CHAPTER 7

PSYCHOLOGICAL ASPECTS OF SPEECH PERCEPTION

IMPLICATIONS FOR RESEARCH AND THEORY

DOMINIC W. MASSARO

I. INTRODUCTION

In a psychology laboratory, a six-week-old infant in a baby chair has a pacifier in his mouth. As he sucks on the pacifier, the experimenter presents the sound /ba/ contingent on the infant's sucking. Given this feedback, the baby increases his sucking rate but soon becomes bored and sucks less. Now, however, the /ba/ sound is changed to /da/ and the infant increases his sucking rate again. The infant must have noticed the sound change from /ba/ to /da/.

A petite 3-year-old girl sits at a table of toy figures. She is told a short story and she must describe the story with the toy figures. To the child, she is playing a game, but to the psychologist and psycholinguist, she is displaying a remarkable ability to perceive and understand language. As an example, the child is told *The fence the horse kicks*. The child takes the horse and has it kick the fence.

A sophomore in college is studying Mandarin Chinese and learns that the syllable /ma/ has four possible meanings depending on its pitch. At first it is difficult to determine which is which. With practice, categorizing these variants becomes second nature.

A senior citizen is watching a talk show on television and is having trouble hearing the participants. He remembers he doesn't have his glasses, retrieves them, and puts them on. Surprisingly, seeing the show better allows him to hear the show better.

Spoken language is an inherent dimension of humanity from the crib to the grave. A worthwhile goal is to describe how we perceive and understand speech. The answer might take several different forms. It might be argued, for example, that a speech "organ" has evolved to carry out this function. A speech organ is necessary because speech is a highly specialized domain that necessarily requires a specialized processing system. In contrast, it might be hypothesized that understanding speech is just one domain of many that require discrimination, categorization, and understanding. We also discriminate, categorize, and interact with everyday objects and events. Why should speech be any different? Of course, other solutions between these two extreme alternatives are possible. We begin with this issue of whether speech perception is specialized.

II. IS SPEECH PERCEPTION SPECIALIZED?

A central issue in speech perception and psycholinguistics has been the so-called modularity of speech and language. Noam Chomsky (1980) has described language as an independent organ (or module), analogous to other organs such as our digestive system. This organ follows an independent course of development in the first years of life and allows the child to achieve a language competence that cannot be explained in traditional learning terms. Thus, a mental organ responsible for the human language faculty is viewed as responsible for our language competence. This organ matures and develops with experience, but the mature system does not simply mirror this experience. The language user inherits rule systems of highly specific structure. This innate knowledge allows us to acquire the rules of the language, which cannot be induced from normal language experience because (advocates argue) of the paucity of the language input. The data of language experience are so limited that no process of induction, abstraction, generalization, analogy, or association could account for our language competence. Somehow, the universal grammar given by our biological endowment allows the child to learn to use language appropriately without learning many of the formal intricacies of language. Other linguists, however, have documented that our language input is not as sparse as the nativists would have us believe (Sampson, 1989).

Although speech has not had an advocate as charismatic and influential as Chomsky, a similar description has been given for speech perception. Some theorists now assume that a speech module is responsible for speech perception (Liberman & Mattingly, 1989). The justification for this module has been analogous to the one for language more generally. Performance is not easily accounted for in terms of the language input. In speech, it is claimed that the acoustic signal is deficient and that typical pattern recognition schemes could not work. Put another way, it is argued that speech exceeds our auditory information processing capabilities. In terms of the modularity view, our speech perception system is linked with our speech production system—and our speech perception is somehow mediated by our speech production. For these theorists (and for the direct-realist perspective of Fowler, 1986), the objects of speech perception are articulatory events or gestures. These gestures are the primitives that the mechanisms of speech production translate into actual articulatory movements, and they are also the primitives that the specialized mechanisms of speech perception recover from the signal.

A. Evolutionary History of Speech

If speech perception were a highly unique and modular function, we would expect it to have a relatively long evolutionary history. That is, a unique process would be expected to have a unique evolutionary history. Speech as we know it, however, appears to be relatively recent in our evolutionary history. Before the artificial speech of the last few decades, speech could be produced only by biological entities. Our speech is critically dependent on the characteristics of our respiratory system and vocal tract. Thus, it is of interest to determine the evolutionary history of the biological system used for speech.

Lieberman (1991) provides a systematic analysis of the evolution of human speech. Using fossil records, he argues that speech as we know it was not possible just over 100,000 years ago. As can be seen in Figure 1, Neanderthal had a larynx positioned high, close to the entrance to the nasal cavity. The tongue was also positioned almost entirely in the mouth, as opposed to being half in the pharynx as it is in our mouths. Using computer modeling, it was discovered that the Neanderthal vocal tract would not form the configurations that are necessary to produce [i], [u], and [a] vowels. Its speech would also be necessarily nasalized (since the nasal cavity could not be blocked off), which would create a less discriminable signal because of the superimposed nasal sounds. The fossils of *Homo sapiens* of around 100,000 years ago appear to have skulls that contain a modern supralaryngeal vocal tract. From this, Lieberman concludes that language as we know it, in terms of having the supralaryngeal vocal tract to support it, is about 100 to 125 thousand years old. Given that speech is so recent in our evolutionary history, it seems unlikely that a unique skill has evolved to perceive speech and understand language. Independent of the issue of the uniqueness of speech perception, we cannot expect an evolutionary description of speech perception to be sufficient. As psycholinguists, we must also be concerned with proximal causes and influences, not just the distal influences described by evolutionary theory.

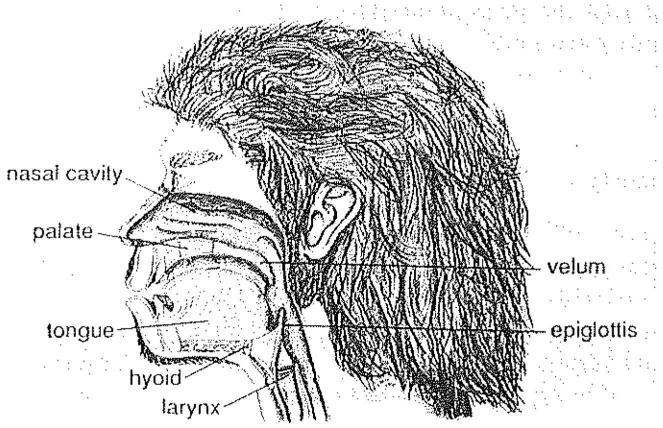

FIG. 1 The reconstructed airway of the La Chapelle-aux-Saints Neanderthal fossil (after Lieberman, 1991).

It appears that the astonishing brain growth of our ancestors occurred before the development of speech and language. This means that it is unlikely that specific brain structures evolved to enable speech production and speech perception. Our gift of language, thought, and culture must be due to exploiting the plasticity of the brain for communication. In addition, spoken language became the higher level programming language of human computer systems.

B. Lack of Invariance between Signal and Percept

One of the original arguments for the specialized nature of speech perception involved the uncertain relationship between properties of the speech signal and a given perceptual (read phonemic) category. It was stressed that, contrary to other domains of pattern recognition, one could not delineate a set of acoustic properties that uniquely defined a phoneme. The classic example involved the dramatic changes in the second-formant transitions of stop consonants in different vowel environments. Although there has been a small but continuous defense of the idea that phonemes do have invariant properties (Cole & Scott, 1974; Blumstein & Stevens, 1981), most investigators have accepted the tremendous variability of phonemes in different contexts (e.g., Wickelgren, 1969).

The argument for lack of invariance has always been articulated in a narrow sense and holds very little force with close scrutiny. First, there has been no questioning of the psychological reality of the phonetic units described in theoretical linguistics—even though the concept is even debatable in that domain. Second, it has been accepted without question that phonemes are perceived. However, a subject report of the syllable /ba/ does not necessarily imply that phoneme perception mediated this behavior. Third, research in many domains has shown that a strict correspondence between signal and perceived pattern is the exception rather than the rule in human pattern recognition. As we will see in more detail, the use of multiple sources of ambiguous information better characterizes pattern recognition in most domains, including speech. Fourth, enlarging the perceptual units of analysis to syllables of V, VC, and CV size greatly overcomes much of the invariance problem (where V is a vowel and C is a consonant or consonant cluster).

C. Nonlinearity of Segmental Units

The nonlinearity of phoneme segments in speech has also been used in the same argument for specialty as the lack of invariance. As dramatized by Hockett (1955), phonemes appear to be eggs run through a wringer so that it is difficult to discern at what point one egg ends and the next begins. This blurring and the contextual variance of phonemes is due to the articulation of one phoneme being influenced by the articulation of preceding and following phonemes. This co-articulation arises from physical necessity—even if the talker intended to articulate discrete phonemes, which can also be challenged. Once again, a strict linearity is not necessary for a nonspecialized pattern recognition process.

Perhaps the most comparable situation is handwriting, in which the visible characteristics of a letter are influenced by its adjacent neighbors.

D. Rate of Speech Processing

One traditional argument for a special processor for speech is that the transmission rate of the speech signal appears to exceed our perceptual capacity. Phonetic segments—the minimum linguistic units of speech that are approximated by the letters of the alphabet—occur at a rate of between 10 and 20 per second. Supposedly, humans cannot identify nonspeech signals at even half this rate. There are several counterarguments to the rate argument, however. First, speech has a fast rate only when phonetic segments are taken as the psychologically real unit of analysis. Although linguists have described the linguistic reality of these phonetic segments, there is no evidence that these segments are psychologically functional in speech perception. If syllables (V, CV, and VC) are assumed to be functional perceptual units in speech perception, then the rate of presentation of these signals is well within the range of our information processing capability.

A second problem with counting the rate of phonetic segments as an index of speech rate is that a word could be recognized without necessarily recognizing the phonetic segments that make it up. Some evidence for this idea has been obtained in the processing of nonspeech sounds (Warren, 1982). If a sequence of arbitrarily selected sounds is presented, listeners have trouble identifying the order of the elements that make up the sequence unless each sound is presented for 0.25 s or so. On the other hand, one sequence could be discriminated from another when the sounds are much shorter—in the range of 5–100 ms. Warren, Bashford, and Gardner (1990) found that subjects could discriminate different sequences of repeated vowels without identifying their order. The emergence of unique words with different words for different sequences was responsible for the discrimination. A conjunction of different sounds has the consequences of a unique percept emerging which can be informative for the perceptual system. Two different sequences of identical components are discriminated from one another because one arrangement sounds different from the other. One might sound "bubbly" and the other "shrill." Subjects can even learn to label these sequences as wholes if appropriate feedback is given. This research is consistent with research on language acquisition. Peters (1983) observed that the child acquires speech segments in terms of a variety of sizes: syllables, words, or even phrases. For example, the child learns to identify the word *through* not in terms of a sequence of three phonetic segments but as a CV syllable of a particular quality.

A final problem with the argument that the rate of speech processing is larger than other forms of auditory information processing is the positive contribution of context (see Section II,I). Our ability to process speech at a fast rate holds only for familiar speech. Even linguists have great difficulty transcribing a language that they do not know. Knowing a language allows us to perceive speech on the basis of a deficient signal or with little processing time. For example, we can perceive the first /s/ in the word *legislatures* even when the relevant segment has been replaced with a noise or a tone (Warren, 1970).

Similarly, we can perceive speech of a language we know when it is speeded up at two or three times its normal rate.

E. Speech Perception by Nonhumans

There is another source of evidence against the hypothesis that speech perception is carried out by a specialized module unique to humans. If speech perception were special and mediated in any way by speech production, then discrimination and recognition of fundamental speech categories should be impossible for nonhumans. However, some nonhuman animals can discriminate fundamental speech segments. Chinchillas (a small rodent with auditory capabilities close to humans) can discriminate fundamental distinctions such as the auditory difference signaling the difference between /ba/ and /pa/. More recently and more impressively, Kluender, Diehl, and Killeen (1987) have shown that quail can learn to discriminate the stop consonant /d/ from the stops /b/ and /g/ (occurring in different vowel environments). Given these results, it appears that there is information in the auditory speech signal that can be processed using normal perceptual processes.

F. Categorical Perception

One of the classic research findings used to support speech as a specialized modular process was categorical perception. Categorical perception occurs when changes along some dimension of the speech signal are not perceived continuously but in a discrete manner. Listeners are supposedly limited in their ability to discriminate differences between different sounds belonging to the same phoneme category. The sounds within a category are only identified absolutely, and discrimination is possible for only those sounds that can be identified as belonging to different categories. For example, small changes can be made in the consonant–vowel syllable /be/ (*bay*) to transform it in small steps into the syllable /de/ (*day*). These syllables are used in identification and discrimination tasks. The results seemed to indicate that subjects can discriminate the syllables only to the extent they recognize them as different categories. These results were contrasted with other forms of perception in which we can discriminate many more signals that we can categorize. Hence, speech perception seemed to qualify as a special type of performance.

There are severe weaknesses in the previous evidence for categorical perception. The results have been interpreted as showing categorical perception because discrimination performance was reasonably predicted by identification performance. It turns out that this relation between identification and discrimination provides no support for categorical perception, for two reasons. First, categorical perception usually provides an inadequate description of the relation between identification and discrimination, and has not been shown to provide a better description than continuous perception. Second, other explanations of the results are possible, and these explanations do not require any special processes for speech (Massaro, 1987, 1989b).

In fact, there is now an abundance of evidence that perceivers are very good at perceiving differences within a speech category. For example, subjects are very good at indicating the degree to which a speech stimulus represents a given speech category. In addition, reaction times of identification judgments illustrate that members within a speech category vary in ambiguity or the degree to which they represent the category (Massaro, 1987). These results indicate that subjects can discriminate differences within a speech category, and they are not limited to just categorical information. Decision processes can transform continuous sensory information into results usually taken to reflect categorical perception. A finding of categorical partitioning of a set of stimuli in no way implies that these stimuli were perceived categorically.

G. The Demise of Categorical Perception

Categorical perception is a belief that will not die or fade away easily. Many textbooks and tutorial articles also state that speech is perceived categorically (J. R. Anderson, 1990; Eimas, 1985; Flavell, 1985; Miller, 1981). However, I have argued in too many places that previous results and more recent studies are better described in terms of continuous perception—a relatively continuous relationship between changes in a stimulus and changes in perception (Massaro, 1987).

There are severe weaknesses in previous evidence for categorical perception. One approach—the traditional one used throughout the almost three decades of research on categorical perception—concerns the relation between IDENTIFICATION and DISCRIMINATION. In the typical experiment, a set of speech stimuli along a speech continuum between two alternatives is synthesized. Subjects identify each of the stimuli as one of the two alternatives. Subjects are also asked to discriminate among these same stimuli. The results have been interpreted as showing categorical perception because discrimination performance was reasonably predicted by identification performance (Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). It turns out that this relation between identification and discrimination provides no support for categorical perception, for two reasons. First, the categorical model usually provides an inadequate description of the relation between identification and discrimination, and has not been shown to provide a better description than continuous models. Second, even if the results provided unequivocal support for the categorical model, explanations other than categorical perception are possible (Massaro, 1987; Massaro & Oden, 1980).

We saw that evidence against categorical perception comes from a direct experimental comparison between categorical and continuous models of perception. Subjects asked to classify speech events independently varying along two dimensions produce identification results consistent with the assumption of continuous information along each of the two dimensions. A model based on categorical information along each dimension gives a very poor description of the identification judgments. In other research, we asked subjects to make repeated ratings of how well a stimulus represents a given category (Massaro & Cohen, 1983). The distribution of the rating judgments to a given stimulus is better described by a continuous model than a categorical one. The best conclusion is to reject all reference to categorical perception of speech and to concentrate instead on the structures and processes responsible for categorizing the world of speech.

Most readers will remain unconvinced so long as no satisfying explanation is given for the sharp category boundaries found in speech perception research. However, it is only natural that continuous perception should lead to sharp category boundaries along a stimulus continuum. Given a stimulus continuum from *A* to *not A* that is perceived continuously, goodness(*A*) is an index of the degree to which the information represents the category *A*. The left panel of Figure 2 shows goodness(*A*) as a linear function of Variable *A*.

An optimal decision rule in a discrete judgment task would set the criterion value at 0.5 and classify the pattern as *A* for any value greater than 0.5. Otherwise, the pattern is classified as *not A*. Given this decision rule, the probability of an *A* response would take the form of the step-function shown in the right panel of Figure 2. That is, with a fixed criterion value and no variability, the decision operation changes the continuous linear function given by the perceptual operation into a step function. Although based on continuous perception, this function is identical to the idealized form of categorical perception in a speech identification task. It follows that a step function for identification is not evidence for categorical perception because it can also occur with continuous information.

If there is noise in the mapping from stimulus to identification, a given level of Variable *A* cannot be expected to produce the same identification judgment on each presentation. It is reasonable to assume that a given level of Variable *A* produces a normally distributed range of goodness(*A*) values with a mean directly related to the level of Variable *A* and a variance equal across all levels of Variable *A*. If this is the case, noise will influence the identification judgment for the levels of Variable *A* near the criterion value more than it will influence

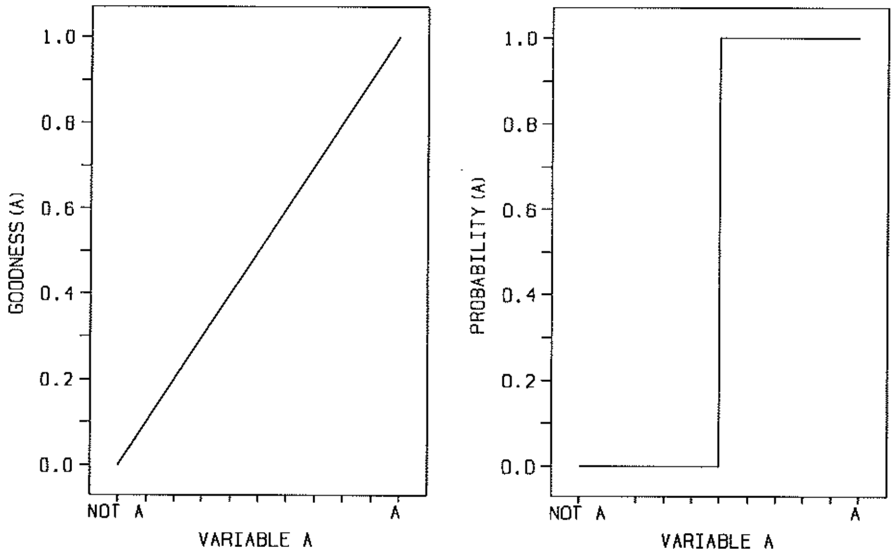

FIG. 2 *Left:* The degree to which a stimulus represents the category *A*, called goodness(*A*), as a function of the level along a stimulus continuum between *not A* and *A*. *Right:* The probability of an *A* response, probability(*A*), as a function of the stimulus continuum if the subject maintains a decision criterion at a particular value of goodness(*A*) and responds *A* if and only if the goodness(*A*) exceeds the decision criterion.

the levels away from the criterion value. Figure 3 illustrates the expected outcome for identification if there is normally distributed noise with the same criterion value assumed in Figure 2.

If the noise is normal and has the same mean and variance across the continuum, a stimulus whose mean goodness is at the criterion value will produce random classifications. The goodness value will be above the criterion on half of the trials and below the criterion on the other half. As the goodness value moves away from the criterion value, the noise will have a diminishing effect on the identification judgments. Noise has a larger influence on identification in the middle of the range of goodness values than at the extremes because variability goes in both directions in the middle and only inward at the extremes. This differential effect of noise across the continuum will produce an identification function that has a sharp boundary. Thus, our hypothetical subject giving this result appears to show enhanced discrimination across the category boundary when, in fact, discrimination was constant across the continuum. The shape of the function resulted from noise at the decision stage.

This example shows that categorical decisions made on the basis of continuous information produce identification functions with sharp boundaries, previously taken to indicate categorical perception. Strictly speaking, of course, categorical perception was considered present only if discrimination behavior did not exceed that predicted from categorization. However, one should not have been impressed with the failure of discrimination to exceed that predicted by categorization if the discrimination task resembled something more akin to categorization than discrimination. That is, subjects will tend to rely on identification labels in discrimination tasks if the perceptual memory is poor (Massaro, 1987).

At the theoretical level, it is necessary to distinguish between sensory and decision processes in the categorization task. What is central for our purposes is that decision processes can transform continuous sensory information into

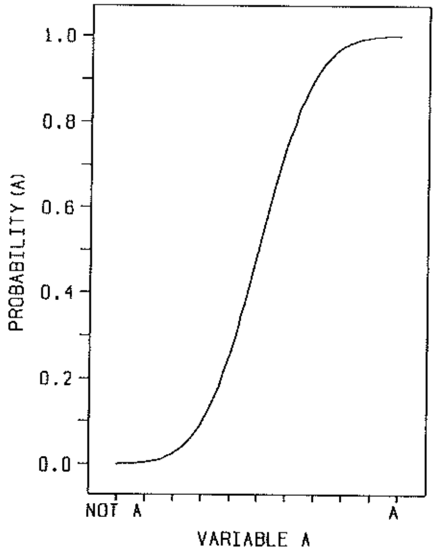

FIG. 3 Probability(A) as a function of Variable A given the linear relationship between goodness(A) and Variable A and the decision criterion represented in Figure 2, but with normally distributed noise added to the mapping of Variable A to goodness(A).

results usually taken to reflect categorical perception. A finding of relatively categorical partitioning of a set of stimuli in no way implies that these stimuli were perceived categorically. Tapping into the process in ways other than simply measuring the identification response reveals the continuous nature of speech perception. Perceivers can rate the degree to which a speech event represents a category, and they can discriminate among different exemplars of the same speech category. Werker (1991) has demonstrated remarkable changes in speech categorization as a function of development and native language. She and others (Kuhl, 1990) have found age-related changes in the sensitivity to nonnative contrasts. These changes are not necessarily evidence for categorical speech perception, however. As Werker (1991, p. 104) states, "However, the fact that adults can still discriminate the nonnative contrasts under certain testing conditions indicates that maintenance is operating at the level of linguistic categories rather than auditory abilities." In addition, reaction times (RT) of identification judgments illustrate that members within a speech category vary in ambiguity or the degree to which they represent the category (Massaro, 1987).

Although speech perception is continuous, there may be a few speech contrasts that qualify for a weak form of categorical perception. This weak form of categorical perception would be reflected in somewhat better discrimination between instances from different categories than between instances within the same category. As an example, consider an auditory /ba/ to /da/ continuum similar to one used in the current experiments. The F2 and F3 transitions were varied in linear steps between the two endpoints of the continuum. The syllable /ba/ is characterized by rising transitions and /da/ by falling transitions. Subjects might discriminate between a rising and falling transition more easily than between two rising or two falling transitions even though the frequency difference is identical in the two cases. Direction of pitch change is more discriminable than the exact magnitude of change. This weak form of categorical perception would arise from a property of auditory processing rather than a special characteristic of speech categories. Thus similar results would be found in humans, chinchillas, and monkeys as well as for nonspeech analogs (as they are, e.g., Kuhl, 1987; Pastore, 1987). However, it is important to note that discrimination between instances within a category is still possible; and although a weak form of categorical perception might exist for a few categories, most do not appear to have this property. We must hence explain continuous rather than categorical speech perception.

Psychology and the speech sciences seem reluctant to give up the notion of categorical perception perhaps, in part, because of phenomenal experience. Our phenomenal experience in speech perception is that of categorical perception. Listening to a synthetic speech continuum between /ba/ and /pa/ provides an impressive demonstration of this. Students and colleagues usually agree that their percept changes qualitatively from one category to the other in a single step or two with very little fuzziness in between. (This author has had similar experiences, hearing certain German phonological categories in terms of similar English ones.) Our phenomenal experience, however, is not enough to confirm the existence of categorical perception. As noted by Marcel (1983), phenomenal experience might be dependent on linking current hypotheses with sensory information. If the sensory information is lost very quickly, continuous information could participate in the perceptual process but might not be readily accessible to introspection. Reading a brief visual display of a word might lead to recognition even though the reader is unable to report certain properties of the type font or even a misspelling of the word. Yet the visual characteristics that subjects cannot report could have contributed to word recognition. Analogously, continuous information could have been functional in speech perception even if retrospective inquiry suggests otherwise. As in most matters of psychological inquiry, we must find methods to tap the processes involved in cognition without depending only on introspective reports.

Dennett (1991) has clarified an important distinction between filling and finding out. We report a variety of experiences such as the apparent motion in the phi phenomenon. The issue for Dennett is whether it is correct to say that the sensory system accomplishes these outcomes by filling in. That is, the sensory system accomplishes an identical outcome in the phi phenomenon that it does in continuous motion. It has been reported that the color of the moving object changes in midstream when a red dot at one location is alternated with a green dot at another location. Does the visual system fill in to give us the impression of a continuously moving dot that changes color? Dennett argues very forcefully that our impressions go beyond the information given, but that our sensory systems do not—that is, they do not fill in.

Dennett's philosophical argument is highly relevant to categorical perception. In the categorical perception viewpoint, there seems to be significant filling in. Categorical perception accomplishes at the sensory/brain level a direct correspondence between some representation and our impression. Categorical perception supposedly occurs because the sensory/perceptual system blurs any stimulus differences within a category and perhaps sharpens stimulus differences between categories. To describe categorical perception in the context of filling in, we perceive two different speech events as the same category because the speech-is-special module makes them equivalent at the sensory/perceptual level. Categorical perception also seems to predict filling in because sensory processing supposedly occurs in such a manner to render the stimuli within a category indiscriminable. This process would be analogous to filling in. On the other hand, it is possible that categorization is simply finding out (as I argue in several places). That is, the goal of speech perception is categorization, and we are able to find out which category best represents the speech event without necessarily modifying the sensory/perceptual representation of that event. In terms of the fuzzy logical model of perception (FLMP), we evaluate, integrate, and make a categorical decision if necessary without necessarily modifying the sensory/perceptual representations of the speech event.

Filling in might also appear to be an attractive explanation of our phenomenal experience of contradictory auditory and visual speech. We are told to report what we hear, and the visible speech biases our experience relative to the unimodal case. Because it is our auditory experience we are reporting, it seems only natural to believe that the representation of the auditory speech has been changed—filled in—by the visual. Another interpretation, however, is that we do not have veridical access of the auditory representation. As Marcel (1983) has pointed out, we report interpretations—finding out—and not representations. Thus, one must be careful about equating phenomenal reports with representations.

Categorical perception has been a popular assumption because it appeared to place certain constraints on the speech perception process—constraints that make speech perception possible and/or easier. If the infant were limited to perceiving only the discrete categories of his or her language, then acquisition of that language would be easier. However, an ability to discriminate within category differences could only hurt speech perception. We know that higher order sentential and lexical information contributes to speech perception. If categorical perception were the case, errors would be catastrophic in that perceivers would access the incorrect category. Categorical perception would also make it difficult to integrate sentential and lexical information with the phonetic information. Continuous information is more naturally integrated with higher order sources of information (Massaro, 1987).

One of the impediments to resolving the controversy is the term PERCEPTION. If perception simply refers to our reported experience, then we cannot deny categorical perception because we naturally attend to the different categories of language. If perception refers to the psychological processing, however, then it is clear that the processing system is not limited to categorical information. One possible reason why categorical perception has been viewed so positively is that scientists misinterpreted the outcome for the processes leading up to the outcome.

Despite our phenomenal experience and the three decades of misinterpreting the relationship between the identification and discrimination of auditory speech, we must conclude that speech is perceived continuously, not categorically. Our work shows that visible and bimodal speech are also perceived continuously. This observation also seems to pull the carpet from under current views of language acquisition that attribute discrete speech categories to the infant and child (Eimas, 1985; Gleitman & Wanner, 1982). Most important, the case for the modularity or specialization of speech is weakened considerably because of its reliance on the assumption of categorical perception. We are now faced with the bigger challenge of explaining how multiple continuous sources of information are evaluated and integrated to achieve a percept with continuous information.

H. Development of Speech Perception

The development of speech perception also speaks to the issue of modularity and the need to assume a specialized processor for speech. Modularity necessarily has a large nativistic component. About two decades ago, investigators presented evidence for this view based on studies of infant speech perception. Early studies seem to find that infants noticed changes between speech categories but not within speech categories (Eimas, 1985). For example, the infant appeared to discriminate an auditory change that changed the signal from /ba/ to /pa/ but not a similar auditory change within either of these two categories. These early studies were misleading, however, and more recent research has shown that infants can discriminate differences within, as well as between, categories (Massaro, 1987). Thus, research with infants reveals that they are capable of discriminating the multiple dimensions of the auditory speech signal, such as the loudness or duration of a speech segment. However, the role these differences play in the language must be learned, and infants are not prewired to categorize the signals into innate phonetic categories. In fact, infants and young children do not appear to discriminate and categorize the speech signal as well as adults. Their caregivers seem to be aware of this limitation because there is also a substantial amount of motherese during the first years of life. In MOTHERESE, the caregiver speaks clearly and slowly to the child. There is also experimental evidence of a slow acquisition of the fundamental distinctions of our spoken language. Children have more difficulty discriminating speech categories, and their ability to discriminate increases gradually across childhood. Even after the onset of schooling, American children have trouble discriminating the segments /v/ and /ð/ (as in *vat* and *that*) (Massaro, 1987).

I. Contextual Effects in Speech Perception

Another strong source of evidence against the modularity of speech perception involves the strong contribution of linguistic and situational context to speech perception. We perceive language more easily when we have some expectation of what the talker is going to say. Many of our conversations involve situations in which we find ourselves predicting exactly what the talker will say next. Experiments have shown that the first words of a sentence can facilitate the recognition of the next word. Another piece of evidence for the positive contribution of context is the finding that trained phoneticians are not able to transcribe a nonnative language accurately. Much of the original detail is lost in the transcription. Not knowing the language or the meaning of the message makes us poorer perceivers.

J. Conclusion

The research that I have reviewed weakens the claim that speech perception requires a specialized module. If speech perception were governed by a specialized module, we would expect no relationship between speech and other skills. However, there is a positive correlation between motor skills and language, and also one between cognitive functioning and vocabulary size. For example, there is a positive correlation between cognitive development and the learning of new words. It seems that speech perception can be considered as one of several perceptual/cognitive functions that can be understood in terms of more general perceptual and learning processes.

III. VARIOUS FORMS OF CONTEXT EFFECTS IN SPEECH PERCEPTION

There is considerable debate concerning how informative the acoustic signal actually is (Blumstein & Stevens, 1979; Cole & Scott, 1974; Liberman, Cooper, Shankweiler, & Studdert-Kennedy 1967; Massaro, 1975b; Massaro & Oden, 1980). Even if the acoustic signal was sufficient for speech recognition under ideal conditions, however, few researchers would believe that the listener relies on only the acoustic signal. It is generally agreed that the listener normally achieves good recognition by supplementing the information from the acoustic signal with information generated through the utilization of linguistic context. A good deal of research has been directed at showing a positive contribution of linguistic context (Cole & Jakimik, 1978; Marslen-Wilson & Welsh, 1978; Pollack & Pickett, 1963). We now review some of this research.

A. Detecting Mispronunciations

Abstracting meaning is a joint function of the independent contributions of the perceptual and contextual information available. In one experiment, Cole (1973) asked subjects to push a button every time they heard a mispronunciation in a spoken rendering of Lewis Carroll's *Through the Looking Glass*. A mispronunciation involved changing a phoneme by 1, 2, or 4 distinctive features (for example, *confusion* mispronounced as *gunfusion, bunfusion,* and *sunfusion,* respectively). The probability of recognizing a mispronunciation increased from 30% to 75% with increases in the number of feature changes, which makes the contribution of the perceptual information passed on by the primary recognition process. The contribution of contextual information should work against the recognition of a mispronunciation since context would support a correct rendering of the mispronounced word. In support of this idea, all mispronunciations were correctly recognized when the syllables were isolated and removed from the passage.

Cole and Jakimik (1978) extended Cole's (1973) mispronunciation task to evaluate how higher order contextual information can influence sentence processing. To the extent that a word is predicted by its preceding context, the listener should be faster at detecting a mispronunciation. This follows from the idea that the quickest way to detect a mispronunciation is to first determine what the intended word is and then notice a mismatch with what was said. Given the sentences *He sat reading a* ***book/bill*** *until it was time to go home for his tea,* mispronouncing the /b/ in *book* as /v/ should be detected faster than the same mispronunciation of *bill.* In fact, listeners were 150 ms faster detecting mispronunciations in highly predictable words than in unpredictable words.

In other experiments Cole and Jakimik (1978) demonstrated similar effects of logical implication. Consider the test sentence *It was the middle of the next day before the killer was caught,* with the /k/ in *killer* mispronounced as /g/. Detection of the mispronunciation should be faster when the text word is implied by the preceding sentence *It was a stormy night when the phonetician was murdered,* compared to the case in which the preceding sentence states that the phonetician merely died. Thematic organization also facilitated recognition of words in their stories. Given an ambiguous story, a disambiguating picture shortened reaction times to mispronunciations of thematically related words but not to mispronunciations of other words that were unrelated to the theme of the story.

Marslen-Wilson (1973) asked subjects to shadow (repeat back) prose as quickly as they heard it. Some individuals were able to shadow the speech at extremely close delays with lags of 250 ms, about the duration of a syllable or so. One might argue that the shadowing response was simply a sound-to-sound mapping without any higher order semantic-syntactic analyses. When subjects make errors in shadowing, however, the errors are syntactically and semantically appropriate given the preceding context. For example, given the sentence *He had heard at the Brigade,* some subjects repeated *He had heard that the Brigade.* The nature of the errors did not vary with their latency; the shadowing errors were always well formed given the preceding context.

B. Limitations of Results

Perceivers have been shown to be efficient exploiters of different types of context to aid in speech perception. However, it might be claimed that the context effects that were observed occurred AFTER speech perception. One might argue, for example, that the rapid shadowing errors observed by Marslen-Wilson (1973) occurred at the stage of speech production rather than speech perception. Analogous to research in other domains, it is essential to locate the stage of processing responsible for experimental findings. A new task has helped address this issue and, more important, the results can be used to reveal how stimulus information and context jointly contribute to word recognition.

C. Gating Task

The gating task (Grosjean, 1980, 1985) has been a recent method developed to assess speech perception and word recognition. As indicated by the name of the task, portions of the spoken message are eliminated or gated out. In a typical task with single words, only the first 50 ms or so of the word is presented. Successive presentations involve longer and longer portions of the word by increasing the duration of each successive presentation by 20 ms. Subjects attempt to name the word after each presentation. Warren and Marslen-Wilson (1987), for example, presented words such as *school* or *scoop*. Figure 4 shows that the probability of correct recognition of a test word increases as additional word information is presented in the gating task.

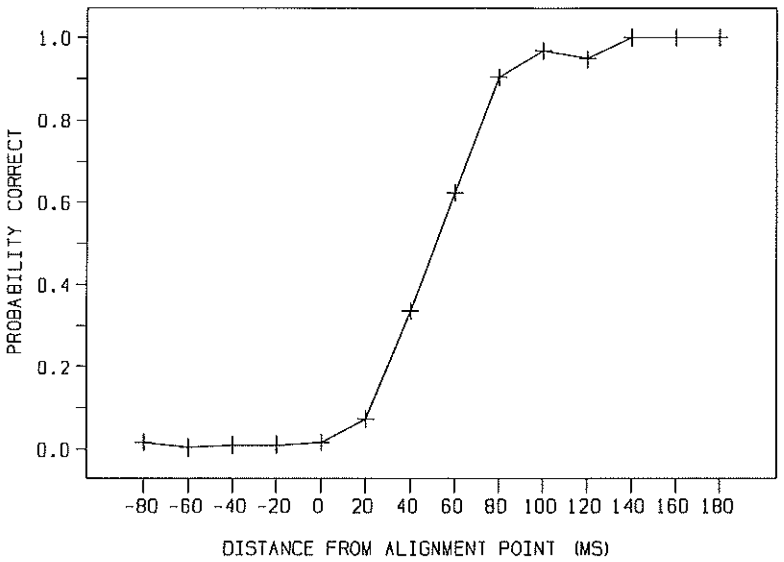

FIG. 4 Probability of correct recognition of the test word as a function of the distance from the alignment point in the test word. The alignment point corresponds to a point near the onset of the final consonant of the word (results adapted from Warren & Marslen-Wilson, 1987).

The gating task appears to have promise for the investigation of speech perception and spoken language understanding. Investigators have worried about two features of the gating task that may limit its external validity. The first feature task is that subjects hear multiple presentations of the test word on a given trial. The standard procedure is to present increasingly larger fragments of the same word on a given trial. The subject responds after each presentation of the fragment. The repeated presentations of the fragment may enhance recognition of the test word relative to the case in which the subject obtains only a single presentation of an item. In visual form perception, for example, it has been shown that repeated tachistoscopic presentations of a test form lead to correct recognition, even though the duration is not increased as it is in the gating task (Uhlarik & Johnson, 1978). The same short presentation of a test form that does not produce correct recognition on its initial presentation can give correct recognition if it is repeated three or four times in the task. This improvement in performance occurs even though the duration of the test stimulus was not increased. These repeated looks at the stimulus can lead to improved performance relative to just a single look. Information from successive presentations can be utilized to improve performance, and therefore multiple presentations lead to better performance than just a single presentation. Based on this result, performance in the gating task might reflect repeated presentations of the test word, in addition to the fact that the successive presentations increased in duration.

Cotton and Grosjean (1984) compared the standard multiple presentation format with the format in which subjects heard only a single fragment from each word in the task. Similar results were found in both conditions. Salasoo and Pisoni (1985) carried out a similar study and found that the average duration of the test word needed for correct identification was only 5 ms less in the task with multiple presentations on a trial than for a single presentation of the test word. Thus, using successive presentations in the gating task appears to be a valid method of increasing the duration of the test word to assess its influence on recognition.

A second question concerning gating tasks has to do with how quickly subjects are required to respond in the task. It could be the case that subjects, given unlimited time to respond in the task, will perform differently from their performance in the on-line recognition of continuous speech. That is, the gating task might be treated as a conscious problem-solving task in which subjects are very deliberate in making their decision about what word was presented. This deliberation would not be possible in a typical situation involving continuous speech, and therefore the results might be misleading. To assess performance under more realistic conditions, Tyler and Wessels (1985) employed a naming response in the gating task. Subjects were required to name the test word as quickly as possible on each trial. In addition, a given word was presented only once to a given subject. The results from this task were very similar to the standard gating test. The durations of the test words needed for correct recognition were roughly the same as that found in the standard gating task. Thus, the experiments exploring the external validity of the gating task have been very encouraging. The results appear to be generalizable to the on-line recognition of continuous speech.

D. Integrating Sentential Context

Tyler and Wessels (1983) used the gating paradigm to assess the contribution of various forms of sentential context to word recognition. Subjects heard a sentence followed by the beginning of the test word (with the rest of the word gated out). The word was increased in duration by adding small segments of the word until correct recognition was achieved. The sentence contexts varied in syntactic and semantic constraints. Some sentence contexts had minimal semantic constraints in that the target word was not predictable in a test given the sentence context and the first 100 ms of the target word. Performance in this condition can be compared to a control condition in which no sentential constraints were present. The experimental question is whether context contributes to recognition of the test word.

Figure 5 gives the probability of correct word recognition as a function of the number of segments in the test word and the context condition. Both variables had a significant influence on performance. In addition, the interaction between the two variables reveals how word information and context jointly influence word recognition. Context influences performance most at intermediate levels of word information. The contribution of context is most apparent when there is some but not complete information about the test word. The lines in Figure 5 give the predictions of the fuzzy logical model of perception

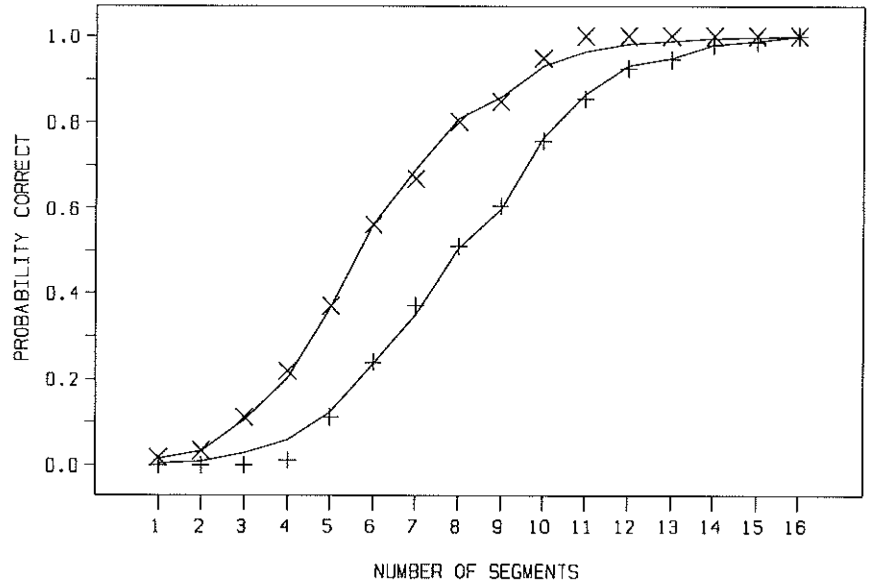

FIG. 5 Observed (*points*) and predicted (*lines*) probability of identifying the test word correctly as a function of the sentential context and the number of segments of the test word. The minimum context (*crosses*) refers to minimum semantic and weak syntactic constraints. The none context (*plusses*) refers to no semantic and weak syntactic constraints (results of Tyler & Wessels, 1983; predictions of the FLMP).

(see Section V,G). The FLMP describes word recognition in terms of the evaluation and integration of word information and sentential context followed by a decision based on the outcome. As can be seen in the figure, the model captures the exact form of the integration of the two sources of information.

A positive effect of sentence context in this situation is very impressive because it illustrates a true integration of word and context information. The probability of correct recognition is zero when context is given with minimum word information. Similarly, the probability of correct recognition is zero with three segments of the test word presented without context. That is, neither the context alone nor the limited word information permits word recognition; however, when presented jointly word recognition is very good. Thus, the strong effect of minimum semantic context illustrated in Figure 5 can be considered to reflect true integration of word and contextual sources of information.

The form of the interaction of stimulus information and context is relevant to the prediction of the cohort model. Marslen-Wilson (1987) assumes that some minimum cohort set must be established on the basis of stimulus information before context can have an influence. In terms of FLMP description, this assumption implies that the evaluation of context should change across different levels of gating. To test this hypothesis, another model was fitted to the results. In this model, context was assumed to have an influence only after some minimum gating interval. Because it is not known what this minimum interval should be, an additional free parameter was estimated to converge on the interval that gave the best description of the observed results. This model did not improve the description of the results, weakening the claim that context has its influence only after some minimum stimulus information has been processed. This result is another instance of the general finding that there are no discrete points in psychological processing. The system does not seem to work one way at one point in time (i.e., no effect of context) and another way in another point in time (i.e., an effect of context).

IV. INDEPENDENT VERSUS INTERACTIVE INFLUENCES OF CONTEXT

We have reached the stage of research in which context effects are well documented. What is important for the next stage is to understand how context and the speech signal come together to support speech perception. There are two general explanations. First, top-down context interacts with bottom-up sensory information to modify the latter's representation. This can be described as a sensitivity effect—context actually modifies the sensitivity of the relevant sensory system. For example, lexical context could change the perceiver's ability to distinguish some speech segment within the word. Second, context might simply provide an additional source of information that supplements the sensory information. In this case, bias is a more appropriate description of the contribution of top-down context. The lexical context biases the speech perception system to perceive some speech segment in the word, but the context does not change the workings of the relevant sensory system. We turn to three studies of context effects with the goal of distinguishing between these two explanations of context effects.

A. Phonemic Restoration

Samuel (1981) reported one of the few other existing experiments addressing sensitivity and bias effects in language processing. He employed a signal detection framework in a study of phonemic restoration. In the original type of phonemic restoration study (Warren, 1970), a phoneme in a word is removed and replaced with some stimulus, such as a tone or white noise. Subjects have difficulty indicating what phoneme is missing. Failure to spot the missing phoneme could be a sensitivity effect or a bias effect. Samuel addressed this issue by creating signal and noise trials. Signal trials contained the original phoneme with superimposed white noise. Noise trials replaced the original phoneme with the same white noise. Subjects were asked to indicate whether or not the original phoneme was present. Sensitivity is reflected in the degree to which the two types of trials can be discriminated and can be indexed by d' within the context of signal detection theory. Bias would be reflected in the overall likelihood of saying that the original phoneme is present.

To evaluate the top-down effects of lexical constraints, Samuel compared performance on phonemes in test words relative to performance on the phoneme segments presented in isolation. A bias was observed in that subjects were more likely to respond that the phoneme was present in the word than in the isolated segment. In addition, subjects discriminated the signal from the noise trials much better in the segment context than the word context. The d' values averaged about two or three times larger for the segment context than for the word context. In contrast to the results of the study of phonological context discussed in Section IV,B, there appears to be a large negative effect of top-down context on sensitivity (changes in sensitivity are equivalent to nonindependent effects of stimulus and context). However, the segment versus word comparison in the Samuel study confounds stimulus contributions with top-down contributions. An isolated segment has bottom-up advantages over the same segment presented in a word. Forward and backward masking may degrade the perceptual quality of a segment presented in a word relative to being presented alone. In addition, the word context might provide co-articulatory information about the critical phoneme which would not be available in the isolated segment.

Samuel carried out a second study that should have overcome the confounding inherent in comparing words and segments. In this study, a word context was compared to a pseudoword context. As an example, the word *living* might be compared to the pseudoword *lathing*, or *modern* might be compared to *madorn*. Samuel also reasoned that pseudowords might show a disadvantage relative to words, simply because subjects would not know what sequence of segments makes up a pseudoword. As an attempt to compensate for this disadvantage for pseudowords, each word or pseudoword was first spoken in intact form (primed) before its presentation as a test item. There was a d' advantage of primed pseudowords over primed words, which Samuel interpreted as a sensitivity effect. Analogous to the difference in the segment and word conditions, a stimulus confounding might also be responsible for the difference between pseudowords and words. Natural speech was used, and therefore an equivalence of stimulus information between the words and pseudowords could not be insured. In fact, the pseudowords averaged about 10%

longer in duration than the words. Longer duration is usually correlated with a higher quality speech signal, which might explain the advantage of the pseudowords over the words.

In a final experiment, Samuel placed test words in a sentence context. The test word was either predicted or not by the sentence context. The results indicated that the predictability of the test word had a significant influence on bias but not sensitivity. The influence of sentence predictability appears to be a valid comparison because there was no apparent stimulus confounding between the predictable and unpredictable contexts. Given the possibility of stimulus confoundings when sensitivity effects were found and no sensitivity effect with a sentence context, it seems premature to conclude that the phonemic restoration paradigm produces sensitivity effects. More generally, top-down effects on sensitivity have yet to be convincingly demonstrated, making the concept of top-down activation unnecessary to explain speech perception.

B. Phonological Context

To study how stimulus information and phonological constraints are used in speech perception, subjects were asked to identify a liquid consonant in different phonological contexts (Massaro, 1989c). Each speech sound was a consonant cluster syllable beginning with one of the three consonants /p/, /t/, or /s/ followed by a liquid consonant ranging (in five levels) from /l/ to /r/, followed by the vowel /i/. The five different levels along the /l/–/r/ continuum differed in terms of the frequency of the third formant (F3) at the onset of the liquid—which is higher for /l/ than /r/. (Formants are bands of energy in the syllable that normally result from natural resonances of the vocal tract in real speech.) There were 15 test stimuli created from the factorial combination of five stimulus levels combined with three initial consonant contexts. Eight elementary school children were instructed to listen to each test syllable and to respond whether they heard /li/ or /ri/.

Figure 6 gives the inverse logistic transform of the average probability of an /r/ response as a function of the two factors. As can be seen in the figure, both factors had a strong effect. The probability of an /r/ response increased systematically with decreases in the F3 transition. Phonological context also had a significant effect on the judgments. Subjects responded /r/ more often given the context /t/ than given the context /p/. Similarly, there were fewer /r/ responses given the context /s/ than given the context /p/. Finally, the significant interaction reflected the fact that the phonological context effect was greatest when the information about the liquid was ambiguous. As will be described in Section V,C, these two factors had independent influences on performance.

C. Lexical Context

Elman and McClelland (1988) carried out an ingenious demonstration of context effects in speech perception. Because of co-articulation—the influence of producing one speech segment on the production of another—a given speech segment has different acoustic forms in different contexts. The phonemes /s/ and /ʃ/ are necessarily produced differently and will differentially influence the production of the following speech segment. Perceivers not only recognize the

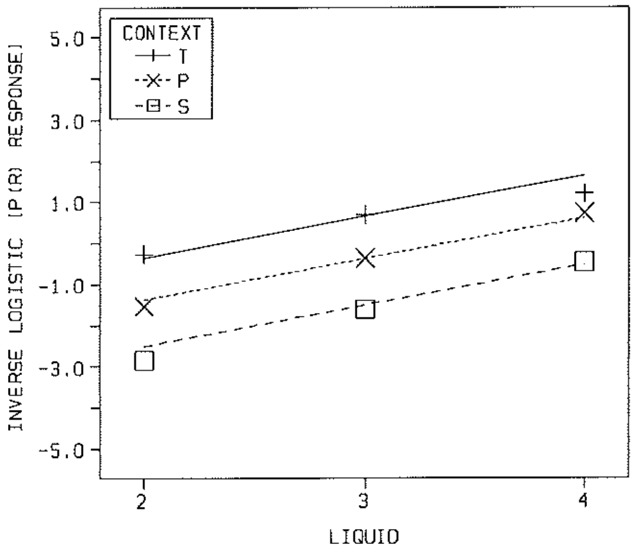

FIG. 6 Predicted (*lines*) and observed (*points*) inverse logistic transformation of the probability of an /r/ identification for the middle three levels of liquid. Context is the curve parameter (results of Massaro, 1989c).

different speech segments /s/ and /ʃ/; they apparently are able to compensate for the influence of these segments in recognizing the following speech segment. During production of speech, co-articulation involves the assimilation of the acoustic characteristics of one sound in the direction of the characteristics of the neighboring sound. The production of /s/ contains higher frequency energy than /ʃ/, and co-articulation will result in the sound following /s/ having higher frequency energy. The energy in /k/ is somewhat lower in frequency than that in initial /t/—the /t/ has a high burst. Thus, /s/ biases the articulation of a following stop in such a way that the stop segment has somewhat higher frequency energy. The segment /ʃ/, on the other hand, biases the articulation of a following stop in such a way that the stop segment has somewhat lower frequency energy. Perceivers apparently take this assimilative coarticulatory influence into account in their perceptual recognition of /t/ and /k/ (and /d/ and /g/) and show a contrast effect. Mann and Repp (1981) showed that recognition of the following segment as /t/ or /k/ is dependent on whether the preceding segment is /s/ or /ʃ/. Given a vowel–fricative syllable followed by a stop–vowel syllable, subjects were more likely to identify the stop as /k/ than /t/ if the preceding fricative was /s/ than if it was /ʃ/ (a contrast effect).

The goal of the Elman and McClelland (1988) study was to induce the same contrast effect but mediated by the lexical identity of a word rather than the acoustic structure of the preceding syllable. Using synthetic speech, a continuum of speech sounds ranging between *tapes* and *capes* was made by varying the onset properties of the sounds. These sounds were placed after the words *Christmas* and *foolish*. As expected from the Mann and Repp (1981) study, there were more judgments of *capes* following *Christmas* than following *foolish*. However, this dependency could have been triggered directly by the acoustic differences between /s/ and /ʃ/. To eliminate this possibility, Elman and McClelland (1988) created an ambiguous sound half way between /s/ and /ʃ/ and replaced the original fricatives in *Christmas* and *foolish* with this ambiguous sound. Given a lexical context effect first reported by Ganong (1980) and also replicated by Connine and Clifton (1987), we would expect that the ambiguous segment would tend to be categorized as /s/ when it occurs in *Christmas* and as /ʃ/ when it occurs in *foolish*. The empirical question is whether the same contrast effect would occur given the same ambiguous segment in the two different words. That is, Would just the lexical identity of the first word also lead to a contrast effect in the recognition of the following speech segment varying between *tapes* and *capes*? In fact, subjects were more likely to report the test word *capes* following the context word *Christmas* than following the context word *foolish,* and this effect was larger when the segmental information about the /k/–/t/ distinction in the test word was ambiguous.

How does an interactive activation model such as TRACE describe this effect? According to Elman and McClelland (1988), the contrast effect can be induced by assuming connections from the phoneme level in one time slice to the feature level in adjacent time slices (as in TRACE I, Elman & McClelland, 1986). In our example, the units corresponding to /s/ and /ʃ/ would be connected laterally and downward to feature units which in turn are connected upward to the phoneme units /t/ and /k/. The downward activation from the fricative phoneme to the feature level would modulate the upcoming upward activation from the feature level to the stop phonemes. To describe the lexical effect for the case in which the two words *Christmas* and *foolish* have the same ambiguous final fricative segment, top-down connections from the word level to the phoneme level would activate the appropriate phoneme unit—/s/ and /ʃ/ in *Christmas* and *foolish,* respectively. These units would then activate downward to the feature level, leading to a contrast effect. Because of the assumed top-down activation modulating the bottom-up activation, interactive activation is central to their explanation.

However, an adequate explanation of the Elman and McClelland results does not require interactive activation. The results simply show that top-down information from the lexical level can influence the amount of information transmitted at the sublexical level. It is the lexical context that disambiguates the final segment of the context word which, in turn, influences identification of the first segment of the following word. We already know that lexical context influences identification of the segments that make up a word (Ganong, 1980). In terms of the FLMP, the lexical context and the segmental information are integrated to achieve perceptual recognition and, therefore, identification of the ambiguous segment. Elman and McClelland (1988) have extended this phenomenon to an indirect measure of identification of the critical segment (/s/ or /ʃ/) by assessing its influence on a following segment (/t/ or /k/). Although this result contributes to the validity of top-down effects on perceptual processing by making the hypothesis of a postperceptual decision less likely, the result appears to be neutral with respect to the existence of interactive activation. In fact, Figure 7 gives the fit of the FLMP to the results (Elman & McClelland, 1988, Experiment 1). Nine free parameters were estimated to predict the 28 data points: seven for the seven levels along the *tapes–capes* continuum, one for /s/ or /ʃ/ in the intact context word condition, and one for

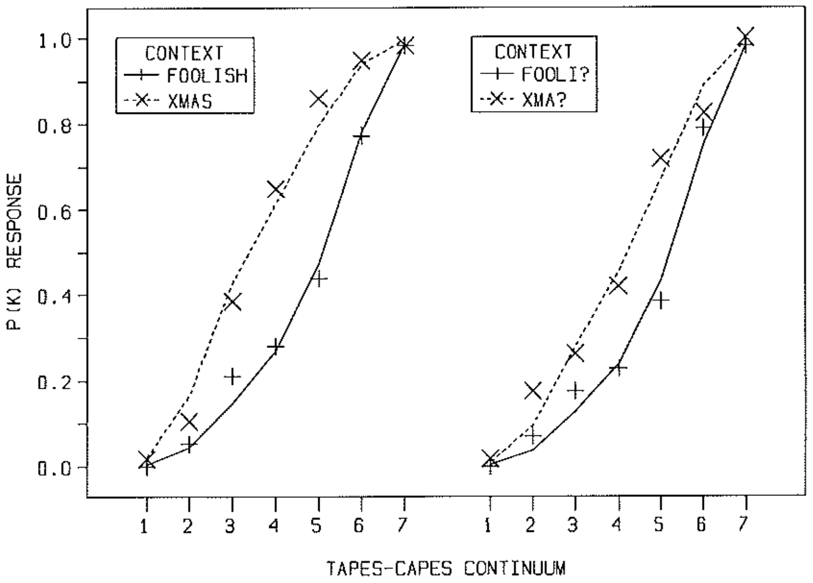

FIG. 7 Observed (*points*) and predicted (*lines*) probability of a /k/ identification as a function of stimulus and preceeding context for original and ambiguous preceding consonant. Results from Elman and McClelland (1988). The predictions are for the FLMP.

lexical context. The pure lexical context effect is seen in the right panel and the combined effect of lexical context and context segment (/s/ or /ʃ/) is shown in the left panel. It should be emphasized that the FLMP explanation is in terms of perceptual processes and is not simply a result of a postperceptual decision mechanism. We now consider extant theories of speech perception and word recognition and evaluate them within the context of empirical evidence.

V. THEORIES OF SPEECH PERCEPTION AND WORD RECOGNITION

Although there are several current theories of spoken word recognition, they can be classified and described fairly easily. All theories begin with the acoustic signal and usually end with access to a word or phrase in the mental lexicon. Six models of word recognition will be discussed to highlight some important issues in understanding how words are recognized. We review several important characteristics of the models to contrast and compare the models. Figure 8 gives a graphical presentation of these characteristics. One important question is whether word recognition is mediated or nonmediated. A second question is whether the perceiver has access to only categorical information in the word recognition process, or whether continuous information is available. A third consideration is whether information from the continuously varying signal is used on line at the lexical stage of processing, or whether there is some delay in initiating processing of the signal at the lexical stage. A fourth characteristic involves parallel versus serial access to the lexical representations in memory. The final characteristic we will consider is whether the word recognition process functions autonomously, or whether it is context-dependent.

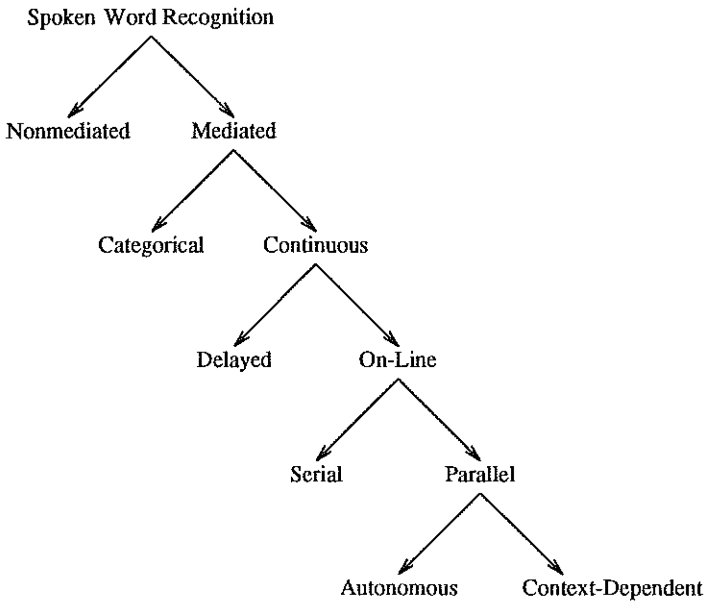

FIG. 8 Tree of wisdom illustrating binary oppositions central to the differences among theories of spoken word recognition.

A. Logogen Model

The logogen model described by Morton (1964, 1969) has had an important influence on how the field has described word recognition. Morton proposed that each word that an individual knows has a representation in long-term memory. To describe this representation, Morton used the term LOGOGEN—*logos,* meaning 'word', and *genēs,* meaning 'born'. Each logogen has a resting level of activity, and this level of activity can be increased by stimulus events. Each logogen has a threshold—when the level of activation exceeds the threshold, the logogen fires. The threshold is a function of word frequency; more frequent words have lower thresholds and require less activation for firing. The firing of a logogen makes the corresponding word available as a response. Figure 9 gives a schematic diagram of the logogen model.

Morton's logogen model can be evaluated with respect to the five characteristics shown in Figure 8. The model is nonmediated because there is supposedly a direct mapping between the input and the logogen. That is, no provision has been made for smaller segments, such as phonemes or syllables, to mediate word recognition. The perceiver of language appears to have continuous information, given that the logogen can be activated to various degrees. On the other hand, one might interpret the theory as categorical because of the assumption of a threshold below which the logogen does not fire. Processing is on-line rather than delayed. With respect to the fourth issue, words are activated in parallel rather than serially. Finally, as can be seen in Figure 9, the logogen allows for the contribution of contextual information in word recognition. Contextual information activates logogens in the same way that information from the stimulus word itself activates logogens. The main limitation in the logogen model is

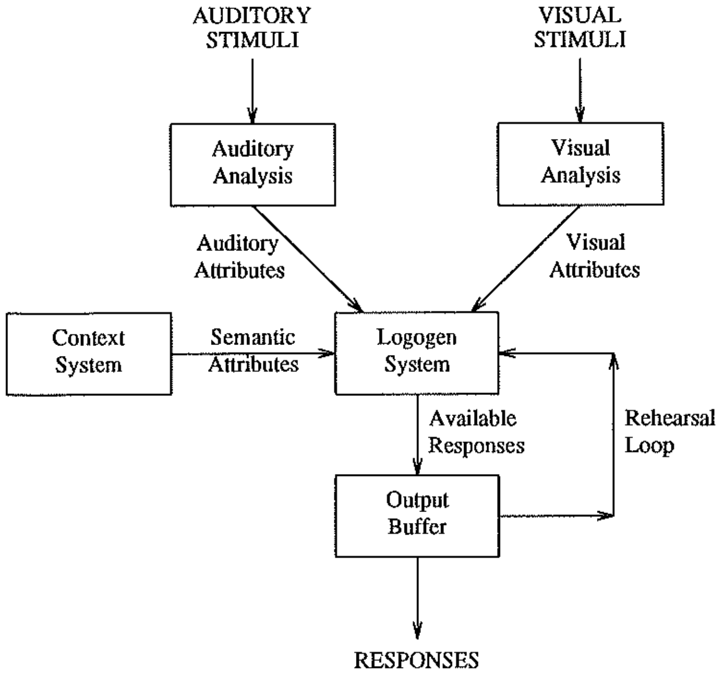

FIG. 9 A schematic diagram of the logogen model. Recognition occurs when the activation in a logogen exceeds a critical level and the corresponding word becomes available as a response (adapted from Morton & Broadbent, 1967).

its nonmediated nature. Thus, the model has difficulty explaining intermediate recognition of sublexical units (e.g., CV syllables) and how nonwords are recognized. However, an important feature of the logogen model is the assumed independence of stimulus information and context in speech perception.

B. Cohort Model

A recent influential model of word recognition is the COHORT model (Marslen-Wilson, 1984). According to this model, word recognition proceeds in a left-to-right fashion on line with the sequential presentation of the information in a spoken word. The acoustic signal is recognized phoneme by phoneme from left to right during the word presentation. Each phoneme is recognized categorically. Word recognition occurs by way of the elimination of alternative word candidates (cohorts). Recognition of the first phoneme in the word eliminates all words that do not have that phoneme in initial position. Recognition of the second phoneme eliminates all the remaining cohorts that do not have the second phoneme in second position. Recognition of phonemes and the elimination of alternative words continues in this fashion until only one word remains. It is at this point that the word is recognized. Figure 10 gives an example illustrating how the corhort model recognizes the word *elephant*.

The corhort model is easy to describe with respect to the five characteristics in Figure 8. The model is mediated, categorical, on-line, parallel, and contextu-

/ɛ/	/ɛl/	/ɛl ə /	/ɛl ə f/	/ɛl ə f ə /
aesthetic	elbow	elegiac	elephant	elephant
any	elder	elegy	elephantine	
.	eldest	element		(1)
.	eleemosynary	elemental	(2)	
ebony	elegance	elementary		
ebullition	elegiac	elephant		
echelon	elegy	elephantine		
.	element	elevate		
.	elemental	elevation		
economic	elementary	elevator		
ecstacy	elephant	elocution		
.	elephantine	eloquent		
.	elevate			
element	elevation	(12)		
elephant	.			
elevate	.			
.				
.	(28)			
entropy				
entry				
.				
.				
extraneous				
.				
(324)				

FIG. 10 "Illustration of how the word *elephant* is recognized, according to the cohort model (Marslen-Wilson, 1984). Phonemes are recognized categorically and on-line in a left-to-right fashion as they are spoken. All words inconsistent with the phoneme string are eliminated from the cohort. The number below each column represents the number of words remaining in the cohort set at that point in processing the spoken word. Note that the example is for British pronunciation in which the third vowel of *elephantine* is pronounced /æ/" (from Massaro, 1989a).

ally dependent to some extent. Word recognition is mediated by phoneme recognition, phonemes are recognized on-line categorically, words are accessed in parallel, and the word alternative finally recognized can be influenced by context. The primary evidence against the cohort model concerns the categorical recognition of phonemes. We have seen that phonemes are not perceptual units and the speech perception is not categorical.

As might be expected, the cohort model has not gone unmodified. Its advocates have acknowledged that the model's integrity is not critically dependent on phonemes as the unit of analysis. Thus, simpler features could be processed as they occur to establish a viable cohort. In addition, the features need not work in an all-or-none fashion, but could provide continuous activation—allowing a fuzzy boundary between words in and out of the cohort. Although the theory allows speech to be processed on line, it can also be modified to allow word recognition to occur somewhat later than the normatively ideal recognition point. These modifications are necessary to bring the model in line with empirical results, but they weaken the model considerably and make it more difficult to test against alternative models.

C. TRACE Model

The TRACE model of speech perception (McClelland & Elman, 1986) is one of a class of models in which information processing occurs through excitatory and inhibitory interactions among a large number of simple processing units. These units are meant to represent the functional properties of neurons or neural networks. Three levels or sizes of units are used in TRACE: feature, phoneme, and word. Features activate phonemes which activate words, and activation of some units at a particular level inhibits other units at the same level. In addition, an important assumption of interactive activation models is that activation of higher order units activates their lower order units; for example, activation of the /b/ phoneme would activate the features that are consistent with that phoneme.

With respect to the characteristics in Figure 8, the TRACE model is mediated, on-line, somewhat categorical, parallel, and context-dependent. Word recognition is mediated by feature and phoneme recognition. The input is processed on-line in TRACE, all words are activated by the input in parallel, and their activation is context-dependent. In principle, TRACE is continuous, but its assumption about interactive activation leads to a categorical-like behavior at the sensory (featural) level. According to the TRACE model, a stimulus pattern is presented, and activation of the corresponding features sends more excitation to some phoneme units than others. Given the assumption of feedback from the phoneme to the feature level, the activation of a particular phoneme feeds down and activates the features corresponding to that phoneme (McClelland & Elman, 1986, p. 47). This effect of feedback produces enhanced sensitivity around a category boundary, exactly as predicted by categorical perception. Evidence against phonemes as perceptual units and against categorical perception is, therefore, evidence against the TRACE model.

The TRACE model is structured around the process of interactive activation between layers at different levels and also competition within layers. Because of this process, the representation over time of one source of information is modified by another source of information. Contrary to independence predicted by the FLMP, TRACE appears to predict nonindependence of top-down and bottom-up sources of information. As discussed in Section IV,B, Massaro (1989c) varied a top-down and a bottom-up source of information in a speech identification task. An important question is whether the top-down context from the lexical level modified the representation at the phoneme level. The TRACE model accounts for the top-down effects of phonological constraints by assuming interactive activation between the word and phoneme levels. Bottom-up activation from the phoneme units activates word units, which in turn activate the phoneme units that make them up. Interactive activation appropriately describes this model because it is clearly an interaction between the two levels that is postulated. The amount of bottom-up activation modifies the amount of top-down activation, which then modifies the bottom-up activation, and so on.

In terms of the logistic results in Figure 6, an independent influence of context should simply change the spread among the curves, whereas a nonindependent effect should differentially influence their slopes. Thus, nonindependence effects would be seen in nonparallel functions, contrary to the results that are observed.

I claimed that the concept of interactive activation, as implemented in TRACE, should produce nonindependence effects (Massaro, 1989b). Take as an example a liquid phoneme presented after the initial consonant /t/. The liquid would activate both /l/ and /r/ phonemes to some degree; the difference in activation would be a function of the test phoneme. There are many English words that begin with /tr/ but none than begin with /tl/, and therefore there would be more top-down activation for /r/ than for /l/. Top-down activation of /r/ would add to the activation of the /r/ phoneme at the phoneme level. What is important for our purposes is that the amount of top-down activation is positively related to the amount of bottom-up activation. Now consider the top-down effects for the two adjacent stimuli along the /l/–/r/ continuum. Both test stimuli activate phonemes to some degree, and these phonemes activate words, which then activate these phonemes. Given that two adjacent syllables along the continuum are different, they have different patterns of bottom-up activation, and therefore, the top-down activation must also differ. The difference in the top-down activation will necessarily change the relative activation of the two phonemes. This relationship between top-down and bottom-up activation should be reflected in a nonindependent effect of top-down context.

Because the TRACE model, as originally formulated, cannot be tested directly against the results, a simulation of the experiment with TRACE was compared the observed results. A simulation allows a test of fundamental properties of TRACE rather than a concern with specific results that are primarily a consequence of the details of the implementation. Differences due to the makeup of the lexicon and specific parameter values are less important than systematic properties of the predictions. Within the current architecture of the TRACE model, the word level appears to play a fundamental role in the discrimination of alternatives at the phoneme level. The most straightforward test of this observation is to simulate results with the standard TRACE model and compare this simulation with the observed results. The simulation used the lexicon, input feature values, and parameter values given in McClelland and Elman (1986, Tables 1 and 3). Three levels of information about the liquid (*l*, *r*, and *L*) were used as three levels of input information. The phoneme /L/ refers to an intermediate level of a liquid phoneme with neutralized diffuse and acute feature specifications. The other feature specifications for /L/ are the same as those for /l/ and /r/. Thus, the input /L/ activates the two liquids more than the other phonemes but activates /l/ and /r/ to the same degree. These three liquids were placed after initial /t/, /p/, and /s/ contexts and followed by the vowel /i/. The simulations, therefore, involved a test of these nine stimulus conditions.

A simulation of TRACE involves presentation of a pattern of activation to the units at the feature level. The input is presented sequentially in successive time slices, as would be the case in real speech. The processing of the input goes through a number of cycles in which all the units update their respective activations at the same time, based on the activations computed in the previous update cycle. The TRACE simulation is completely deterministic; a single run is sufficient for each of the three initial consonant conditions. The activation of the /l/ and /r/ units at the phoneme level occurred primarily at the 12th time slice of the trace, and these values tended to asymptote around the 54th cycle of the simulation run. Therefore, the activations at the 12th time slice after the 54th cycle were taken as the predictions of the model. These activations cannot be taken as direct measures of the question of the independence of top-down and bottom-up sources of information. In order to assess this question, it is necessary to map these activation levels into predicted responses.

The predicted proportion of /l/ and /r/ responses is not given by the activations directly. McClelland and Elman (1986) assume that the activation a_i of a phoneme unit is transformed by an exponential function into a strength value S_i,

$$S_i = e^{ka_i} \tag{1}$$

The parameter k is assumed to be 5. The strength value S_i represents the strength of alternative i. The probability of choosing an alternative i, $P(R_i)$, is based on the activations of all relevant alternatives, as described by Luce's (1959) choice rule.

$$P(R_i) = \frac{S_i}{\sum} \tag{2}$$

where $\sum$ is equal to the sum of the strengths of all relevant phonemes, derived in the manner illustrated for alternative i. The activation values were translated into strength values by the exponential function given by Eq. (1). The constant k was set equal to 5. The probability of an /r/ judgment was determined from the strength values using Eq. (2).

To determine if top-down context makes an independent or nonindependent contribution, the response proportions were translated into logistic values. This analysis is analogous to the Braida and Durlach (1972) and Massaro (1979) analyses, except the logistic rather than the Gaussian transform is used. The two transforms are very similar to one another. In addition, the present analysis of independence versus nonindependence parallels the question of sensitivity versus bias in those previous studies and in Massaro (1989b). These logistic values are given in Figure 11. As can be seen in the figure, the predicted curves are not parallel. In terms of the present analysis, the contribution of top-down context is nonindependent. Thus the simulation is consistent with the intuition that interactive activation between the word and phoneme levels in TRACE produces nonindependent changes at the phoneme level (Massaro, 1988).

At first glance, the effect of the context /p/ seems strange, because there is a strong bias for /r/ rather than for /l/. One might have expected very little difference, because initial /p/ activates both /pr/ and /pl/ words. However, the makeup of the lexicon used in the simulation favored /r/ much more than /l/. In this case, the /p/ context functions more like the /t/ context.

The predictions of TRACE were also determined for other values of the constant k used in Eq. (1) that maps activations into strength values. Eight values of k were used, giving a total of eight simulated subjects. The values of k were 0.5, 1, 2, 3.5, 5, 7.5, 10, and 15. For each value of k, there was a nonindependent effect of context. Given that TRACE has been shown to predict

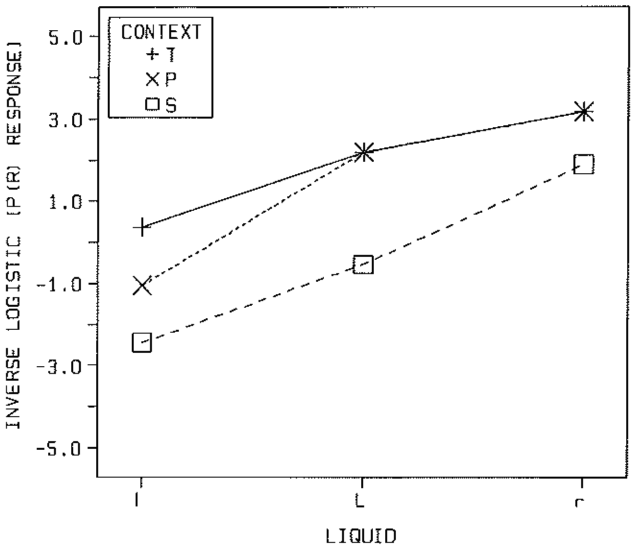

FIG. 11 Inverse logistic of the probability of an /r/ identification as a function of liquid and context. Predictions of the TRACE model (from Massaro, 1989c).

nonindependence of stimulus and context, the predictions are falsified by the actual results in Figure 6.

D. Autonomous Search Model

A fourth model is an autonomous search model that has been proposed to describe word recognition (Forster, 1979, 1981, 1985). The model involves two stages: an initial access stage and a serial search stage. This model was developed for the recognition of written words rather than for recognizing spoken words. However, advocates of the model have begun to apply its basic assumptions to spoken word recognition (Bradley & Forster, 1987). For ease of presentation, we present the model in terms of recognizing a written word.

The first stage in processing a written stimulus involves recognizing the letters that make up a word. The abstract representation of this information serves as an access code to select some subset of the lexicon. The distinctive feature of this model is that words within this subset must be processed serially. The serial order of processing is determined by the frequency of occurrence of the words in the language. After making a match in the search stage of processing, a verification or postsearch check is carried out against the full orthographic properties of the word. If a match is obtained at this stage, the relevant contents of the lexical entry are made available.

The autonomous search model can be described with respect to the five characteristics in Figure 8. The model is mediated, categorical, on-line, serial, and contextually independent. Written word recognition is mediated by letter recognition, letters are recognized on line categorically, final recognition of a word requires a serial search. All this processing goes on without any influence from the context at other levels, such as the sentence level. The autonomous

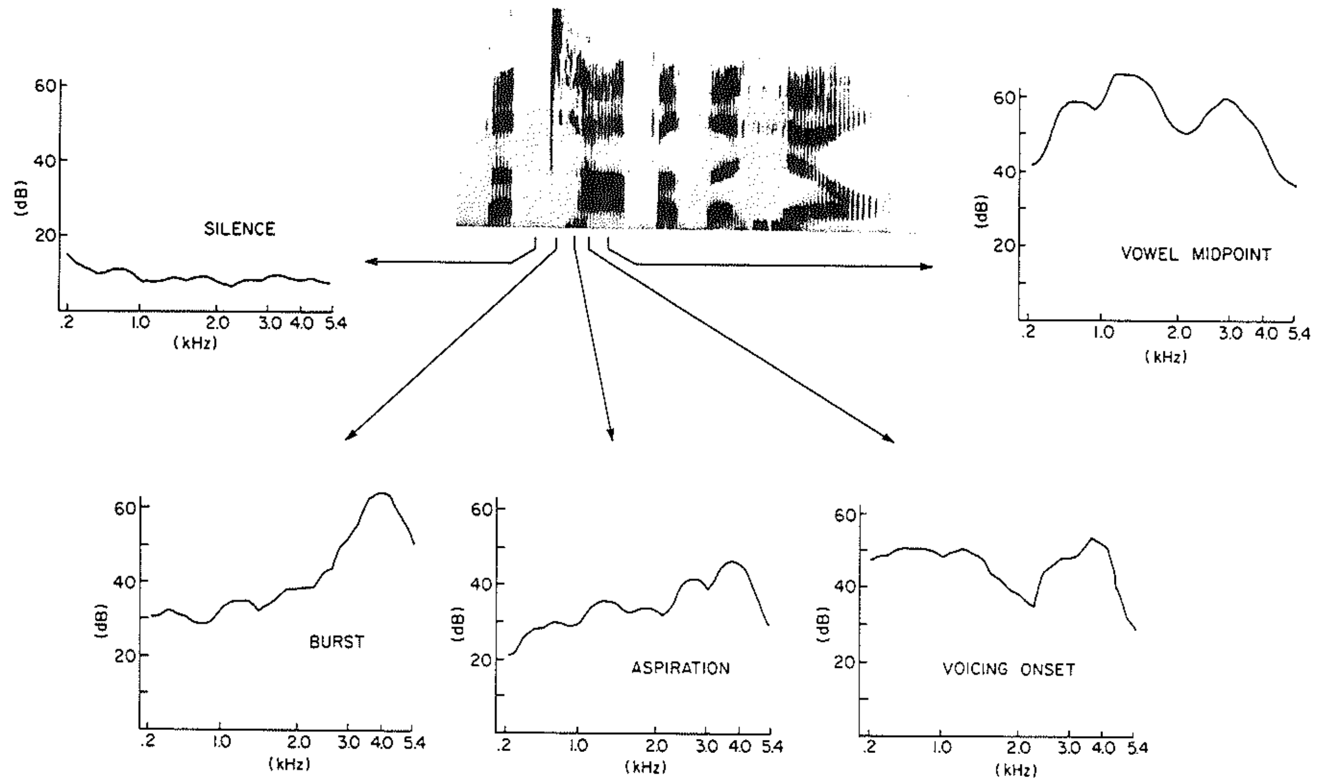

Fig. 12 "The phonetic transition from the middle of [t] to the middle of [a] . . . has been approximated by a sequence of five static critical-band spectra" (from Klatt, 1989, p. 193).

search model appears to fail on at least two counts: categorical perception and contextually independent processing. We have reviewed evidence for continuous perception, and there is convincing evidence for the influence of context in word recognition (see section on Lexical Context).

E. Lexical Access from Spectra Model

Klatt (1979) developed a lexical access from spectra (LAFS) model that bypasses features and segments as intermediate to word recognition. The expected spectral patterns for words and for cross-word boundaries are represented in a very large decoding network of expected sequences of spectra. Figure 12 illustrates how each word is first represented phonemically, then all possible pronunciations are determined by phonetic recording rules specifying alterna-

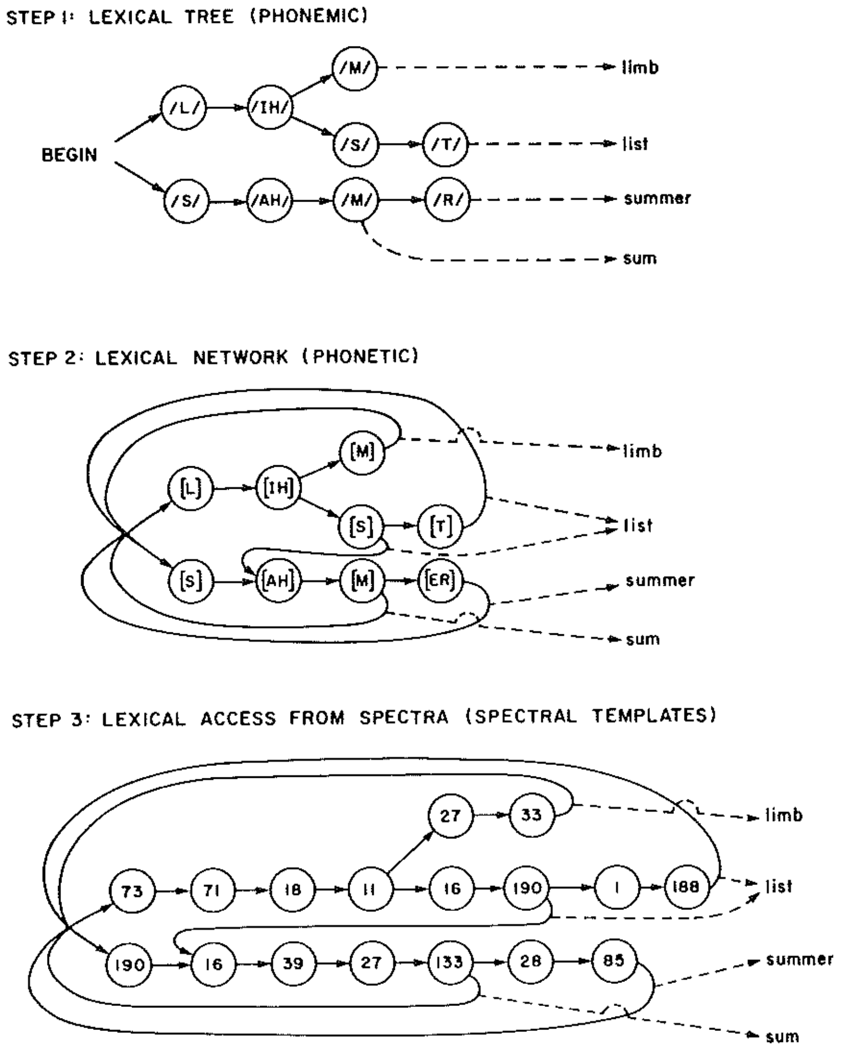

FIG. 13 The lexical tree, lexical network, and lexical access from spectra of the LAFS model (from Klatt, 1989, p. 195).

tive pronunciations within and across word boundaries, and these phonetic representations are converted to sequences of spectral templates like those shown in Figure 13. Figure 12 shows a sequence of 5 static critical-band spectra corresponding to the middle of [t] to the middle of [a].

Central to the LAFS model are the assumptions that running spectra fully represent speech and that the differences among spectra can differentiate among the meaningful differences in real speech. With respect to the five characteristics in Figure 8, the model is mediated, continuous, on line, parallel, and contextually independent. A goodness-of-match is determined for each word path based on the running spectra of the speech stimulus. The goodness-of-match provides continuous and not just categorical information. Multiple alternatives can be evaluated in parallel and on line as the speech signal arrives. Finally, the contextual dependencies built into the representation are phonologically based, and therefore there is no provision for semantic and syntactic constraints. That is, the contribution of linguistic context is limited to its effects on articulation and, therefore, properties of the speech signal. Constraints over and above this influence are not accounted for in the model. Thus, the model could not easily account for the positive contribution of linguistic context.

F. Lexical Access from Features (LAFF) Model

Stevens (1986; 1988, cited in Klatt, 1989) has articulated a model describing lexical access via acoustic correlates of linguistic binary phonetic features (LAFF). These features are language universal and binary (present or absent). The display in (1) includes a conventional featural representation of the word *pawn*. At the right it is modified to reflect expectations as to the temporal locations within the syllable of acoustic information important to the detection of feature values. In addition, features not specified by a plus or minus are deemed not critical to the lexical decision. This model is driven by parsimony in that the features are assumed to be binary and robust. Binary features allow the integration process to be shortcircuited in that multiple ambiguous sources of information do not have to be combined. With respect to the five characteristics in

(1) Conventional and Modified Lexical Representation (Stevens, 1988, cited in Klatt, 1989)

	Conventional			Modified						
Features	*p*	*ɔ*	*n*	*p*			*ɔ*			*n*
high	−	−	−					−		
low	−	+	−					+		
back	−	+	−					+		
nasal	−	−	+						+	
spread glottis	+	−	−			+				
sonorant	−	+	+		−					
voiced	−	+	+		−					
strident	−	−	+							
consonantal	+	−	+	+						+
coronal	−	−	+	−						+
anterior	+	−	+	+						+
continuant	−	+	−	−						−

Figure 8, the model is mediated, categorical, delayed, parallel, and contextually independent. A goodness-of-match is determined for each word path based on the distinctive features assembled from the speech input. As we discussed in Section V,C, there is strong evidence against categorical information at the feature level. The goodness-of-match provides just categorical information with respect to each feature. Continuous information could be derived from the number of features that match each word in memory. Multiple alternatives can be evaluated in parallel, but the matching process cannot perform reliably until the complete word has been presented. Finally, the contextual dependencies built into the representation are phonologically based, and therefore there is no prescribed provision for linguistic constraints. Thus, the model has difficulty with the positive contribution of linguistic context.

G. Fuzzy Logical Model of Perception

The central thesis of this present framework is that there are multiple sources of information supporting speech perception, and the perceiver evaluates and integrates all these sources to achieve perceptual recognition. Within just the auditory signal, there are many different sources of information or cues that the listener uses to decode the message. For example, investigators have listed 16 different acoustic properties that distinguish /igi/ from /iki/. As noted earlier, perceivers also use situational and linguistic context to help disambiguate the signal. Finally, it has been repeatedly demonstrated that perceivers use information from other modalities in face-to-face communication. Both lip movements and hand gestures have been shown to aid in speech perception.

According to the fuzzy logical model of perception, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns (Massaro, 1987). Similar to other approaches, it is assumed that speech is processed through a sequence of processing stages (Pisoni & Luce, 1987). The model has received support in a wide variety of domains and consists of three operations in perceptual (primary) recognition: feature evaluation, feature integration, and decision. Continuously valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness-of-match of the stimulus information with the relevant prototype descriptions.

Central to the FLMP are summary descriptions of the perceptual units of the language. These summary descriptions are called prototypes, and they contain a conjunction of various properties called features. A prototype is a category, and the features of the prototype correspond to the ideal values that an exemplar should have if it is a member of that category. The exact form of the representation of these properties is not known and may never be known. However, the memory representation must be compatible with the sensory representation resulting from the transduction of the audible and visible speech. Compatibility is necessary because the two representations must be related to one another. To recognize the syllable /ba/, the perceiver must be able to relate the information provided by the syllable itself to some memory of the category /ba/.

Prototypes are generated for the task at hand. In speech perception, for example, we might envision activation of all prototypes corresponding to the perceptual units of the language being spoken. For ease of exposition, consider a speech signal representing a single perceptual unit, such as the syllable /ba/. The sensory systems transduce the physical event and make available various sources of information called features. During the first operation in the model, the features are evaluated in terms of the prototypes in memory. For each feature and for each prototype, featural evaluation provides information about the degree to which the feature in the speech signal matches the featural value of the prototype.

Given the necessarily large variety of features, it is necessary to have a common metric representing the degree of match of each feature. The syllable /ba/, for example, might have visible featural information related to the closing of the lips and audible information corresponding to the second and third formant transitions. These two features must share a common metric if they eventually are going to be related to one another. To serve this purpose, fuzzy truth values (Zadeh, 1965) are used because they provide a natural representation of the degree of match. Fuzzy truth values lie between zero and one, corresponding to a proposition being completely false and completely true. The value .5 corresponds to a completely ambiguous situation, whereas .7 would be more true than false, and so on. Fuzzy truth values, therefore, not only can represent continuous rather than just categorical information, they can also represent different kinds of information. Another advantage of fuzzy truth values is that they couch information in mathematical terms (or at least in a quantitative form). This allows the natural development of a quantitative description of the phenomenon of interest.

Feature evaluation provides the degree to which each feature in the syllable matches the corresponding feature in each prototype in memory. The goal, of course, is to determine the overall goodness of match of each prototype with the syllable. All the features are capable of contributing to this process, and the second operation of the model is called feature integration. That is, the features (actually, the degrees of matches) corresponding to each prototype are combined (or conjoined, in logical terms). The outcome of feature integration consists of the degree to which each prototype matches the syllable. In the model, all features contribute to the final value, but with the property that the least ambiguous features have the most impact on the outcome.

The third operation during recognition processing is decision. During this stage, the merit of each relevant prototype is evaluated relative to the sum of the merits of the other relevant prototypes. This relative goodness-of-match gives the proportion of times the syllable is identified as an instance of the prototype. The relative goodness-of-match could also be determined from a rating judgment indicating the degree to which the syllable matches the category. The pattern classification operation is modeled after Luce's (1959) choice rule. In pandemonium-like terms (Selfridge, 1959), we might say that it is not how loud some demon is shouting but rather the relative loudness of that demon in the crowd of relevant demons. An important prediction of the model is that one feature has its greatest effect when a second feature is at its most ambiguous level. Thus, the most informative feature has the greatest impact on the judgment.

Figure 14 illustrates the three stages involved in pattern recognition. The three stages are shown to illustrate their necessarily successive but overlapping processing. Different sources of information are represented by capital letters. The evaluation process transforms these into psychological values (indicated by lowercase letters) that are then integrated to give an overall value. The classification operation maps this value into some response, such as a discrete decision or a rating. The model confronts several important issues in describing speech perception. One fundamental claim is that multiple sources of information are evaluated in speech perception. The sources of information are both bottom-up and top-down. Two other assumptions have to do with the evaluation of the multiple sources of information. Continuous information is available from each source, and the output of evaluation of one source is not contaminated by the other source. The output of the integration process is also assumed to provide continuous information. With respect to the contrasts in Figure 8, spoken word recognition is mediated, continuous, on-line, serial and parallel, and both autonomous and context-dependent.

The theoretical framework of the FLMP has proven to be a valuable framework for the study of speech perception. Experiments designed in this framework have provided important information concerning the sources of information in speech perception and how these sources of information are processed to support speech perception. The experiments have studied a broad range of information sources, including bottom-up sources such as audible and visible characteristics of speech and top-down sources, including phonological, lexical, syntactic, and semantic constraints (Massaro, 1987).

Although the FLMP has not explicitly addressed the general problem of word recognition, its principles can easily be extended to describe spoken word recognition. The additional assumptions needed would resemble the properties already discussed in other models. Perhaps the most compatible is a neighborhood activation model (Luce, 1986), in which word recognition reduces to

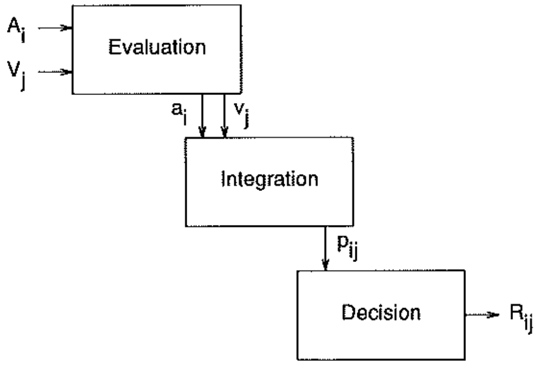

FIG. 14 Schematic representation of the three stages involved in perceptual recognition. The three stages are shown to illustrate their necessarily successive but overlapping processing. The sources of information are represented by capital letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j). These sources are then integrated to give an overall degree of support for a given alternative p_{ij}. The decision operation maps this value into some response R_{ij}, such as a discrete decision or a rating.

finding the best match in a set of activated word candidates (a process analogous to the decision process of the FLMP). This model is grounded in relative goodness-of-match (or activation) rather than absolute activation—again an important similarity between it and the FLMP.

H. Conclusion

We have reviewed seven current models of speech perception and word recognition. Models of speech perception are confronted with several characteristics of speech perception that are apparently easy for humans but difficult for models and machines. Some of these characteristics are listed in Table I. We now illustrate the simultaneous influence of multiple influences in a bimodal speech perception experiment.

VI. A PROTOTYPICAL EXPERIMENT

There is valuable and effective information afforded by a view of the speaker's face in speech perception and recognition by humans. Visible speech is particularly effective when the auditory speech is degraded because of noise, bandwidth filtering, or hearing impairment. As an example, the perception of short sen-

TABLE I
Aspects of Speech Perception That Cause Problems for Models but Not Listeners

Problem	FLMP solution
1. Context dependency of feature values. The value of a feature is not constant for different segments. For example, the voice onset time (VOT) for a voiced stop /gi/ is roughly equal to the VOT for the voiceless stop /pa/.	Prototypes represent V, CV, and VC syllables. The feature values for these units are relatively constant. A different VOT value can be specified for the syllables /gi/ and /pa/.
2. Spectral characteristics of speech segments are influenced by gender, age, rate of speaking, and background noise.	A prototype can consist of a disjunction of several descriptions of each syllable. There is evidence that a dozen or so descriptions could represent individuals of all ages and sexes, for example (Wilpon & Rabiner, 1985).
3. Characteristics of consonants change as a function of their vowel environment.	Same solution as 1.
4. The formant transitions of a CV vary as a function of the vowel preceding the CV.	This variation does not appear to be psychologically meaningful (Massaro & Oden, 1980).
5. Formant values for vowels change with the phonetic environment. For example, a lax front vowel (e.g., /ɪ/) changes dramatically when followed by /l/.	Same solution as 1.
6. Formant values for vowels depend on vowel duration.	There is experimental evidence that these two sources of information are processed independent of one another (Massaro, 1984).
7. Vowel duration is influenced by nonsegmental properties, such as rate of speaking, syntax, and semantics.	Parallel influences can be described by the fuzzy integration of the FLMP.

tences that have been bandpass-filtered improves from 23% correct to 79% correct when subjects are permitted a view of the speaker. This same type of improvement has been observed in hearing-impaired listeners and patients with cochlear implants (Massaro, 1987). The strong influence of visible speech is not limited to situations with degraded auditory input, however. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight (McGurk & MacDonald, 1976). If an auditory syllable /ba/ is dubbed onto a videotape of a speaker saying /da/, subjects often perceive the speaker to be saying /ða/.

To study how perceivers use both auditory and visual speech, we carried out an experiment manipulating auditory and visual information in a cross-linguistic speech perception task (Massaro, Tsusaki, Cohen, Gesi, & Heredia, 1993). Five levels of audible speech varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. The audible and visible speech also are presented alone, giving a total of 25 + 5 + 5 = 35 independent stimulus conditions. This test procedure is called an expanded factorial design.

A five-step auditory /ba/ to /da/ continuum was synthesized by altering the parametric information specifying the first 80 ms of the consonant–vowel syllable. Using an animated face, control parameters are changed over time to produce a realistic articulation of a consonant–vowel syllable. By modifying the parameters appropriately, a five-step visible /ba/ to /da/ continuum was synthesized. The presentation of the auditory synthetic speech was synchronized with the visible speech for the bimodal stimulus presentations. All the test stimuli were recorded on videotape for presentation during the experiment. Six unique test blocks were recorded with the 35 test items presented in each block. Subjects were instructed to listen and to watch the speaker, and to identify the syllable as either /ba/ or /da/.

The points in Figure 15 give the average results for a group of Spanish-speaking subjects in the experiment (all instructions and the test were in Spanish for these subjects). As can be seen in the figure, both the auditory and visual sources influenced identification performance. There was also a significant interaction because the effect of one variable was larger to the extent that the other variable was ambiguous. The lines give the predictions of the FLMP. This model is able to capture the results of several influences on identification performance. The FLMP also predicts the results of individual subjects.

It is important to evaluate the results of individual subjects because group results can be misleading (Massaro & Cohen, 1993). All the model tests were carried out on individual subjects. The points in Figure 16 give the mean proportion of identifications for a typical Japanese-speaking subject in the same experiment. The identification judgments changed systematically with changes in the audible and visible sources of information. The likelihood of a /da/ identification increases as the auditory speech changes from /ba/ to /da/, and analogously for the visible speech. Each source has a similar effect in the bimodal conditions relative to the corresponding unimodal condition. In addition, the influence of one source of information is greatest when the other source is ambiguous.

To describe the results, the important assumption of the FLMP is that the auditory source supports each alternative to some degree and analogously for the visual source. Each alternative is defined by ideal values of the auditory

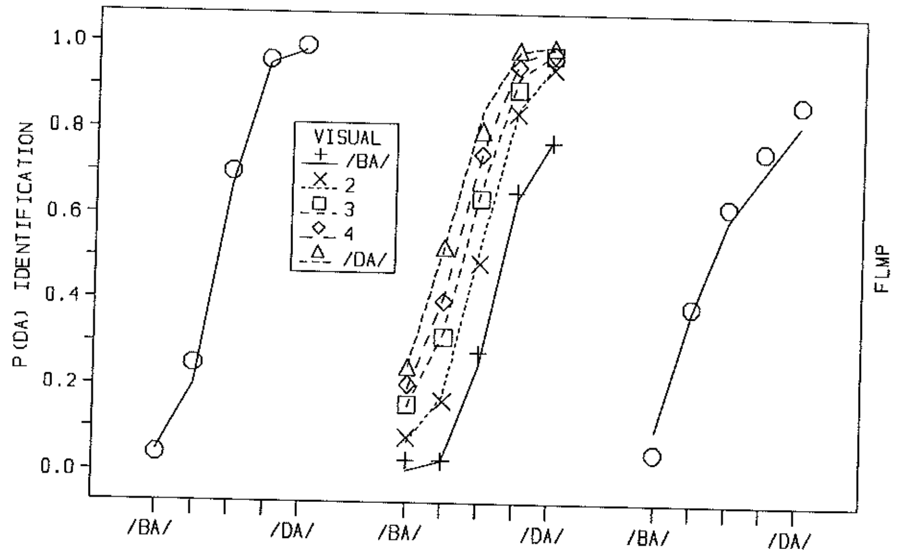

FIG. 15 Observed (*points*) and predicted (*lines*) proportion of /da/ identifications for the auditory-alone (*left*), the factorial auditory-visual (*center*), and visual-alone (*right*) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/ (Spanish-speaking subjects). The lines give the predictions for the FLMP (after Massaro et al., 1993).

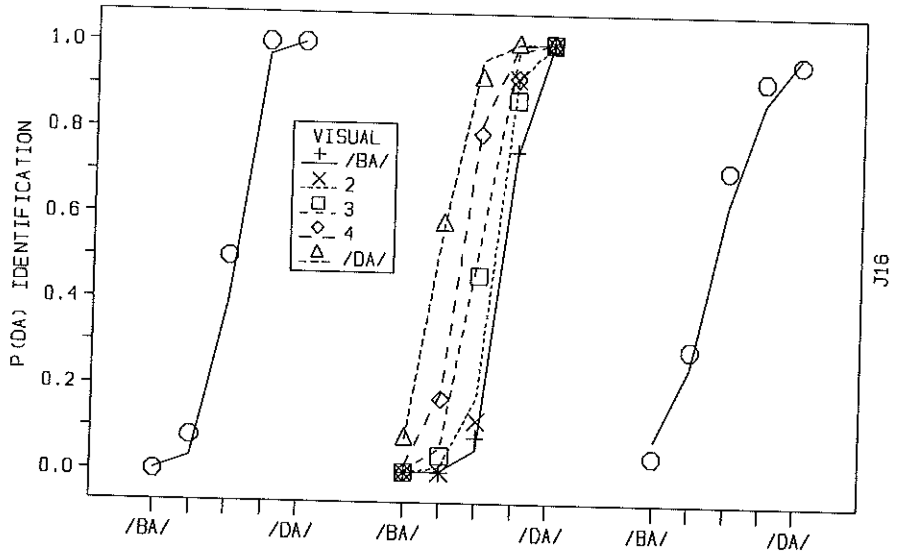

FIG. 16 The points give the mean proportion of /da/ identifications for a typical Japanese subject in the experiment as a function of the auditory and visual sources of information. The lines give the predictions of FLMP (after Massaro et al., 1993).

and visual information. Each level of a source supports each alternative to differing degrees represented by feature values. The feature values representing the degree of support from the auditory and visual information for a given alternative are integrated following the multiplicative rule given by the FLMP. The decision operation gives the response by determining the relative goodness-of-match of the relevant response alternatives. The formal model, as tested against the results, requires five parameters for the visual feature values and five parameters for the auditory feature values. The lines in Figures 15 and 16 give the predictions of FLMP. The model provides a good description of the identifications of both the unimodal and bimodal syllables.

VII. CONCLUSION

Our short review of research and theory indicates that speech perception might be best understood in terms of general perceptual, cognitive, and learning processes. The guiding assumption for the proposed framework is that humans use multiple sources of information in the perceptual recognition and understanding of spoken language. In this regard, speech perception resembles other forms of pattern recognition and categorization because integrating multiple sources of information appears to be a natural function of human endeavor. Integration appears to occur to some extent regardless of the goals and motivations of the perceiver. A convincing demonstration for this fact is the Stroop color-word test. People asked to name the color of the print of words that are color names printed in different colors become tongue-tied and have difficulty naming the colors. We cannot stop ourselves from reading the color word, and this interferes with naming the color of the print.

VIII. SYNTHETIC VISIBLE SPEECH AND COMPUTER ANIMATION

Given the importance of visible speech and the perceiver's natural ability to integrate multiple sources of information, our current research goal is to develop an animation system for visible speech synthesis. A critical assumption concerns the experimental, theoretical, and applied value of synthetic speech. Auditory synthetic speech has proven to be valuable in all three of these domains. Much of what we know about speech perception has come from experimental studies using synthetic speech. Synthetic speech gives the experimenter control over the stimulus in a way that is not always possible using natural speech. Synthetic speech also permits the implementation and test of theoretical hypotheses, such as which cues are critical for various speech distinctions. The applied value of auditory synthetic speech is apparent in the multiple everyday uses for text-to-speech systems for both normal and visually impaired individuals.

We believe that visible synthetic speech will prove to have the same value as audible synthetic speech. Synthetic visible speech will provide a more fine-grained assessment of psychophysical and psychological questions not possible with natural speech. For example, testing subjects with synthesized syllables intermediate between several alternatives gives a more powerful measure of integration relative to the case of unambiguous natural stimuli. It is also obvious that synthetic visible speech will have a valuable role to play in alleviating some of the communication disadvantages of the deaf and hearing-impaired.

A main objective of our research is to identify the facial properties that are informative by evaluating the effectiveness of various properties in a synthetic animated face. Analogous to the valuable contribution of using auditory speech synthesis in speech perception research, visible speech synthesis permits the type of experimentation necessary to determine (a) what properties of visible speech are used, (b) how they are processed, and (c) how this information is integrated with auditory information and other contextual sources of information in speech perception. Our experimental and theoretical framework has been validated in several domains (Massaro, 1987), and it is ideal for addressing these questions about facial information in speech perception.

One attractive aspect of providing or using audible and visible speech jointly is the complementarity of audible and visible speech. Visible speech is usually most informative for just those distinctions that are most ambiguous auditorily. For example, places of articulation (such as the difference between /b/ and /d/) are difficult via sound but easy via sight. Voicing, on the other hand, is difficult to see visually but is easy to resolve via sound. Thus, audible and visible speech not only provide two independent sources of information; these two sources are often productively complementary. Each is strong when the other is weak.

The development of a realistic, high-quality facial display provides a powerful tool for investigation of a number of questions in auditory-visual speech perception. The analysis of the articulation of real speakers guides our development of the visible speech synthesis. In addition, perception experiments indicate how well the synthesis simulates real speakers. We expect that the results of our research will also have applications in the area of automatic lip-reading to enhance speech recognition by machine. If perceivers achieve robust recognition of speech by using multiple sources of information, the same should be true for machine recognition.

One applied value of visible speech is its potential to supplement other (degraded) sources of information. Visible speech should be particularly beneficial in poor listening environments with substantial amounts of background noise. Its use is also important for hearing-impaired individuals because it allows effective communication within speech—the universal language of the community. Just as auditory speech synthesis has proved a boon to our visually impaired citizens in human–machine interaction, visual speech synthesis should prove to be valuable for the hearing-impaired. We expect that lip-reading a stylized representation of the face is also possible. The successful use of stylized representations would make animation much easier and would allow the visible speech to be transmitted over a low-bandwidth channel such as a telephone line.

Finally, synthetic visible speech is an important part of building synthetic "actors" (Thalmann & Thalmann, 1991). We can also be confident that synthetic visible speech will play a valuable role in the exciting new sphere of virtual reality.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon* Tech. Rep. No. 6. Bloomington: Indiana University, Research on Speech Perception.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69,* 548–558.

Marcel, A. J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology, 15,* 238–300.

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244,* 522–523.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition: A tutorial review. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 125–150). Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. In U. H. Frauenfelder & L. K. Tyler (Eds.), *Spoken word recognition* (pp. 71–102). Cambridge, MA: MIT Press.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

Massaro, D. W. (1975a). *Experimental psychology and information processing*. Chicago: Rand McNally.

Massaro, D. W. (Ed.). (1975b). *Understanding language: An information processing analysis of speech perception, reading, and psycholinguistics*. New York: Academic Press.

Massaro, D. W. (1979). Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 595–609.

Massaro, D. W. (1984). Time's role for information, processing, and normalization. *Annals of the New York Academy of Sciences, Timing and Time Perception, 423,* 372–384.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1988). Ambiguity in perception and experimentation. *Journal of Experimental Psychology: General, 117,* 417–421.

Massaro, D. W. (1989a). *Experimental psychology: An information processing approach*. San Diego, CA: Harcourt Brace Jovanovich.

Massaro, D. W. (1989b). Multiple book review of *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Behavioral and Brain Sciences, 12,* 741–794.

Massaro, D. W. (1989c). Testing between the TRACE model and the fuzzy logical model of perception. *Cognitive Psychology, 21,* 398–421.

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.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3, pp. 129–165). New York: Academic Press.

Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics, 21,* 445–478.

Mattingly, I. G., & Studdert-Kennedy, M. (Eds.). (1991). *Modularity and the motor theory of speech perception*. Hillsdale, NJ: Erlbaum.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1–44.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* (*London*), *264,* 746–748.

Miller, G. A. (1981). *Language and speech*. San Francisco: Freeman.

Morton, J. (1964). A preliminary functional model for language behavior. *International Audiology, 3,* 216–225.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165–178.

Morton, J., & Broadbent, D. E. (1967). Passive versus active recognition models, or is your hommunculus really necessary? In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 103–110). Cambridge, MA: MIT Press.

Pastore, R. E. (1987). Categorical perception: Some psychophysical models. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 29–52).

Peters, A. M. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition, 25,* 21–52.

Pollack, I., & Pickett, J. M. (1963). The intelligibility of excerpts from conversation. *Language and Speech, 6,* 165–171.

Salasoo, A., & Pisoni, D. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Cognition, 2,* 210–231.

Sampson, G. R. (1989). Language acquisition: Growth or learning? *Philosophical Papers, 18,* 203–240.

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

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Mechanisation of thought processes, Proceedings of a symposium held at the National Physical Laboratory on 24–27 November 1958* (pp. 511–526). London: H. M. Stationery Office.

Stevens, K. N. (1986). Models of phonetic recognition II: A feature-based model of speech recognition. In P. Mermelstein (Ed.), *Proceedings of the Montreal satellite symposium on speech recognition, Twelfth International Congress on Acoustics.*

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review, 77,* 234–249.

Thalmann, N. M., & Thalmann, D. (1991). *Computer animation '91.* Heidelberg: Springer-Verlag.

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34,* 409–420.

Tyler, L. K., & Wessels, J. (1985). Is gating an on-line task? Evidence from naming latency data. *Perception & Psychophysics, 38,* 217–222.

Uhlarik, J., & Johnson, R. (1978). Development of form perception in repeated brief exposures to visual stimuli. In R. D. Pick & H. L. Pick, Jr. (Eds.), *Perception and experience* (pp. 347–360). New York: Plenum.

Warren, P., & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics, 41,* 262–275.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167,* 392–393.

Warren, R. M. (1982). *Auditory perception: A new synthesis.* New York: Pergamon.

Warren, R. M., Bashford, J. A., & Gardner, D. A. (1990). Tweaking the lexicon: Organization of vowel sequences into words. *Perception & Psychophysics, 47,* 423–432.

Werker, J. (1991). The ontogeny of speech perception. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 91–109). Hillsdale, NJ: Erlbaum.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory and serial order in speech behavior. *Psychological Review, 76,* 1–15.

Wilpon, J. G., & Rabiner, L. R. (1985). A modified k-means clustering algorithm for use in speaker-independent isolated word recognition. *IEEE Transactions ASSP-33,* 587–594.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338–353.