after hearing /ab/ in the second member of a stimulus pair, they somehow had to take the CV portions of the stimuli 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 (nonphonetic) 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 phoneme perception has been a research topic of considerable interest since the discovery by McGurk and MacDonald [101] that subjects presented with certain 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 articulatory acts 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 nerve 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 ultimatelv 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 effect reported by Ganong [58] and replicated by Fox [57], which causes a relative shift in the category boundaries on acoustic word-nonword (e.g., DASH-TASH versus DASK-TASK) continua in favor of word percepts; and the "fluent restorations" in rapid shadowing of semantically anomalous passages [94]. 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, 86]. 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 metaphoric 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 information integration are independent of 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 often-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, Pickett 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 /abda/, 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 "s" 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 optically 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], 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 nonspeech 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 nonspeech 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 [11].) 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, this is exactly what listeners report hearing; that is, there is no /p/ percept any more. Similarly, when

nasal-consonant-vowel syllables such as /mi/ 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, or in any case do not perform better 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 nonspeech) 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 nonspeech 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 more often than 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 Scott [24] rapidly iterated fricative-vowel syllables and found 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 /ska/ by Diehl et al. [40] who then used their findings to explain the different effects of /spa/ or /ska/ 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 /s/ 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 various findings underline 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 [176]). More recently, Brokx and Nooteboom [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. [43] found that when 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 critical silence, suggesting 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 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 in parallel, and largely over the same time-frequency channels, 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 segregation of 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 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., [163, 166, 173]). There have been relatively few perceptual studies on this topic; the general assumption has been that it is sufficient to define 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 rapidly 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 by no means unique in this respect).

Analogous experiments have been conducted on normalization in the temporal domain--that 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 segregate lexical tones from the overall intonation contour [25] and from speaker-dependent variation in F0 [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 neurophysiological compartmentalization. A general conclusion to be drawn from research on perceptual normalization is 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 /ʃ/ in isolation is perceived as /s/ when followed by /u/ but as /ʃ/ when followed by /a/ [91]. One way of describing this finding is that listeners "know" that anticipatory liprounding for /u/ may lower the spectrum of a preceding fricative noise, so they adopt a different criterion for the /s/-/ʃ/ distinction in that context. This view, which emphasizes the role of tacit phonetic knowledge in speech perception, has recently been elaborated by several authors (e.g., [48, 138]). The perceptual accomplishment seems more integrative than segregative from that perspective.

An alternative view, having an equally long history, has a recent proponent in Fowler [53, 54, 55] who has likened the separation of overlapping segmental information to mathematical vector analysis. According to her theory, listeners literally subtract or factor out the influences of one segment on another, so that invariant segments are "heard." Fowler conceives of phonetic segments as articulatory events, not as abstract mental categories (see the exchange on coarticulation between Fowler [50, 52] and Hammarberg [68]), though listeners are assumed to be able to judge their "sound" [53]. Several experiments [51, 53, 55] were intended to demonstrate this. They showed that subjects judge acoustically different representations of a segment to be more similar than acoustically identical ones if the former occur in their original contexts while the latter have been spliced into inappropriate contexts. However, since only the former match what listeners expect to hear in a given context, these results are also compatible with an alternative account based on tacit knowledge of contextual effects in speech production [129, 138]. That is, rather than having access to the sound of segments [53], listeners may have made their judgments on the basis of the discrepancy of the input from context-sensitive mental norms or prototypes.

Other recent experiments in a similar vein have addressed the separation of nasality and vowel height information in nasalized vowels. Kawasaki [71a] showed that English listeners judge vowels in /m_m/ environment as increasingly nasal as the surrounding nasal murmurs are attenuated; that is, when the nasal consonants are intact, the vowel nasality is attributed to (coarticulation with) the nasal consonants, as it were, and is "factored out" from the vowel percept. Building on this result, Beddor et al. [5] first established that there are different category boundaries on synthesized /bɛd/-/bæd/ and /bɛ̃d/-/bæ̃d/ continua. English listeners apparently interpret some of the spectral consequences of nasalization as a change in vowel height. However, when an appropriate "conditioning environment" was added in the form of a postvocalic /n/, the category boundary on the resulting /bɛ̃nd/-/bæ̃nd/ continuum was identical with that on the /bɛd/-/bæd/ continuum, as if listeners attributed the vowel nasality to (coarticulation with) the nasal consonant and "factored it out" in Fowler's sense. The result is equally compatible, however, with a theory that postulates context-sensitive vowel (or syllable) prototypes. Indeed, it may be difficult to come up with any decisive experiments. Mentalism and realism may simply represent different metatheoretical perspectives.

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 the same 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's 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 Mattingly, Susan Nittrouer, and Michael Studdert-Kennedy for helpful comments on an earlier version.

REFERENCES

[1] Assmann, P. F., & Nearey, T. M. (1987). Perception of front vowels: The role of harmonics in the first formant region. *Journal of the Acoustical Society of America*, *81*, 520-534.

[2] Atal, B. S. (1983). Efficient coding of LPC parameters by temporal decomposition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (Boston)*, 81-84.

[3] Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 536-563.

[4] Beddor, P. S. (1984). Formant integration and the perception of nasal vowel height. *Haskins Laboratories Status Report on Speech Research*, *SR-77/78*, 107-120.

[5] Beddor, P. S., Krakow, R. A., & Goldstein, L. M. (1986). Perceptual constraints and phonological change: a study of nasal vowel height. *Phonology Yearbook*, *3*, 197-218.

[6] Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, *29*, 191-211.

[7] Bladon, A. (1982). Arguments against formants in the auditory representation of speech. In R. Carlson & B. Granström (Eds.), *The representation of speech in the peripheral auditory system* (pp. 95-102). Amsterdam: Elsevier Biomedical Press.

[8] Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, *67*, 648-662.

[9] Breeuwer, M., & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America*, *79*, 481-499.

[10] Bregman, A. S. (1978). The formation of auditory streams. In J. Requin (Ed.), *Attention and performance VII* (pp. 63-76). Hillsdale, NJ: Erlbaum.

[11] Bregman, A. S. (1987). The meaning of duplex perception: Sounds as transparent objects. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception*. The Hague: Martinus Nijhoff.

[12] Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, *32*, 19-31.

[13] Bregman, A. S., & Steiger, H. (1980). Auditory streaming and vertical localization: Interdependence of "what" and "where" decisions in audition. *Perception & Psychophysics*, *28*, 539-546.

[14] Broadbent, D. E., & Ladefoged, P. (1952). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, *29*, 708-710.

[15] Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219-254.

[16] Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*, 23-36.

[17] Bryan, W. L., & Harter, N. (1899). Studies in the physiology and psychology of the telegraphic language. The acquisition of a hierarchy of habits. *Psychological Review*, *6*, 345-375.

[18] Carlson, R., Fant, G., & Granström, B. (1975). Two-formant models, pitch, and vowel perception. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 55-82). London: Academic Press.

[19] Carlson, R., Granström, B., & Fant, G. (1970). Some studies concerning perception of isolated vowels. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *2-3*, 19-35 (Stockholm: Royal Technical University).

[20] Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, *62*, 961-970.

[21] Celmer, R. D., & Bienvenue, G. (1987). Critical bands in the perception of speech by normal and sensorineural hearing loss listeners. in M. E. H. Schouten (Ed.), *The psychophysics of speech perception*. The Hague: Martinus Nijhoff.

[22] Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. *Journal of the Acoustical Society of America*, *77*, 789-805.

[23] Chistovich, L. A., & Lublinskaja, V. V. (1979). The center of gravity effect in vowel spectra and critical distance between the formants. *Hearing Research*, *1*, 185-195.

[24] Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology*, *27*, 441-449.

[25] Connell, B. A., Hogan, J. T., & Rozsypal, A. J. (1983). Experimental evidence of interaction between tone and intonation in Mandarin Chinese. *Journal of Phonetics*, *11*, 337-351.

[26] Cowan, N. (in press). Auditory sensory storage in relation to the growth of sensation and acoustic information extraction. *Journal of Experimental Psychology: Human Perception and Performance*.

[27] Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, *83*, 114-140.

[28] Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, *33A*, 185-207.

[29] Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America*, *76*, 1636-1647.

[30] Darwin, C. J., & Gardner, R. B. (1985). Which harmonics contribute to the estimation of first formant frequency? *Speech Communication*, *4*, 231-235.

[31] Darwin, C. J., & Seton, J. (1983). Perceptual cues to the onset of voiced excitation in aspirated initial stops. *Journal of the Acoustical Society of America*, *74*, 1126-1135.

[32] Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *The Quarterly Journal of Experimental Psychology*, *36A*, 193-208.

[33] Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, *8*, 195-210.

[34] Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America*, *68*, 843-857.

[35] Delgutte, B., & Kiang, N. Y. S. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. *Journal of the Acoustical Society of America*, *75*, 897-907.

[36] Derr, M. A., & Massaro, D. W. (1980). The contribution of vowel duration, Fo contour, and frication duration as cues to the /juz/-/jus/ distinction. *Perception & Psychophysics*, *27*, 51-59.

[37] Deutsch, D. (1978). Lateralization by frequency for repeating sequences of dichotic 400- and 800-Hz tones. *Journal of the Acoustical Society of America*, *63*, 184-186.

[38] Deutsch, D., & Roll, P. (1976). Separate "what" and "where" decision mechanisms in processing a dichotic tonal sequence. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 23-29.

[39] Diehl, R. (1987). Auditory constraints on the perception of speech. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception*. The Hague: Martinus Nijhoff.

[40] Diehl, R. L., Kluender, K. R., & Parker, E. M. (1985). Are selective adaptation and contrast effects really distinct? *Journal of Experimental Psychology: Human Perception and Performance*, *11*, 209-220.

[41] Disner, S. F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, *67*, 253-261.

[42] Dorman, M. F., Cutting, J. E., & Raphael, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 121-129.

[43] Dorman, M. F., Raphael, L. J., & Liberman, A. M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, *65*, 1518-1532.

[44] Dreschler, W. A., & Plomp, R. (1980). Relation between psychophysical data and speech perception for hearing-impaired subjects. I. *Journal of the Acoustical Society of America*, *68*, 1608-1615.

[45] Eggermont, J. J. (1985). Peripheral auditory adaptation and fatigue: A model oriented review. *Hearing Research*, *18*, 57-71.

[46] Elman, J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice. Vol. 10* (pp. 337-373). New York: Academic Press.

[47] Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, *27*, 343-350.

[48] Flege, J. E. (in press). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), *Human communication and its disorders, Vol. 1*. Norwood, NJ: Ablex.

[49] Fodor, J. A. 1983). *The modularity of mind*. Cambridge, MA: MIT Press.

[50] Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*, 113-133.

[51] Fowler, C. A. (1981). Production and perception of coarticulation among stressed

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

[52] Fowler, C. A. (1983). Realism and unrealism: a reply. Journal of Phonetics, 11, 303-322.

[53] Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. Perception & Psychophysics, 36, 359-368.

[54] Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics, 14, 3-28.

[55] Fowler, C. A., & Smith, M. R. (1986). Speech perception as "vector analysis": An approach to the problems of invariance and segmentation. In J. S. Perkell & D. H. Klatt (Eds.), Invariance and variability in speech processes (pp. 123-135). Hillsdale, NJ: Erlbaum.

[56] Fox, R. A. (1983). Perceptual structure of monophthongs and diphthongs in English. Language and Speech, 26, 21-60.

[57] Fox, R. A. (1984). Effect of lexical status on phonetic categorization. Journal of Experimental Psychology: Human Perception and Performance, 10, 526-540.

[58] Ganong, W. F., III. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.

[59] Garner, W. R. (1974). The processing of information and structure. Potomac, MD: Erlbaum.

[60] Gibson, J. J. (1966). The senses considered as perceptual systems. Boston: Houghton Mifflin.

[61] Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

[62] Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. Journal of the Acoustical Society of America, 77, 671-677.

[63] Green, D. M. (1983). Profile analysis. A different view of auditory intensity discrimination. American Psychologist, 38, 133-142.

[64] Green, D. M., & Kidd, G., Jr. (1983). Further studies of auditory profile analysis. Journal of the Acoustical Society of America, 73, 1260-1265.

[65] Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. Perception & Psychophysics, 38, 269-276.

[66] Halle, M., Hughes, G. W., & Radley, J.-P. A. (1957). Acoustic properties of stop consonants. Journal of the Acoustical Society of America, 29, 107-116.

[67] Halwes, T. G. (1969). Effects of dichotic fusion on the perception of speech. Haskins Laboratories Status Report on Speech Research (Supplement).

[68] Hammarberg, R. (1982). On redefining coarticulation. Journal of Phonetics, 10, 123-137.

[69] Jaeger, J. J. (1980). Testing the psychological reality of phonemes. Language and Speech, 23, 233-253.

[70] Jenkins, J. J. (1985). Acoustic information for objects, places, and events. In W. H. Warren & R. E. Shaw (Eds.), Persistence and change. Proceedings of the First International Conference on Event Perception (pp. 115-138). Hillsdale, NJ: Erlbaum.

[71] Jongman, A., Blumstein, S. E., & Lahiri, A. (1985). Acoustic properties for dental and alveolar stop consonants: a cross-language study. Journal of Phonetics, 13, 235-251.

[71a] Kawasaki, H. (1986). Phonetic explanation of phonological universals: The case of distinctive vowel nasalization. In J. J. Ohala & Jaeger, J. J. (Eds.), Experimental phonology (pp.81-104). New York: Academic Press.

[72] Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. Journal of Phonetics, 14, 29-59.

[73] Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 73, 322-335.

[74] Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. Journal of the Acoustical Society of America, 73, 1779-1793.

[75] Klatt, D. H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. Journal of Phonetics, 7, 279-312.

[76] Kohler, K. J. (1979). Dimensions in the perception of fortis and lenis plosives. Phonetica, 36, 332-343.

[77] Kubovy, M. (1981). Concurrent-pitch segregation and the theory of indispensable attributes. In M. Kubovy & J. R. Pomerantz (Eds.), Perceptual organization (pp. 55-98). Hillsdale, NJ: Erlbaum.

[78] Kunisaki, O., & Fujisaki, H. (1977). On the influence of context upon perception of voiceless fricative consonants. Annual Bulletin of the Research Institute for Logopedics and Phoniatrics, 11, 85-91 (University of Tokyo).

[79] Kurowski, K., & Blumstein, S. E. (in press). Acoustic properties for place of articulation in nasal consonants. Journal of the Acoustical Society of America.

[80] Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America, 29, 98-104.

[81] Lahiri, A., Gewirth, L., & Blumstein, S. E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. Journal of the Acoustical Society of America, 76, 391-404.

[82] Leather, J. (1983). Speaker normalization in perception of lexical tone. Journal of Phonetics, 11, 373-382.

[83] Lehiste, I. (1972). The units of speech perception. In J. H. Gilbert (Ed.), Speech and cortical functioning (pp. 187-236). New York: Academic Press.

[84] Liberman, A. M. (1979). Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), Proceedings of the Ninth International Congress of Phonetic Sciences (pp. 468-473). Copenhagen: Institute of Phonetics, University of Copenhagen.

[85] Liberman, A. M. (1982). On finding that speech is special. American Psychologist, 37, 148-167.

[86] Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.

[87] Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. Language, 33, 42-49.

[88] Lisker, L. (1978). On buzzing the English /b/. Haskins Laboratories Status Report on Speech Research, SR-55/56, 181-188.

[89] MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. Perception & Psychophysics, 24, 253-257.

[90] Mann, V. A. (1986). Phonological awareness: The role of reading experience. Cognition, 24, 65-92.

[91] Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. Perception & Psychophysics, 28, 213-228.

[92] Marcus, S. M., & Atal, B. S. (1986). Decoding the spech code--applications of temporal decomposition. Journal of the Acoustical Society of America, 80 (Suppl. 1), S17.

[93] Marcus, S. M., & Van Lieshout, R. A. J. M. (1984). Temporal decomposition of speech. IPO Annual Progress Report, 25-31 (Nijmegen, The Netherlands).

[94] Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. Speech Communication, 4, 55-73.

[95] Massaro, D. W. (1975). Preperceptual images, processing time, and perceptual units in speech perception. In D. W. Massaro (Ed.), Understanding language. An information-processing analysis of speech perception, reading, and psycholinguistics (pp. 125-150). New York: Academic Press.

[96] Massaro, D. W. (in press). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.

[97] Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9, 753-771.

[98] Massaro, D. W., & Oden, G. C. (1980). Evaluation and integration of acoustic features in speech perception. Journal of the Acoustical Society of America, 67, 996-1013.

[99] Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.), Language by ear and by eye. The relationships between speech and reading (pp. 133-148). Cambridge, MA: MIT Press.

[100] McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. Journal of the Acoustical Society of America, 77, 678-685.

[101] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

[102] McNeill, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. Journal of Verbal Learning and Verbal Behavior, 12, 419-430.

[103] Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, NJ: Erlbaum.

[104] Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics, 25, 457-465.

[105] Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidth and excitation patterns. Journal of the Acoustical Society of America, 74, 750-753.

[106] Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331.

[107] Mushnikov, V. N., & Chistovich, L. A. (1973). Experimental testing of the band hypothesis of vowel perception. Soviet Physics-Acoustics, 19, 250-254.

[108] Obrecht, D. H. (1965). Three experiments in the perception of geminate consonants in Arabic. Language and Speech, 8, 31-41.

[109] Parker, E. M., Diehl, R. L., & Kluender, K. R. (1986). Trading relations in speech and nonspeech. Perception & Psychophysics, 39, 129-142.

[110] Peters, R. W., Moore, B. C. J., & Glasberg, B. R. (1983). Pitch of components of complex tones. Journal of the Acoustical Society of America, 73, 924-929.

[111] Pickett, J. M., & Decker, L. R. (1960). Time factors in perception of a double consonant. Language and Speech, 3, 11-17.

[112] Pisoni, D. B. (1987). Auditory perception of complex sounds: Some comparisons of speech vs. non-speech signals. In W. A. Yost and C. S. Watson (Eds.), Auditory processing of complex sounds (pp. 247-256). Hillsdale, NJ: Erlbaum.

[113] Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. Perception & Psychophysics, 34, 314-322.

[114] Plomp, R. (1964). The ear as a frequency analyzer. Journal of the Acoustical Society of America, 36, 1628-1636.

[115] Pols, L. C. W., van der Kamp, L. J. Th., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. Journal of the Acoustical Society of America, 46, 458-467.

[116] Port, R. F. (1979). The influence of tempo

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

[117] Price, P. J., & Levitt, A. G. (1983). The relative roles of syntax and prosody in the perception of the /š/-/č/ distinction. Language and Speech, 26, 291-304.

[118] Rakerd, B., Dechovitz, D. R., & Verbrugge, R. R. (1982). An effect of sentence finality on the phonetic significance of silence. Language and Speech, 25, 267-282.

[119] Rakerd, B., & Verbrugge, R. R. (1985). Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels. Journal of the Acoustical Society of America, 77, 296-301.

[120] Rand, T. C. (1974). Dichotic release from masking for speech. Journal of the Acoustical Society of America, 55, 678-680.

[121] Repp, B. H. (1976a). Effects of fundamental frequency contrast on discrimination and identification of dichotic CV syllables at various temporal delays. Memory & Cognition, 4, 75-90.

[122] Repp, B. H. (1976b). Identification of dichotic fusions. Journal of the Acoustical Society of America, 60, 456-469.

[123] Repp, B. H. (1978a). "Coperception": Influence of vocalic context on same-different judgments about intervocalic stop consonants. Unpublished manuscript (available from the author).

[124] Repp, B. H. (1978b). Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. Perception & Psychophysics, 24, 471-485.

[125] Repp, B. H. (1979a). Influence of vocalic environment on perception of silence in speech. Haskins Laboratories Status Report on Speech Research, SR-57, 267-290.

[126] Repp, B. H. (1979b). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. Language and Speech, 22, 173-189.

[127] Repp, B. H. (1980). Accessing phonetic information during perceptual integration of temporally distributed cues. Journal of Phonetics, 8, 185-194.

[128] Repp, B. H. (1981). Two strategies in fricative discrimination. Perception & Psychophysics, 30, 217-227.

[129] Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.

[130] Repp, B. H. (1983a). Bidirectional contrast effects in the perception of VC-CV sequences. Perception & Psychophysics, 33, 147-155.

[131] Repp, B. H. (1983b). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. Speech Communication, 2, 341-362.

[132] Repp, B. H. (1984a). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), Speech and language: Advances in basic research and practice. Vol. 10 (pp. 243-335). New York: Academic Press.

[133] Repp. B. H. (1984b). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. Language and Speech, 27, 245-254.

[134] Repp, B. H. (1984c). The role of release bursts in the perception of [s]-stop clusters. Journal of the Acoustical Society of America, 75, 1219-1230.

[135] Repp, B. H. (1985a). Can linguistic boundaries change the effectiveness of silence as a phonetic cue? Journal of Phonetics, 13, 421-431.

[136] Repp, B. H. (1985b). Perceptual coherence of speech: Stability of silence-cued stop consonants. Journal of Experimental Psychology: Human Perception and Performance, 11, 799-813.

[137] Repp, B. H. (1987a). On the possible role of auditory short-term adaptation in perception of the prevocalic [m]-[n] contrast. Manuscript submitted for publication.

[138] Repp, B. H. (1987b). The role of psychophysics in understanding speech perception. In M.E.H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[139] Repp, B. H., & Bentin, S. (1984). Parameters of spectral/temporal fusion in speech perception. Perception & Psychophysics, 36, 523-530.

[140] Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. N. Harnad (Ed.), Categorical perception. New York: Cambridge University Press.

[141] Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 4, 621-637.

[142] Repp, B. H., Milburn, C., & Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. Perception & Psychophysics, 33, 333-337.

[143] Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. Nature, 291, 150-152.

[144] Ross, E. D., Edmondson, J. A., & Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. Journal of Phonetics, 14, 283-302.

[145] Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. Journal of Experimental Psychology: General, 110, 474-494.

[146] Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. Journal of Verbal Learning and Verbal Behavior, 9, 295-302.

[147] Sawusch, J. R., & Jusczyk, P. (1981). Adaptation and contrast in the perception of voicing. Journal of Experimental Psychology: Human Perception and Performance, 7, 408-421.

[148] Sawusch, J. R., & Nusbaum, H. C. (1983). Auditory and phonetic processes in place perception for stops. Perception & Psychophysics, 34, 560-568.

[149] Scheffers, M. T. M. (1983). Sifting vowels. Auditory pitch analysis and sound segregation. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

[150] Schubert, E. D. (1982). On hearing your own performance. In V. L. Lawrence (Ed.), Transcripts of the Eleventh Symposium Care of the Professional Voice (pp. 161-185). New York: The Voice Foundation.

[151] Schwartz, J. L., & Escudier, P. (1987). Does the human auditory system include large scale spectral integration? In M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[152] Stelmachowicz, P. G., Jesteadt, W., Gorga, M. P., & Mott, J. (1985). Speech perception ability and psychophysical tuning curves in hearing-impaired listeners. Journal of the Acoustical Society of America, 77, 620-627.

[153] Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 64, 1358-1368.

[154] Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (eds.), Perspectives in the study of speech (pp. 1-38). Hillsdale, NJ: Erlbaum.

[155] Studdert-Kennedy, M. (1985). Perceiving phonetic events. In W. H. Warren, Jr., & R. E. Shaw (Eds.), Persistence and change: Proceedings of the First International Conference on Event Perception (pp. 139-156). Hillsdale, NJ: Erlbaum.

[156] Summerfield, Q. (1979). Use of visual information for phonetic perception. Phonetica, 36, 314-331.

[157] Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: Human Perception and Performance, 7, 1074-1095.

[158] Summerfield, Q. (1983). Audio-visual speech perception, lipreading, and artificial stimulation. In M. E. Lutman & M. P. Haggard (Eds.), Hearing science and hearing disorders (pp. 131-182). London: Academic Press.

[159] Summerfield, Q. (in press). Preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), Hearing by eye. Hillsdale, NJ: Erlbaum.

[160] Summerfield, Q., & Assmann, P. (1987). Auditory enhancement and speech perception. In M.E.H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[161] Summerfield, Q., & Haggard, M. P. (1975). Vocal tract normalization as demonstrated by reaction times. In G. Fant & M. A. A. Tatham (Eds.), Auditory analysis and perception of speech (pp. 115-142). London: Academic Press.

[162] Summerfield, Q., Haggard, M., Foster, J., & Gray. S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. Perception & Psychophysics, 35, 203-213.

[163] Suomi, K. (1984). On talker and phoneme information conveyed by vowels: A whole spectrum approach to the normalization problem. Speech Communication, 3, 199-209.

[164] Suomi, K. (1985). The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables. Journal of Phonetics, 13, 267-285.

[165] Swinney, D. A. (1982). The structure and time-course of information interaction during speech comprehension: Lexical segmentation, access, and interpretation. In J. Mehler, E. C. T. Walker, & M. Garrett (Eds.), Perspectives on mental representation. Experimental and theoretical studies of cognitive processes and capacities (pp. 151-167). Hillsdale, NJ: Erlbaum.

[166] Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. Journal of the Acoustical Society of America, 79, 1086-1100.

[167] Tartter, V. C., & Blumstein, S. E. (1981). The effects of pitch and spectral differences on phonetic fusion in dichotic listening. Journal of Phonetics, 9, 251-259.

[168] Tartter, V. C., Kat, D., Samuel, A. G., & Repp, B. H. (1983). Perception of intervocalic stop consonants: The contributions of closure duration and formant transitions. Journal of the Acoustical Society of America, 74, 715-725.

[169] Thomas, I. B., Hill, P. B., Carrol, F. S., & Garcia, D. (1970). Temporal order in the perception of vowels. Journal of the Acoustical Society of America, 48, 1010-1013.

[170] Tillmann, H. G., Pompino-Marschall, B., & Porzig, U. (1984). Zum Einfluss visuell dargebotener Sprechbewegungen auf die Wahrnehmung der akustisch kodierten Artikulation. Forschungsbericht des Instituts fur Phonetik und Sprachliche Kommunikation der Universität München, 19, 318-336.

[171] Tomiak, G. R., Mullennix, J. W., & Sawusch, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. Journal of the Acoustical Society of America, 81, 755-764.

[172] Traunmüller, H. (1982). Perception of timbre: Evidence for spectral resolution bandwidth different from critical band? In R. Carlson & B. Granström (Eds.), The representation of speech in the peripheral auditory system (pp. 103-108). Amsterdam: Elsevier Biomedical Press.

[173] Traunmüller, H. (1984a). Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. Speech Communication, 3, 49-61.

[174] Traunmüller, H. (1984b). Die spektrale Auflösung bei der Wahrnehmung der Klangfarbe von Vokalen. Acustica, 54,237-246.

[175] Traunmüller, H. (1987). Some aspects of the sound of speech sounds. In M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[176] Treisman, A. M. (1969). Strategies and models of selective attention. Psychological Review, 76, 282-299.

[177] Verbrugge, R. R., & Rakerd, B. (1986). Evidence for talker-independent information for vowels. *Language and Speech*, *29*, 39-57.

[178] Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, *60*, 198-212.

[179] Warren, W. H., Jr., & Shaw, R. E. (1985). *Persistence and change: Proceedings of the First International Conference on Event Perception*. Hillsdale, NJ: Erlbaum.

[180] Warren, W. H., Jr., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 704-712.

[181] Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392-393.

[182] Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American*, *233*, 30-36.

[183] Watson, C. S., & Foyle, D. C. (1985). Central factors in the discrimination and identification of complex sounds. *Journal of the Acoustical Society of America*, *78*, 375-380.

[184] Weintraub, M. (1987). Sound separation and auditory perceptual organization. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception*. The Hague: Martinus Nijhoff.

[185] Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America*, *69*, 275-282.

[186] Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, *35*, 49-64.

[187] Whalen, D. H., & Liberman, A. M. (1987). *Speech perception takes precedence over nonspeech perception*. Manuscript submitted for publication.

[188] Whalen, D. H., & Samuel, A. G. (1985). Phonetic information is integrated across intervening nonlinguistic sounds. *Perception & Psychophysics*, *37*, 579-587.

[189] Wright, H. N. (1964). Temporal summation and backward masking. *Journal of the Acoustical Society of America*, *36*, 927-932.

[190] Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, *68*, 1523-1525.

[191] Zwicker, U. T. (1984). Auditory recognition of diotic and dichotic vowel pairs. *Speech Communication*, *3*, 265-277.

[192] Zwislocki, J. J. (1969). Temporal summation of loudness: An analysis. *Journal of the Acoustical Society of America*, *46*, 431-440.