Sawashima M, Hirose H 1983 Laryngeal gestures in speech production. In: MacNeilage P F (ed.) *The Production of Speech.* Springer-Verlag, New York

H. Hirose

The Encyclopedia of Language and Linguistics
Pergamon Press Ltd.
Headington Hill Hall
Oxford OX3 0BW
UK

1994

Speech Perception

This article gives a functional account of speech perception—how people discriminate and categorize the objects of spoken language. This functional account also includes the dynamics or time course of the processes taking the perceiver from spoken language to its understanding. Beginning with the issue of the functional units in speech perception, the article turns to a detailed discussion of categorical speech perception: a classic study to illustrate how speech perception is studied, with what appeared to be surprising results. These results motivated a theory that remained dominant for several decades. The article then gives a functional account of the results and develops a new theoretical framework, assesses theories of word recognition within a common framework, presents the state of the art in research and theory, and closes with some remarks about what remains to be learned about speech perception in the 1990s.

Speech perception is one of the most impressive demonstrations of auditory information processing. It can be described as a pattern-recognition problem. Given some speech input, the perceiver must determine which message best describes the input. An auditory stimulus is transformed by the auditory receptor system and sets up a neurological code, called a preperceptual auditory storage. This storage holds the information in a preperceptual form for roughly 250 msec, during which time the recognition process must take place. The recognition process transforms the preperceptual image into a perceptual experience, called a synthesized percept. One issue given this framework is, what are the patterns that are functional in the recognition of speech? These sound patterns are referred to as perceptual units.

1. Perceptual Units in Speech

One reasonable assumption is that every perceptual unit in speech has a representation in long-term memory, which is called a prototype. The prototype contains a list of acoustic features that define the properties of the sound pattern as they would be represented in preperceptual auditory storage. As each sound pattern is presented, its corresponding acoustic features are held in preperceptual auditory storage. The recognition process operates to find the prototype in long-term memory which best matches the acoustic features in preperceptual auditory storage. The outcome of the recognition process is the transformation of the preperceptual auditory image of the sound stimulus into a synthesized percept held in synthesized auditory memory. Fig. 1 presents a schematic diagram of the recognition process.

According to this model, preperceptual auditory storage can hold only one sound pattern at a time for a short temporal period. Recognition masking studies have shown that a second sound pattern can interfere with the recognition of

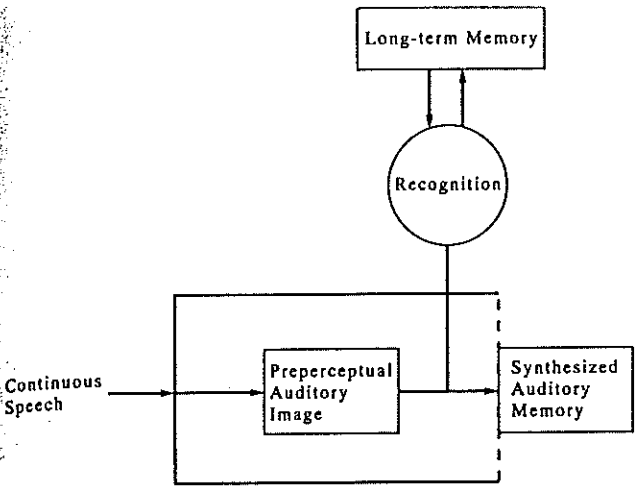

Figure 1. A schematic diagram illustrating the recognition process in speech transforming a preperceptual auditory image into synthesized auditory memory.

an earlier pattern if the second is presented before the first is recognized. Each perceptual unit in speech must occur within the temporal span of preperceptual auditory storage and must be recognized before the following one occurs for accurate speech processing to take place. Therefore, the sequence of perceptual units in speech must be recognized one after the other in a successive and linear fashion. Finally, each perceptual unit must have a relatively invariant acoustic signal so that it can be recognized reliably. If the sound pattern corresponding to a perceptual unit changes significantly within different speech contexts, recognition could not be reliable, since one set of acoustic features would not be sufficient to characterize that perceptual unit. Perceptual units in speech as small as the phoneme or as large as the phrase have been proposed.

1.1 Phonemes

The first candidate considered for the perceptual unit is the phoneme. Phonemes represent the smallest functional difference between the meaning of two speech sounds. Given the word *ten*, its meaning can be changed merely by changing the consonant /t/ to /d/. The two sounds form two different words when they are combined with *-en*; they are therefore different phonemes. On the other hand, sounds are said to be within the same phoneme class if substitution of one for the other does not change the meaning of the sound pattern. One example is the word *did*. The two *d's* in the word are not the same acoustically and, if their sound patterns were extracted and interchanged with each other, the word would not sound the same. Yet they are not functionally different since interchanging them should still give the word *did*. In this case, they are called different allophones of the same phoneme. Thus, if the substitution of one minimal sound for another changes the meaning of the larger unit, then the two sounds are phonemes. If such substitution does not change the meaning of the larger unit, then the different sounds are allophones of the same phoneme class.

Consider the acoustic properties of vowel phonemes. Unlike some consonant phonemes, whose acoustic properties change over time, the wave shape of the vowel is considered to be steady-state or tone-like. The wave shape of the vowel repeats itself anywhere from 75 to 200 times per second. In normal speech, vowels last between 100 and 300 msec, and during this time the vowels maintain a fairly regular and unique pattern. It follows that, by the above criteria, vowels could function as perceptual units in speech.

Now consider consonant phonemes. Consonant sounds are more complicated than vowels and some of them do not seem to qualify as perceptual units. It has been noted that a perceptual unit must have a relatively invariant sound pattern in different contexts. However, some consonant phonemes appear to have different sound patterns in different speech contexts. Fig. 2 shows that the stop consonant phoneme /d/ has different acoustic representations in different vowel contexts. Since the steady-state portion corresponds to the vowel sounds, the first part, called the transition, must be responsible for the perception of the consonant /d/. As can be seen in the figure, the acoustic pattern corresponding to the /d/ sound differs significantly in the syllables. Hence, one set of acoustic features would not be sufficient to recognize the consonant /d/ in the different vowel contexts. Therefore, linguists must either modify their definition of a perceptual unit or eliminate the stop consonant phoneme as a candidate.

1.2 CV Syllables

There is another reason why the consonant phoneme /d/ cannot qualify as a perceptual unit. According to the model perceptual units are recognized in a linear fashion. Research has shown, however, that the consonant /d/ cannot be recognized before the vowel is also recognized. If the consonant were recognized before the vowel, then it should be possible to decrease the duration of the vowel portion of the syllable so that only the consonant would be recognized. Experimentally, the duration of the vowel in the consonant-vowel syllable (CV) is gradually decreased and the subject is asked when she hears the stop consonant sound alone. The CV syllable is perceived as a complete syllable until the vowel is eliminated almost entirely (Liberman, et al. 1967). At that point, however, instead of the perception changing to the consonant /d/, a nonspeech whistle is heard. Liberman, et al. show that the stop consonant /d/ cannot be perceived independently of perceiving a CV syllable. Therefore, it seems unlikely that the /d/ sound would be perceived before the vowel sound; it appears, rather, that the CV syllable is perceived as an indivisible whole or gestalt.

These arguments lead to the idea that syllables function as perceptual units rather than containing two perceptual units each. One way to test this hypothesis is to employ the CV syllables in a recognition-masking task. Liberman, et al., found that subjects could identify shortened versions of the CV syllables when most of the vowel portion is eliminated. Analogous to interpretation of vowel perception, recognition of these shortened CV syllables also should take time. Therefore, a second syllable, if it follows the first soon enough, should interfere with perception of the first. Consider the three CV syllables /ba/, /da/, and /ga/ (/a/ pronounced as in *father*), which differ from each other only with respect to the consonant phoneme. Backward recognition masking, if found with these sounds, would demonstrate that the consonant sound is not recognized before

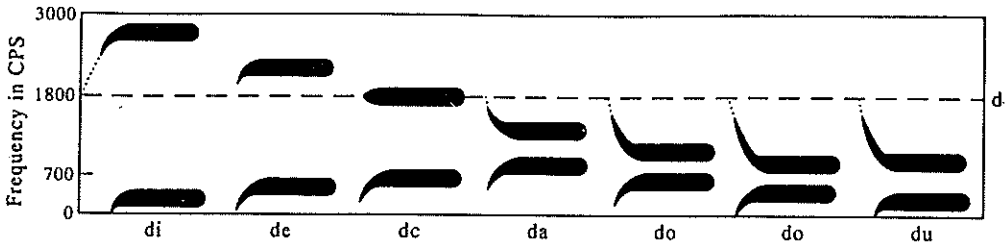

Figure 2. Second-formant transitions appropriate for /d/ before various vowels.

the vowel occurs and also that the CV syllable requires time to be perceived.

1.3 Recognition Masking

Newman and Spitzer (1987) conducted such an experiment, employing as test items three synthetic CV syllables /ba/, /da/, /ga/, each 40 msec long with 20 msec of that duration comprising the transition and the remainder the steady-state vowel. The masking stimulus was a 40 msec steady-state vowel /a/.

Figure 3 shows the percentage of correct recognitions for 8 observers as a function of the silent interval between the test and masking CVs. The results show that recognition of the consonant is not complete at the end of the CV transition, nor even at the end of the short vowel presentation. Rather, correct identification of the CV syllable requires perceptual processing after the stimulus presentation. These results support the hypothesis that the CV syllable must have functioned as a perceptual unit, because the syllable must have been stored in preperceptual auditory storage, and recognition involved a transformation of this preperceptual storage into a synthesized percept of a CV unit. The acoustic features necessary for recognition must, therefore, define the complete CV unit. An analogous argument can be made for VC syllables also functioning as perceptual units.

It is also necessary to ask whether perceptual units could be larger than vowels, CV, or VC syllables. George Miller argued that the phrase of two or three words might function as a perceptual unit. According to the above criteria for a perceptual unit, it must correspond to a prototype in long-term memory which has a list of features describing the acoustic features in the preperceptual auditory image of that perceptual unit. Accordingly, preperceptual auditory storage must last on the order of one or two seconds to hold perceptual units of the size of a phrase. But the recognition-masking studies usually estimate the effective duration of preperceptual storage to be about 250 msec. Therefore, perceptual units must occur within this period, eliminating the phrase as the perceptual unit.

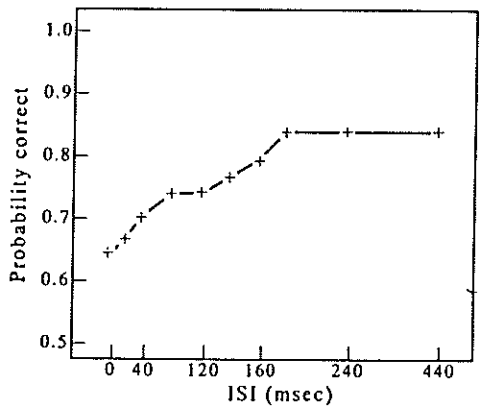

Figure 3. Probability of correct recognitions of the test CV syllables as a function of the duration of the silent intersyllable interval (SII) in a backward recognition-masking task (results of Newman and Spitzer 1987).

The recognition-masking paradigm developed to study the recognition of auditory sounds has provided a useful tool for determining the perceptual units in speech. If preperceptual auditory storage is limited to 250 msec, the perceptual units must occur within this short period. This time period agrees nicely with the durations of syllables in normal speech.

2. Categorical Perception

One persistent and popular belief is that speech is perceived categorically. In fact, the study of speech perception has almost been synonymous with the study of how it is perceived categorically. Perception is said to be categorical if the subject can only make judgments about the name of a stimulus, not its particular sound quality. For example, the same speaker may repeat the same syllable a number of times. The acoustic patterns representing this syllable would differ from each other since a speaker cannot repeat the same sound exactly. A listener who perceives the sounds categorically would not be able to discriminate any difference in the particular sound quality of each repetition of the syllable. The same listener, on the other hand, would be able to recognize a difference between any of these sounds and another syllable spoken by the same speaker. In categorical perception, the listener can recognize differences when the syllables have different names but not when they have the same name.

Subjects are certainly not limited in this way in the processing of nonspeech. They are able to discriminate two tones as different even though they can not differentially label them. This is true for all sound dimensions: subjects can discriminate many more differences than they can identify successfully. This phenomenon, in fact, was one of the observations that convinced George Miller that, although, subjects can make many discriminations along a unidimensional stimulus continuum, they can identify accurately about 7 ± 2 of these stimuli. In this case, discrimination is not limited by identification, since subjects can discriminate differences along a stimulus continuum which they cannot identify absolutely.

2.1 A Seminal Study

A seminal study established the experimental paradigm for the study of categorical speech perception. Liberman, et al. (1957) used synthetic speech to generate a series of 14 CV

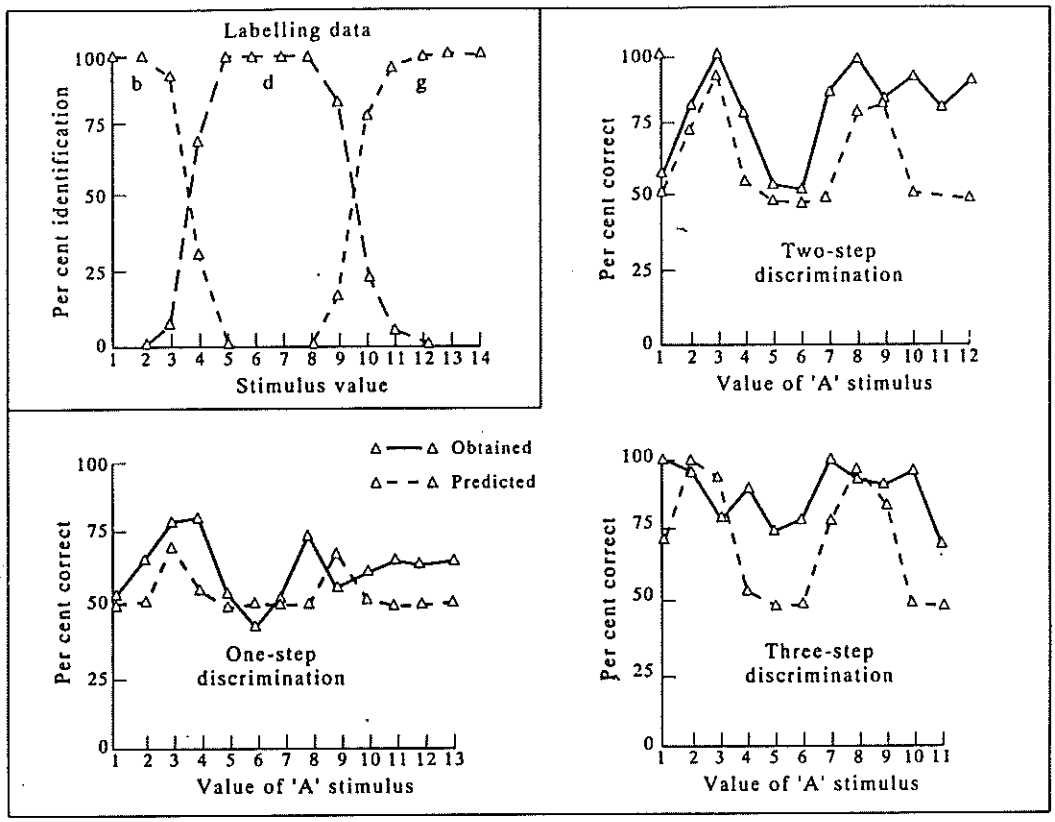

Figure 4. Probability of identification in the labeling task (top left panel) and observed and predicted probability of discrimination in the ABX task (after Liberman, et al., 1957).

syllables as in Fig. 4, i.e., where the second formant (F2) onset was varied in 120 Hz steps and the steady-state vowel was /e/. Consistent with prior studies, listeners identified these stimuli as /be/, /de/, or /ge/ as shown in Fig. 4a, which gives the results for one subject. They then tested listeners' ability to discriminate pairs of these stimuli using the ABX method, i.e., where three stimuli were presented in the order ABX: A and B differ and X was identical to either A or B; listeners had to indicate whether X was identical to A or B. This judgment was supposedly one of auditory discrimination since the subjects were instructed to use 'any cues' they could hear. The results of the same subject's discrimination of stimuli pairs immediately adjacent on the stimulus continuum ('one-step'), as well as those that were 2 steps and 3 steps apart are shown in Figs. 4b–d. The authors also compared such discrimination scores with those predicted on the basis of the identification functions. The authors concluded that discrimination was fairly well predicted by identification and thus that speech was perceived categorically. This view is still widely held today.

But with the advantage of hindsight, this conclusion can be criticized on at least two points. First, the ABX paradigm may encourage the verbal encoding of the stimuli A and B since it would be very difficult to remember their sound quality. Therefore, subjects might simply be performing the ABX discrimination task as if it were an identification task and thus it should not come as a surprise if subjects show poor discrimination of different syllables that have the same label. Second, quantitative measures of the 'goodness-of-fit' of the predicted and obtained discrimination functions are not very impressive; in fact discrimination was generally better than that predicted and this is evident in the data presented in Fig. 4.

2.2 Negative Evidence

Subsequently experimental evidence has been obtained to show that listeners can discriminate auditory differences between stop consonants that are given the same label in identification.

Using a F2-onset stimulus continuum similar to that in Fig. 4, but with the V = /æ/, Barclay (1972) first obtained listeners' identification of the initial consonants as /b/, /d/, or /g/, and got results similar to that in Fig. 4a. He then later asked the same subjects to listen to the same continuum but reduced the eligible responses to /b/ and /g/. These subjects were successful in differentiating stimuli they had earlier assigned to the /d/ category, that is, now assigning either to the /b/ or the /g/ category.

Pisoni and Lazarus (1974) also demonstrated that subjects could discriminate between stimuli that they would give the same name to by first providing subjects more extensive training on the items in the stimuli continuum and second by employing a discrimination method that did not tax auditory memory as much. They present two pairs of stimuli, one pair always being the same and the other different; subjects simply had to indicate which pair had different stimuli. Their results show that discriminating between sounds that are usually given different names was not significantly better than that between sounds usually given the same name.

There have been many demonstrations of continuous perception of speech since these initial studies. Without a doubt, the task of the speech perceiver is to categorize. The child must decide whether the adult said, *Get the ball* or *Get the doll.* However, the decision appears to be based on continuous information provided by the speech signal. As in other domains of categorization, speech recognition involves the evaluation and integration of continuous, not categorical, features. It is particularly important that the information is maintained in a noncategorical form because it can then be supplemented with other types of information. If the child had insufficient acoustic information to distinguish between *ball* and *doll*, a nod or hand gesture by the speaker toward one of the objects could help disambiguate the instruction.

2.3 Categorical Partition

It is still a common mistake to interpret categorization behavior as evidence for categorical perception. 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 Fig. 5 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 this value. Otherwise, the pattern is classified as *not A*. Given this decision rule, the probability of an *A* response would take the step-function form shown in the right panel of Fig. 5. 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 occur given continuous information.

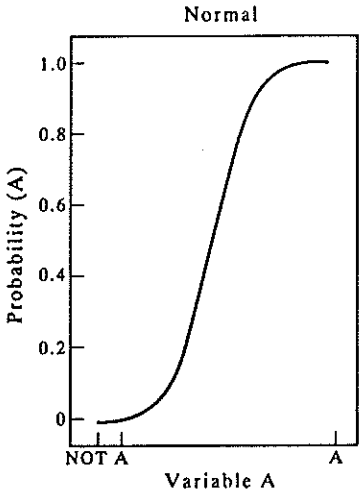

Figure 6. Probability(*A*) as a function of Variable *A* given the linear relationship between GOODNESS(A) and Variable *A* and the decision criterion represented in Fig. 5, but with normally distributed noise added to the mapping of Variable *A* to GOODNESS(A).

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 the levels away from the criterion value. Figure 6 illustrates the expected outcome for identification if there is normally distributed noise with the same criterion value assumed in Fig. 5.

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

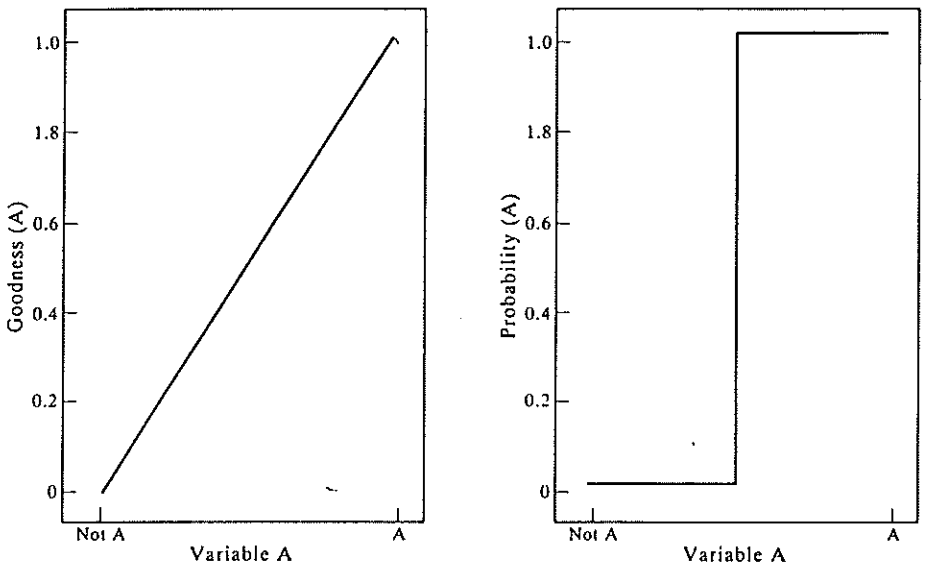

Figure 5. Left Panel: The degree to which a stimulus represents the category *A*, called GOODNESS(A) as a function of the level along a stimulus continuum between *not A* and *A*. Right Panel: The probability of an *A* response, Probability(*A*), as a function of the stimulus continuum if the subject 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.

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 example shows that categorical decisions made on the basis of continuous information produce identification functions with sharp boundaries, previously taken to represent 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.

Drawing upon a broad range of methodological, theoretical, and experimental issues, an attempt has been made to present the evidence against the theory of categorical perception. At the methodological level, it has been shown that the relation between identification and discrimination provides no support for categorical perception. First, the categorical model usually provides an inadequate description of the results, and it has not been shown to provide a better description than alternative models. Second, even if the results were to provide unequivocal support for the categorical model, other explanations than categorical perception would be possible.

At the theoretical level, it is necessary to distinguish between sensory and decision processes in the categorization task. What is central is that decision processes can transform continuous sensory information into results usually taken to reflect categorical perception. Finding relatively categorical partitioning of a set of stimuli in no way implies that these stimuli were perceived categorically. Tapping into the process in other ways 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 easily discriminate among different exemplars of the same speech category. In addition, *RT*s of identification judgments illustrate that members within a speech category vary in ambiguity or the degree to which they represent the category.

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 experiments described above. 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 a rising from a falling transition more easily than discriminate 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 be due to a fundamental characteristic of auditory processing and would not be a result of having speech categories. Thus similar results would be found in humans, chinchillas, and monkeys and for nonspeech analogs. However, it is important to note that discrimination between instances within a category is still possible. Although a weak form of categorical perception might exist for a few distinctions, most distinctions do not appear to have this property, and the linguist is left with explaining continuous rather than categorical speech perception.

3. Theories of Word Recognition

Although there are several 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. Seven models of word recognition will be discussed to highlight some important issues in understanding how words are recognized. Several important characteristics of the models will be reviewed to contrast and compare the models. Figure 7 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 only to 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 to consider is whether the word recognition process functions autonomously, or whether it is context dependent.

3.1 Logogen Model

The logogen model described by Morton (1964) 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 genus, meaning 'birth.' 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 8 gives a schematic diagram of the logogen model.

Morton's logogen model can be evaluated with respect to the five characteristics shown in Fig. 7. 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

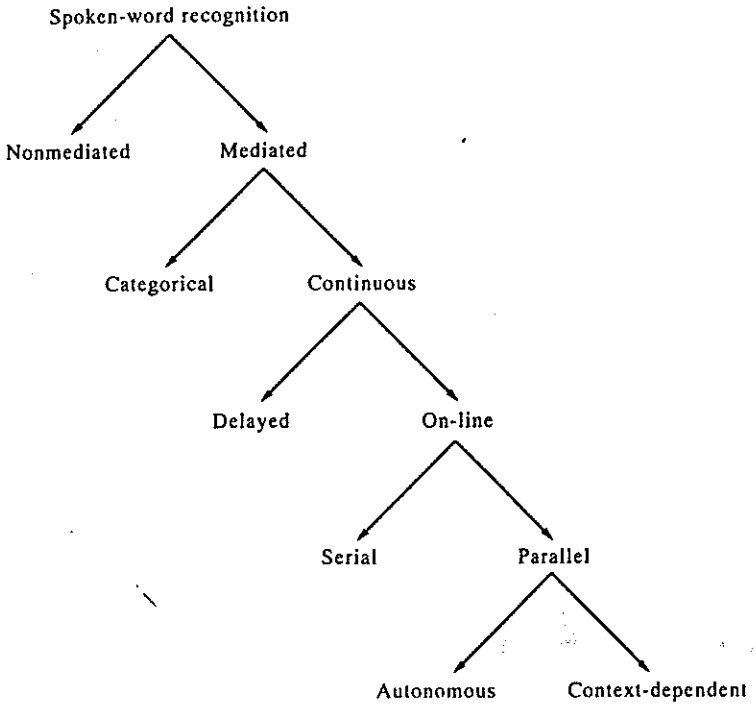

Figure 7. Tree of wisdom illustrating binary oppositions central to the differences among theories of spoken-word recognition.

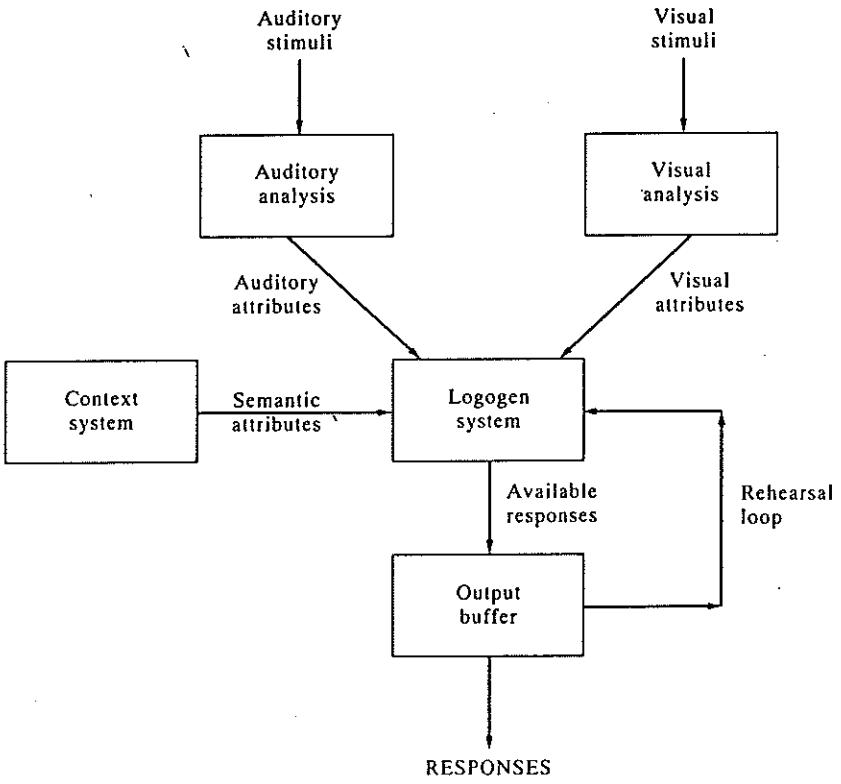

Figure 8. A schematic diagram of the logogen model. Recognition occurs when the activation in a logogen exceeds a critical level and the corresponding word becomes available as a response.

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 Fig. 8, 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 its nonmediated nature. Thus, the model has difficulty explaining intermediate recognition of sublexical units (e.g. CV syllables) and how nonwords are recognized.

3.2 Cohort Model

An 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 of 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 9 gives an example illustrating how the cohort model recognizes the word 'elephant.'

The cohort model is easy to describe with respect to the five characteristics in Fig. 7. The model is mediated, categorical, on-line, parallel, and contextually 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 is that phonemes are not perceptual units and that speech perception is not categorical.

/ɛ/	/ɛl/	/ɛlə/	/ɛləf/	/ɛləfə/
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		
ecstasy	elephant	elocution		
.	elephantine	eloquent		
.	elevate	———		
element	elevation	(12)		
elephant	.			
elevate	.			
.	———			
.	(28)			
entropy				
entry				
.				
.				
extraneous				
.				
———				
(324)				

Figure 9. 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/.

3.3 TRACE Model

The TRACE model of speech perception (McClelland and 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 Fig. 7, 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 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 and Elman 1986: 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.

3.4 Autonomous-search Model

A fourth model of word recognition is an autonomous-search model of word recognition. 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 and Forster 1987). For ease of presentation, the model will be presented in terms of recognizing a written word.

The first stage in processing a written stimulus is in terms of 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 Fig. 7. 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, and final recognition of a word requires a serial search. All of this processing goes on without any influence from the context at other levels, such as the sentence level. The autonomous-search model appears to fail on at least two counts: categorical perception and contextually independent processing. Evidence for continuous perception has been reviewed and there is convincing evidence for the influence of context in word recognition (see Sect. 4).

3.5 Lexical Access from Spectra (LAFS) Model

Klatt (1979) developed a LAFS (lexical access from spectra) 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 large decoding network of expected sequences of spectra. Figure 10 illustrates how each word is first represented phonemically, then all possible pronunciations are determined by phonetic recording rules specifying alternative pronunciations within and across word boundaries, and these phonetic representations are converted to sequences of spectral templates like those shown in Fig. 11. Figure 10 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 is the assumption 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 Fig. 7, 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 contribution of linguistic context.

3.6 LAFF Model

Stevens Kenneth has articulated a model describing lexical access via acoustic correlates of linguistic binary phonetic features. These features are language universal and binary (present or absent). Table 1 gives a featural representation of the word 'pawn.'

This model is driven by parsimony in that the features are assumed to be binary and robust. Binary features allow the integration process to by shortcircuited in that multiple ambiguous sources of information do not have to be combined. With respect to the five characteristics in Fig. 7, 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. 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 com-

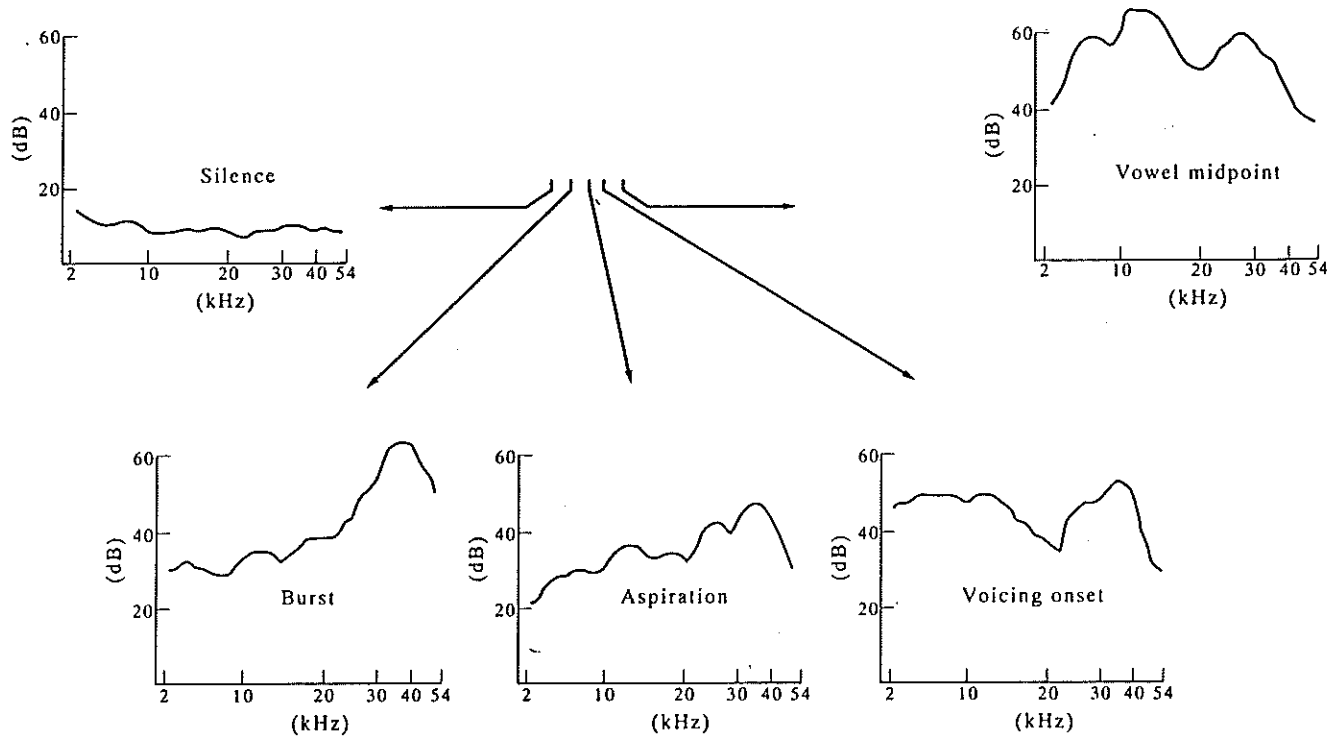

Figure 10. The phonetic transition from the middle of [ta] to the middle of [a] has been approximated by a sequence of five static critical-band spectra (after Klatt 1979).

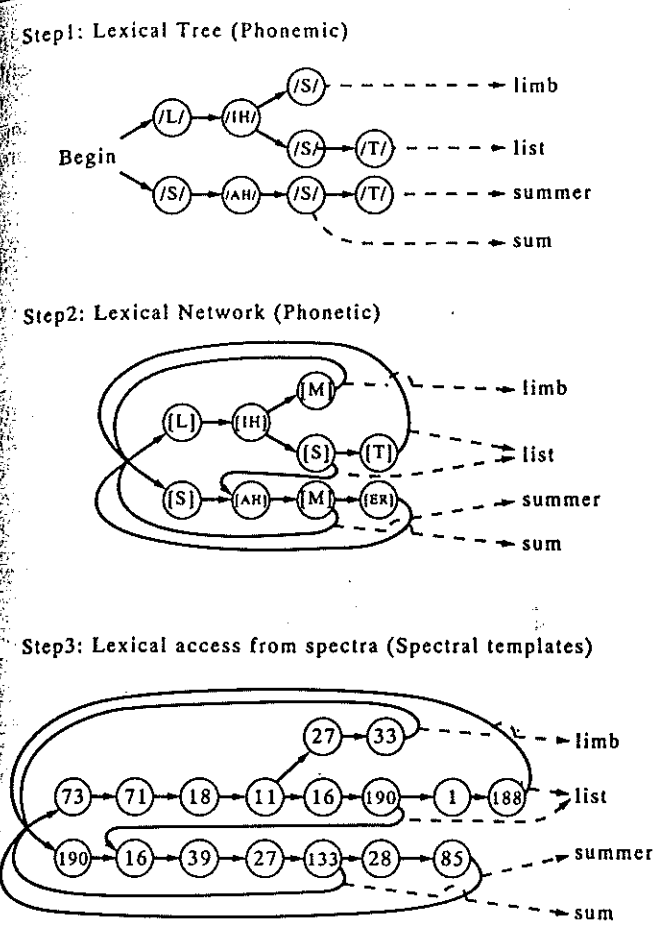

Figure 11. The lexical tree, lexical network, and lexical access from spectra of the LAFS model (after Klatt 1979).

plete word has been presented. Finally, the contextual dependencies built into the represention are phonologically-based and, therefore, there is no prescribed provision for linguistic constraints.

3.7 Fuzzy Logical Model of Perception (FLMP)

According to the fuzzy logical model of perception (FLMP), well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns (Massaro 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 description 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. However, the memory representation must be compatible with the sensory representation

Table 1. A conventional lexical representation for the English word 'pawn,' shown at the top, has been modified below 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. (Klatt (1989).

Conventional lexical representation

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

Modified lexical representation

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

resulting from the transduction of the 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, the linguist might envision activation of all prototypes corresponding to the perception 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 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 0.5

corresponds to a completely ambiguous situation whereas 0.7 would be more true than false and so on. Fuzzy truth values, therefore, not only can represent continuous rather than just categorical information, they also can 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 of 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. 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 12 illustrates the three stages involved in pattern recognition.

Different sources of information are represented by uppercase letters. The evaluation process transforms these into psychological values (indicated by lowercase letters) that are then integrated to give an overall value. The decision operation maps this value into some response, such as 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 Fig. 7, spoken-word recognition is mediated, continuous, on-line, serial and parallel, and both autonomous and context-dependent.

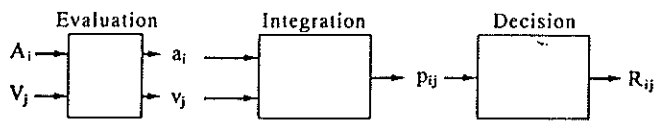

Figure 12. Schematic representation of the three operations involved in perceptual recognition.

The theoretical framework of the FLMP has proven valuable 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; they 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).

Seven models of speech perception and word recognition have been reviewed. All such models have trouble dealing with aspects of speech that are apparently easily dealt with by humans. In short, it is difficult to accommodate the extreme contextual variation in speech. The same acoustic feature is sometimes associated with one phoneme, sometimes another, e.g. the voice onset time (VOT) for the voiced stop in the syllable /gi/ is roughly equivalent to the VOT of the voiceless stop in /pa/. On the other hand, the same phoneme has many different acoustic characteristics depending on its context, e.g., the formant values for the high front lax vowel /I/ change dramatically when followed by the lateral /l/. Conceivably, human listeners solve this by storing a much larger inventory of phonological units than has previously been entertained, i.e., separate prototypes for each V, CV, and VC sequence.

4. Linguistic Context

There is considerable debate concerning how informative the acoustic signal actually is. Even if the acoustic signal is sufficient for speech recognition under ideal conditions, 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.

4.1 Detecting Mispronunciations

Abstracting meaning is a joint function of the independent contributions of the available perceptual and contextual information. 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 percent with increases in the number of feature changes, which reflects 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.

The detection of mispronunciation technique was also used to demonstrate that additional higher-order contextual redundancy is also used in perception, e.g. more accurate detection of mispronunciation in the word 'killer' if a prior sentence included the word 'murder.'

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 msec, 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.

The field of speech perception is an area of research rich in both theories and methodologies, and developments in these two domains feed each other. This happens for several reasons: what is valued more than one experimental methodology that yields reliable results relevant to a given theoretical issue is two or more methodologies whose results show convergence. Furthermore, as new methods emerge and yield results new questions arise which often lead to new theories. It is not possible in the space of this article to cover all the issues and methods that occupy researchers in this area. However, a brief case study of the interaction of one particular methodology and certain points of theory may be illustrative.

4.2 Limitations of Results

Perceivers have been shown to be efficient exploiters of different types of context to aid in speech perception. The autonomous-search, the LAFS, and the LAFF models have difficulty in accounting for the contribution of context because it assumes that speech perception goes on without any help of context. Even these models are not necessarily falsified by the context effects, however, because it can be claimed that the context effects that were observed occurred after speech perception. It might be argued, 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 helped address this issue and, more importantly, the results can be used to reveal how stimulus information and context jointly contribute to word recognition.

4.3 Gating Task

In the gating task, portions of the spoken message are eliminated or gated out. In a typical task with single words, only the first 50 msec 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 msec. Subjects attempt to name the word after each presentation. Warren and Marslen-Wilson (1987), for example, presented words such as 'school' or 'scoop.' Figure 13 shows that the probability of correct recognition of a test word increases as additional word information is presented in the gating task.

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 potentially controversial feature of the 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. 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.

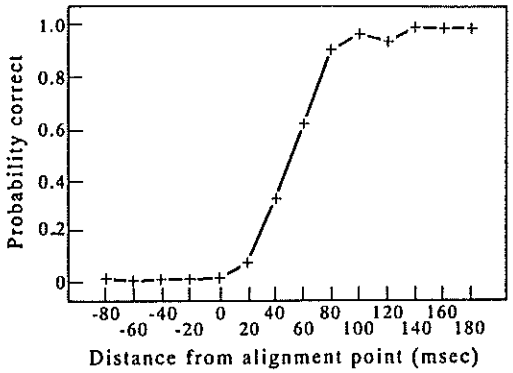

Figure 13. Probability of correct recognition of the test word as a function of the distance from the alignment point in the test word. The alignment point corresponds to a point near the onset of the final consonant of the word (results adapted from Warren and Marslen-Wilson 1987).

The standard multiple presentation format has been compared with the format in which subjects heard only a single fragment from each word in the task. Similar results were found in both conditions. A similar study found that the average duration of the test word needed for correct identification was only 5 msec 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 to increase 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 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.

4.4 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 msec 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 14 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 Fig. 14 give the predictions of the fuzzy logical model of perception (FLMP). 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 3 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 Fig. 14 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, which assumes that some minimum cohort set must be

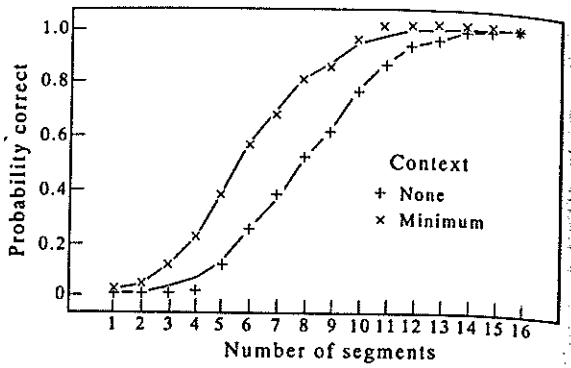

Figure 14. Observed (points) and predicted (lines) probability of identifying the test word correctly as a function of the sentential context and the number of segments of the test word. The minimum context refers to minimum semantic and weak syntactic constraints. The none context refers to no semantic and weak syntactic constraints (after Tyler and Wessels 1983).

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 fit 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).

The functional account of speech perception should continue to be the goal for speech research in the 1990s. It has been seen how perceivers have continuous rather than categorical information from the speech signal in speech perception. There is also good evidence that both the speech signal and the linguistic and contextual context influence speech perception. Given these multiple influences, speech perception necessarily involves the evaluation and integration of a variety of somewhat ambiguous information sources. A decision process is also required to use the outcome of these processes in an optimal fashion. The nature and dynamics of these processes offer an immediate challenge to researchers and students.

See also: Phonetics, Articulatory; Quantal Theory of Speech; Speech: Biological Basis; Speech Perception, Direct Realist Theory; Speech Processing: Auditory Models; Speech: Acoustics; Phonetics, Descriptive Acoustics; Intelligibility and Speech Evaluation; Speech Technology: Overview.

Bibliography

Barclay J R 1972 Non-categorical perception of a voiced stop: A replication. *Perception and Psychophysics* **11**: 269–73

Bradley D C, Forster K I 1987 In: Frauenfelder U H, Tyler L K (eds.) *Spoken Word Recognition.* MIT Press, Cambridge, MA

R A 1973 Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics* 13: 153–56
D H 1979 Speech perception: A model of acoustic–phonetic analysis and lexical access. *JPhon* 7: 279–312
D H 1989 Review of selected models of speech perception. Marslen-Wilson W D (ed.) *Lexical Representation and Process*. MIT Press, Cambridge, MA
A L, Harris K S, Hoffman H S, Griffith B C 1957 The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54: 358–68
Marslen-Wilson W D 1973 Linguistic structure and speech shadowing at very short latencies. *Nature* 244: 522–23
Marslen-Wilson W D 1984 Function and process in spoken word-recognition. In: Bouma H, Bouwhuis D G (eds.) *Attention and Performance. Vol. X: Control of Language Processes*. Lawrence Erlbaum, Hillsdale, NJ
Marslen-Wilson W D 1987 Functional parallelism in spoken word recognition. In: Frauenfelder U H, Tyler L K (eds.) *Spoken Word Recognition*. MIT Press, Cambridge, MA
Massaro D (ed.) 1975 *Understanding Language. An Information-processing Analysis of Speech Perception, Reading, and Psycholinguistics*. Academic Press, New York
Massaro D W 1987 *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum, Hillsdale, NJ
McClelland J L, Elman J L 1986 The TRACE model of speech perception. *Cognitive Psychology* 18: 1–86
Morton J A 1964 A preliminary functional model for language behavior. *International Audiology* 3: 216–25
Newman C W, Spitzer J B 1987 Monotic and dichotic presentation of phonemic elements in a backward recognition-masking paradigm. *Psychological Research* 49: 31–36
Pisoni D B, Lazarus J H 1974 Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America* 55: 328–33
Tyler L K, Wessels J 1983 Quantifying contextual contributions to word-recognition processes. *Perception and Psychophysics* 34: 409–20
Warren P, Marslen-Wilson W D 1987 Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics* 41: 262–75

D. W. Massaro

Speech Perception: Direct Realist Theory

The direct realist theory of speech perception has been developed to explain a listener's ability to recover phonetic and phonological information from a speech utterance. It makes two distinctive theoretical claims. First, perception is 'direct,' in being unmediated either by an internal representation of stimulation constructed in perception or by such processes as inference-making and hypothesis-testing, presumed necessary in some theories to deal with apparent discrepancies between information about phonetic properties in the acoustic speech signal and the ostensibly richer phonetic message recoverable by listeners. Second, the theory is a 'realist' theory because of its claim that linguistic objects of speech perception occur as real physical events in the world, outside the minds of perceivers and talkers. That is, in the theory, consonants and vowels are physical actions of the vocal tract that have linguistic significance because of the functions they serve in communicative exchanges among members of a language community. Information in an acoustic speech signal can specify consonants and vowels to listeners for two reasons. The consonants and vowels directly cause disturbances in the air that serve as stimulation for the ear and hence as information for the perceiver, and the disturbances that each perceptually distinguishable phonetic property of an utterance causes are specific to that property.

The theory addresses 'bottom-line' perceiving in the following sense. The claim is not that consonants and vowels are *the* perceptual objects of speech; rather, perceivers recover a tiered set of objects, generally focusing their attention on the larger domains that convey a communicative message. The claim, instead, is that the gestures which constitute consonants and vowels are the smallest perceivable linguistic units in speech, because they are the smallest linguistically significant actions of the vocal tract that cause disturbances in the air. They are perceptual objects for at least one group of listeners, namely language learners, who must learn to do with their vocal tracts what they hear skilled talkers doing with theirs.

The claims of the theory are controversial. Whereas, in a direct realist theory, perceptual objects are the articulatory causes of the acoustic signal, in most theories (except the motor theory of speech perception), perceptual objects are acoustic structures mapped onto abstract phonetic and/or phonological categories stored in the mind of a language user. Whereas, in a direct realist theory, objects of speech perception can at once be physical actions of the vocal tract and components of a linguistic message, in other theories, neither the physical actions of the vocal tract nor the acoustic signal are composed of units of a linguistic message themselves; rather, they signal linguistic units, while the units themselves are categories in the mind. Finally, claims of a direct realist theory that perception can be direct are generally considered *un*realistic, because, ostensibly, the acoustic speech signal does not specify the consonants and vowels of a spoken message, and because listeners sometimes report hearing sequences of consonants and vowels other than those produced by a speaker.

The theory of speech perception derives from a larger, universal theory of perception developed by James J. Gibson. Understanding why the theory takes the form that it does, therefore, requires one to look outside the particular domain of a theory of speech perception to the larger context of the general theory of perception in which it is imbedded.

1. Perception in General, and Gibson's 'Ecological' Theory

Perception constitutes the only means by which animals, including humans, can come to know the environment in which they participate as actors. Crossmodally, perception is acquaintance with environmental objects, surfaces, and events by means of the causal effects that they have on media that can stimulate the sense organs of a perceiver.

Consider visual perception in its public aspect. (Public aspects of perceiving refer to those things in the environment that are perceivable by a particular species of perceiver, what information about the environment is available in stimulation at the sense organs, and what subset of that information is perceptually effective. Private or covert aspects of perceiving refer to any processes inside perceivers