A Stage Model of Reading and Listening

Visible Language, XII 1 (Winter 1978), pp. 3-26.
Author's address: Department of Psychology, University of Wisconsin, Madison, WI 53706
0022-2224/78/0100-0003$02.00/0 ©1978 Visible Language
Visible Language, Box 1972 CMA, Cleveland, OH 44106

Language processing is the abstraction of meaning from a physical signal such as a printed text or sequence of speech sounds. The goal of an information-processing model is to describe how language is processed, not simply what the reader or listener must know to understand language. Language processing is viewed as a sequence of internal processing stages or operations that occur between the language stimulus and meaning. The operations of a particular stage take time and transform the information in some way, making the transformed information available to the following stage of processing. In the present model the storage component describes the nature of the information at a particular stage of processing whereas the functional component describes the operations of a stage of processing. The information-processing model is used heuristically to incorporate data and theory from a variety of studies of language processing.

I. Introduction

Does it matter that I wrote this contribution rather than spoke it? Or does it make a difference that you are reading rather than listening? Or are you in fact, not only reading the article but simultaneously hearing it being spoken by the little homunculus in your head? Regardless of the modality of the input, this special journal issue does not offer convenient solutions to these and other important problems in communication. We will, however, present some recent research and theory on the psychological processes involved in listening to speech and reading printed text. Our goal is to stimulate your interest and involvement in our study.

One of the persistent questions about understanding language is whether the modality of input (what might be called the true surface structure) makes a difference. The first answer that comes to mind is why should it. The purpose of all language is to communicate and understand a message (or for some to camouflage a message). Language production and understanding processes must solve the same problems in both visible and audible language. Although reading and listening may have developed independently, the processes involved may still be analogous in the same

way that analogous biological processes have developed in convergent evolution. It is commonly accepted that two organisms may develop similar solutions to the problem of survival even though they evoled independently of one another. As an example, the eye of the octopus (a cephalopod) and the eye of man (a mammal) function very similarly although they evolved completely independently of one another (Blakemore, 1977). Following this logic, the assumption that reading and listening can be viewed as similar processes does not necessitate an assumption of a common phylogenetic or ontogenetic evolution. Given that reading and listening solve the same problem, it is not unreasonable to assume that they are analogous rather than hierarchical solutions to the problem. Wrolstad (1976) presents a similar argument and supporting evidence.

Recent research on the processing of manual-visual languages such as American Sign Language (ASL) indicates that analogous solutions to language understanding extend beyond reading and listening. On every dimension that has been explored, remarkable parallels have been found between understanding signs and understanding speech. Lane, Boyes-Braem, and Bellugi (1976) found that perceptual confusions among signs can be described utilizing a distinctive feature system, analogous to systems developed for perceptual confusions in speech. There is also evidence that grammatical structure in sign language plays the same functional role that it does in spoken language (Tweney, Heiman, and Hoemann, 1977). These results support the claim that the processes of language understanding are relatively general and abstract—not tried uniquely to the input modality. The work on ASL encourages the belief that there are similar and analogous processes in all forms of language understanding.

Although it might seem reasonable to assume that understanding spoken and written language exploits similar or analogous comprehension processes and structures, the early stages of decoding the input should reveal some basic differences. This follows from the fact that modality-specific processes are necessary to transform the sound vibrations of speech and the light waves of print. Several other obvious differences come to mind. Spoken language comes in one ear and goes out the other, whereas the print remains available at the beck and call of a regressive eye movement or a flip of the page. It is true that some compulsive listener might record the message and capitalize on the rewind and play option for particularly difficult sections of a spoken message. In the information processing model presented here, however, we will draw similarities between even the earliest modality-specific stages of language processing. Returning to our argument of convergent evolution, it is not unreasonable that the same or analogous processes are exploited for decoding spoken and written language.

II. Information-Processing Model

Reading and listening can be defined as the abstraction of meaning from printed text and from speech, respectively. To derive or arrive at meaning from a spoken or written message requires a series of transformations of the energy signal arriving at the appropriate receptors. Language processing can be studied as a sequence of processing stages or operations that occur between the energy stimulus and meaning. In this framework, language processing can be understood only to the extent that each of these processing stages is described. In a previous effort an information-processing model was utilized for a theoretical analysis of speech perception, reading, and psycholinguistics (Massaro, 1975b). The model was used heuristically to incorporate data and theory from a variety of approaches to the study of language processing. The model should be conceptualized as an organizational structure for the state of the art in language processing. In this paper I will present a general overview of the information-processing model, and use the model to describe and incorporate some recent research.

Figure 1 presents a flow diagram of the temporal course of reading and listening. At each stage the system contains storage and functional components. The storage component represents the information available at a particular stage of processing. The functional component specifies the procedures and processes that operate on the information held in the corresponding storage component. The model distinguishes four functional components: feature detection, primary recognition, secondary recognition, and rehearsal-recoding. The corresponding storage component represents the information available to each of these stages of processing.

III. Feature Detection and Primary Recognition

The feature detection process transforms the energy pattern created by the language stimulus and transduced by the appropriate receptor system into a set of features held in preperceptual storage. Primary recognition evaluates and integrates these features into a percept which is held in synthesized memory. In speech, for example, the changes in sound pressure set the eardrums in motion and these mechanical vibrations are transduced into a set of neural impulses. It is assumed that the signal in the form of continuous changes in vibration pattern is transformed into a set of relatively discrete features. Features do not have to be relatively primitive such as the amount of energy in a particular frequency band, but they may include information about the direction and rate of frequency change. It would be possible, for example, to have a feature detector that responds to the rising first formant transition that is characteristic of the class of voiced stop consonants.

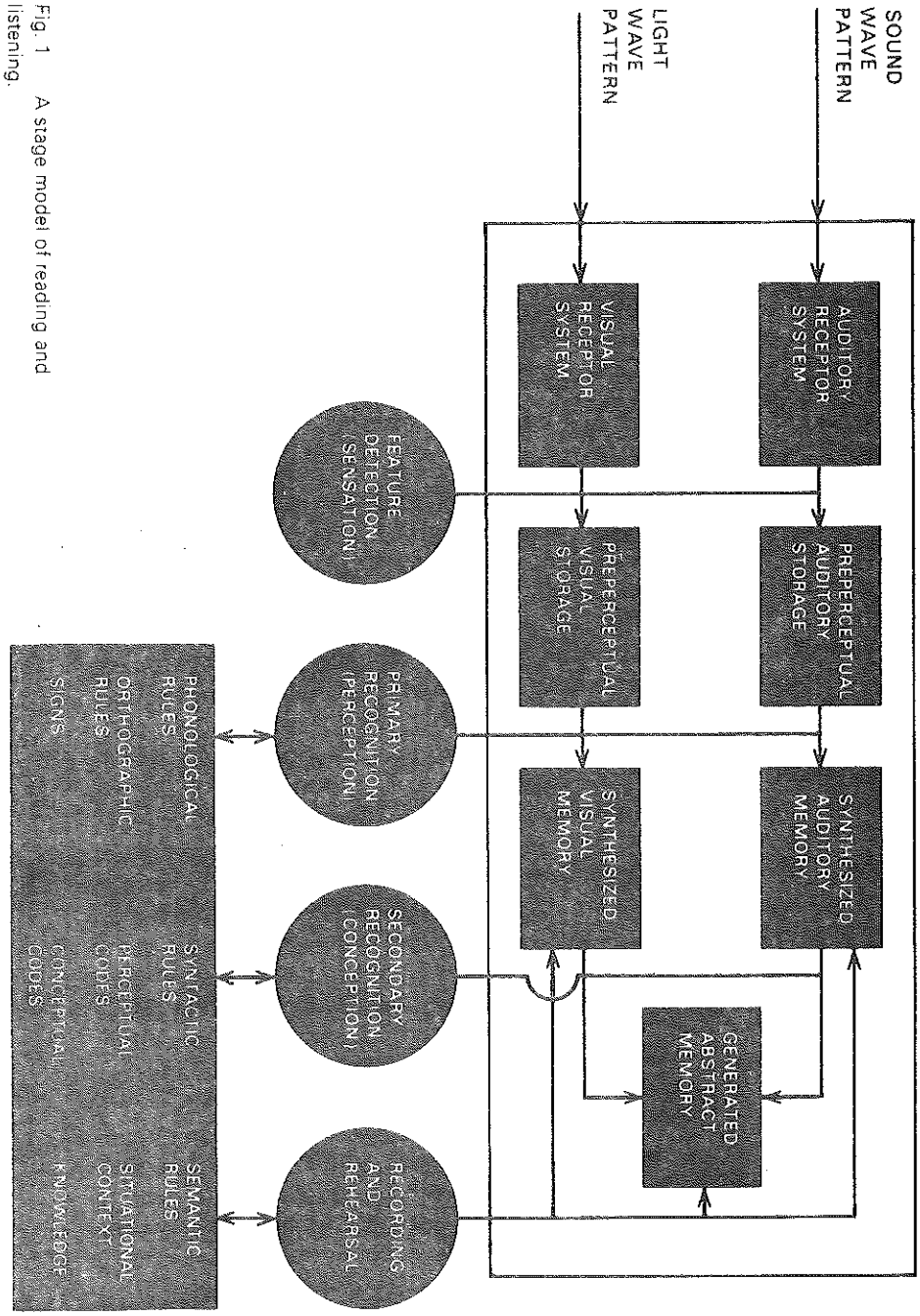


Fig. 1 A stage model of reading and listening.

A. Audible features

One traditional concern in speech research has been to determine the acoustic features that are utilized in perception. In terms of our model the feature detection process places features in a brief temporary storage called preperceptual auditory storage (PAS), which holds information from the feature detection process for about 250 msec. The primary recognition process integrates these features into a synthesized percept which is placed in synthesized auditory memory. One critical question is what features are utilized and a second important question is how are all of the features integrated together. Does the listener only process the least ambiguous feature and ignore all others, or are the features given equal weight, and so on? Despite the overwhelming amount of research on acoustic features, very little is known about how the listener puts together the multitude of acoustic features in the signal in order to arrive at a synthesized percept.

The integration of acoustic features has not been extensively studied for two apparent reasons. The first is that research in this area was highly influenced by linguistic descriptions of speech sounds in terms of binary all-or-none distinctive features (Jakobson, Fant, & Halle, 1961). One of the goals of distinctive feature theory was to describe all of the functional differences among speech sounds by a minimal number of distinctive features of the language. Therefore, distinctive features were designed to be general: if a distinctive-feature difference distinguished two phonemes in the language, that same distinction would also distinguish several other phoneme pairs. Given the distinctive feature of voicing, for example, the distinction of voiced versus voiceless can account for the differences between /z/ and /s/, /v/ and /f/, /b/ and /p/, and so on. The integration of information from two or more binary dimensions is a trivial problem. Integrating binary features from voicing and place of articulation, for example, could be carried out by simple logical conjunction. If the consonant /b/ were represented as voiced and labial and /p/ were represented as voiceless and labial, the identification of voiced labial sound would be /b/ whereas the identification of a voiceless labial sound would be /p/.

A second reason for the neglect of the integration problem is methodological. The primary method of study involved experiments in which the speech sound was varied along a single relevant dimension. In a typical study of voicing all voicing cues were made neutral except one, such as voice onset time and then this dimension was varied through the relevant values. Similarly, place of articulation was studied by neutralizing all cues but one, and then varying the remaining dimension through the appropriate values. Very few experiments independently varied both voicing cues and place cues within a particular experiment so that little information was available about how these cues were integrated into a synthesized percept.

More recently, we have initiated a series of experiments that are aimed more directly at the study of the integration of acoustic features in speech perception (Massaro & Cohen, 1976; Oden & Massaro, 1977). In contrast to the traditional linguistic description, we assume that the acoustic features held in preperceptual auditory storage (PAS) are continuous, so that a feature indicates the degree to which the quality is present in the speech sound. Rather than assuming that a feature is present or absent in PAS, it is necessary to describe a feature as a function of its degree of presence in PAS. This assumption is similar to Chomsky and Halle's (1968) distinction between the classificatory and phonetic function of distinctive features. The features are assumed to be binary in their classificatory function, but not in their phonetic or descriptive function. In the latter, features are multivalued representations that describe aspects of the speech sounds in the perceptual representation. Similarly, Ladefoged (1975) has also distinguished between the phonetic and phonemic level of feature description. A feature describing the phonetic quality of a sound has a value along a continuous scale whereas a feature classifying the phonemic composition is given a discrete value. In our framework the continuous features in PAS are transformed into discrete percepts in synthesized auditory memory (SAM) by the primary recognition process.

Given this theoretical description, acoustic features in PAS must be expressed as continuous values. That is to say, the listener will be able to hear the degree of presence or absence of a particular feature, even though his judgment in a forced choice task will be discrete. Oden and Massaro (1977) have used this description to describe acoustic features as fuzzy; that is to say, varying continuously from one speech sound to another. In this representation features are represented as fuzzy predicates which may be more or less true rather than only absolutely true or false (Zadeh, 1971). In terms of the model, fuzzy predicates represent the feature detection and evaluation process; each predicate is applied to the speech sound and specifies the degree to which it is true that the sound has a relevant acoustic feature. For example, rather than assuming that a sound is voiced or voiceless, the voicing feature of a sound is expressed as a fuzzy predicate.

$$P(\text{voiced}(S_{ij})) = .65 \qquad (1)$$

The predicate given by Equation 1 represents the fact that it is .65 true that speech sound S_{ij} is perceived to be voiced. In terms of our model, then, the feature detection process makes available a set of fuzzy predicates at the level of PAS. In addition to being concerned with the acoustic features in preperceptual storage this analysis of the feature evaluation process makes apparent that an important question in speech perception research is how the various continuous features are integrated into a synthesized percept.

As an example of the study of acoustic features, consider the dimension of voicing of speech sounds. In English the stops, fricatives, and affricates can be grouped into cognate pairs that have the same place and manner of articulation but contrast in voicing. The question of interest is what acoustic features are responsible for this distinction and how the various features are integrated together in order to provide the perceptual distinction. The integration question has not been extensively studied, however, since the common procedure in these experiments is to study just a single acoustic feature at a time. Consider two possible cues to the voicing distinction in stop consonant syllables: voice onset time (VOT), the time between the onset of the syllable and the onset of vocal cord vibration, and the fundamental frequency (F_0) of vocal cord vibration at its onset. Each of these cues has been shown to be functional in psychophysical experiments when all other cues have been held constant at neutral values. However, it is difficult to generalize these results to the perception of real speech, since no information is provided about the weight that these features will carry when other features are also present in the signal. To overcome this problem it is necessary to independently vary two or more acoustic features in the signal. The results of this type of experiment not only provide information about the cue value of one feature when other features are present in the signal, but also allow the investigator to evaluate how the various acoustic features are combined into an integrated percept. (For a further discussion see Massaro & Cohen, 1976, 1977; Oden & Massaro, 1977).

B. Audible features in fluent speech

The success of finding acoustic features in perception of isolated speech sounds might lead one to expect that perception of fluent speech is a straightforward process. Sound segments could be recognized on the basis of their features and the successive segments could be combined into higher-order units of words, phrases, and sentences. However, the acoustic structure of words in fluent speech differ significantly from the same words spoken in isolation. Two sources contribute to the large variation of words in fluent speech: coarticulation and psychological parsimony (Cole, & Jakimik, 1977; Ross, 1975).

In fluent speech the speech articulators must assume an ordered series of postures corresponding to the intended sounds, and the articulators cannot always reach their intended targets because of the influence of adjacent movements. Coarticulation refers to altering the articulation of one sound because of neighboring sounds. The words *did* and *you* spoken as /dId/ and /ju/ in isolation will be articulated as /dIdʒ u/ in combination because of palatalization. The alveolar stop followed by a front glide when combined produce the front-palatal affricate /dʒ /, even though a word boundary intervenes. Psychological

parsimony, sometimes called laziness (Ross, 1975), refers to the minimization of effort when we speak (Lieberman, 1967; Ross, 1975). Extending our example, *did you* can be further modified to give /dɪdʒə / or just /dʒə / in the utterance *Did you want to go?* Therefore, we get the message when a close friend asks /dʒə wanə go/ or even /jə wanə go/?

Luckily, the speaker is not only lazy but also intelligent. One anticipates the linguistic competence of the listener and the contextual constraints in the message (Lieberman, 1967). For example, a speaker will usually tend to give the listener a better acoustic signal for words that have high information content. Lieberman (1963) asked listeners to identify words excised from continuous speech. Identification was good to the degree that the excised word was unpredictable in the original utterance. The word "nine" was recognized about twice as often when it was excised from the sentence, "The number you will hear is nine," than when it was taken from "A stitch in time saves nine." If a word is not highly predictable from context, the speaker compensates by providing the listener with a better acoustic signal. Umeda (1977) measured the temporal properties of consonant sounds in 20 minutes of speech. Content words had longer durations than function words, and she interprets these results in terms of the high information value of content relative to function words.

C. Visible features

One of the oldest areas of reading-related research is the study of the functional cues used in recognizing printed characters. Much of this work was performed by typographers and artists concerned with the relative merits of good design and good legibility in type fonts (Spencer, 1968). Although many of the early conclusions remain valid today, they are almost totally ignored in the contemporary study of letter recognition. The primary influence in extant studies has been the well-known neurophysiological findings that the responses of cells in the visual cortex demonstrate an amazing stimulus selectivity (Hubel & Wiesel, 1962). For example, there appear to be specialized detectors in the visual system for lines of specific size and orientation (see Blakemore, 1973, and Lindsay & Norman, 1977, for intelligible reviews of this work).

Consistent with an all-or-none response of specialized cells, psychological descriptions have centered around binary all-or-none features. Feature sets usually consist of the presence or absence of horizontal, vertical, or oblique lines, curves, intersections, angles, and so on. The feature sets are typically derived from and tested against the recognition confusions of upper-case letters (see Massaro, 1975b, chapter 6). In contrast to the idea of binary all-or-none features, however, visible features, like audible features, may be fuzzy. Rather than a feature being present or absent, the information in preperceptual visual storage (PVS) could represent

the degree to which a given feature is present in the signal.

Given the idea of fuzzy information, it is important to carry out research that manipulates the degree to which a feature is present in a letter. Recently Blesser and his colleagues have developed and studied ambiguous characters (Blesser, Shillman, Kuklinski, Cox, Eden, and Ventura, 1974; Shillman, Cox, Kuklinski, Ventura, Blesser, and Eden, 1974). A completely ambiguous character is one that would be assigned to either letter class with equal probability. As an example, a *v* can be gradually transformed into a *y* by continuously increasing the right oblique line below the intersection (Naus & Shillman, 1976). This work with ambiguous characters is consistent with the idea of fuzzy visible features, since each feature must be represented in terms of the degree to which it is present in the character. Analogous to the recent work in speech perception, the theoretical notion of fuzzy information and the experimental methods of factorial designs and functional measurement techniques should advance the study of visible features in reading.

D. Visible features in printed text

Javal (cited in Huey, 1908) showed that reading is much easier when the top half rather than the bottom half of a line of print is exposed. There are seven ascending letters and five descending letters and, in addition, ascending letters are about five times more frequent in text than are descending letters (Mayzner and Tresselt, 1965). Therefore, it is not surprising the top half is more important than the bottom half of a line of print. It would be interesting to repeat Javal's experiment with a type font that equates the vertical extent for all of the letters. This font was actually developed by Andrew Tuer in the 1880's (cited in Spencer, 1968) and is only now being used in some computer printouts of lowercase type.

Are words perceived by way of the letters that make them up? This old and familiar question addresses whether word recognition can be described in terms of component letter recognition or whether a word is recognized on the basis of supraletter features without reference to the letters that make it up. If words are recognized via the letters that make them up, Cosky (1976) reasoned that the ease of word recognition should be a direct function of the ease of recognition of the component letters. He measured the time to name letters presented alone and the time to discriminate between letters to create an index of letter legibility. Then he composed words of the 13 most legible letters and the 13 least legible letters. Employing these words in a naming task, he found no effect of letter legibility and concluded that letter perception does not mediate word perception. However, in addition to interpreting differences in naming times as a direct index of differences in recognition time there are a number of limitations to Cosky's study. Most importantly, Cosky (1976) did not

show that his grouping of letters in terms of their legibility could predict performance on multi-letter items that are not words. If the uniqueness of words is responsible for Cosky's negative finding, then positive results should occur in nonword strings. Until this is demonstrated, Cosky's results can only be taken as a failure to find proof for the letter-mediation model; it cannot be taken as disproof.

In a well-known experiment carried out by Reicher (1969), subjects presented with either a single letter, a four-letter word, or a four-letter nonword flashed in a tachistoscope had to report what they saw. Reicher's contribution to this century-old task was to constrain the subject's choice by presenting two letter alternatives after each trial. Both alternatives would complete the display spelling words in the word condition so that performance on the word trials would not benefit from a simple guessing strategy. Even with these constraints, Reicher found a 10% advantage for recognition of a letter in a word over recognition of a letter in a nonword or a letter presented alone. These results have been described both in terms of whole-word and single-letter perceptual units. In terms of a perceptual unit the size of a word, it has supraletter features such as overall word shape which facilitate direct contact with the appropriate memory representation. Words are recognized better than letters or nonwords because the unique visual features of a word allow for easier recognition than the features of a single letter or a nonword.

The advantage of words over single letters and nonwords is not incompatible with the idea that the letter is a basic perceptual unit, however (Massaro, 1975b). In the present model the primary recognition process operates on a number of letters in parallel. The visual features read out at each letter position define a candidate set of possible letters for that position. The recognition process is not limited to featural information, but can also utilize knowledge about the orthographic structure of English spelling. The letter that is synthesized at each position, therefore, will not only correspond to the visual information that is available from feature detection and evaluation, but will also correspond to the orthographic constraints in the language. For example, consider the case in which the subject is given the lowercase string *coig* and has resolved just the circular envelope of the first letter and all of the last three letters. Given that *c, e,* and *o* are the only letters that are consistent with the circular envelope, these are the only possible letters at this position. If the reader further assumes that the string must conform to English orthography, only *c* is possible since the strings *ooig* and *eiog* are illegal English spellings. In this case the reader can synthesize *coig* since it is the only valid alternative. When the single letter *c* is presented, on the other hand, the perception of the envelope does not allow an unambiguous choice among *c, e,* and *o.* Accordingly, the reader is less likely to synthesize the correct alternative and will be correct only one out of three times. Although a

word is recognized via its component letters, familiarity with the orthographic structure of words facilitates primary recognition of its letters relative to a single-letter or nonword.

IV. Secondary Recognition

Secondary recognition transforms synthesized percepts into meaningful forms in generated abstract memory. In speech perception it is assumed that the input is analyzed syllable by syllable for meaning. In reading letter sequences are closed off in word units. In both cases the secondary recognition process makes the transformation from percept to meaning by finding the best match between the perceptual information and the lexicon in long-term memory. Each word in the lexicon contains perceptual and conceptual codes. The concept recognized is a function of at least two independent sources of information: the perceptual information in synthesized memory and the semantic/syntactic context in the message.

A. Perceptual and contextual contributions to listening

Our conceptualization of speech processing is one that is perceptually, and, therefore, acoustically driven. We assume that the secondary recognition process operates syllable by syllable on the output of primary recognition. However, contextual constraints also exert a strong influence at this stage of processing, so that both contributions must be accounted for in describing how meaning is imposed on the spoken message. A series of recent studies has shown that abstracting meaning is a joint function of the 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 one-feature mispronunciation was .3 whereas a four-feature change was recognized with probability .75. This result makes apparent the contribution of the perceptual information passed on by the primary recognition process. In our view some of the mispronunciations went unnoticed because the contribution of contextual information worked against the recognition of a mispronunciation. The syntactic/semantic context of the story would support a correct rendering of the mispronounced word, outweighing the perceptual information. In support of this idea all mispronunciations were correctly recognized when the syllables were isolated and removed from the passage.

Cole and Jakimik (1977) reasoned that the listener should be faster at detecting a mispronunciation to the extent that a word is predicted by its preceding context. This follows from the idea that the quickest way to detect a mispro-

nunciation is first to determine what the intended word should be 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 msec faster detecting mispronunciations in highly predictable relative to unpredictable words.

In other experiments Cole and Jakimik (1977) 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 test 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.

A second paradigm that has been used to study speech processing is the shadowing task, in which the listener repeats back the message as it is heard. It is well-known that shadowing performance improves with increases in the syntactic/semantic constraints in the message (Rosenberg & Lambert, 1974; Treisman, 1965). Recent research has been directed at how these higher-order constraints are integrated with the ongoing perceptual analyses in order to arrive at the meaning of the message. Marslen-Wilson (1973) asked subjects to shadow 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. When subjects made errors in shadowing, the errors were 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." In this example, *that* shares acoustic information with *at* and is also syntactically/semantically appropriate in the same position in the sentence.

In another experiment (Marslen-Wilson, 1975) subjects shadowed sentences that had one of the syllables mispronounced in a three-syllable word. Subjects never restored the word, that is, repeated back what should have been said when the mispronunciations occurred in the first syllable. With mispronunciations in the second syllable and third syllable, a significant proportion of restorations occurred. If the mispronounced word was syntactically and semantically anomalous, however, restorations did not occur for any mispronounced syllable. These results indicate that restorations will not occur if the shadower does not have sufficient acoustic information and syntactic/semantic context to make the restoration appropriate. If con-

text were the exclusive and overriding factor, we might expect subjects to replace the syntactically-semantically anomalous word with the appropriate word. This did not occur, however, showing that both context and acoustic information influenced speech processing.

Marslen-Wilson and Welsh (1978) asked observers to shadow (repeat back) spoken passages from a popular novel. The words of the passage were read to the subjects at a rate of 160 words per minute. The subjects were told to repeat back exactly what they heard. At random throughout the passage, common three-syllable words were mispronounced. When the words were mispronounced, only a single consonant phoneme was changed to a new consonant phoneme. The new phoneme differed from the original by one or three phonemic distinctive features, based on Keyser and Halle's (1968) classification system. Independently of the degree of feature change the changes could occur in the first or third syllable of the three-syllable word. Finally, the mispronounced words were either highly predictable or unpredictable given the preceding portion of the passage. Subjects were not told that words could be mispronounced although they probably became aware of this early in the experiment. All subjects shadowed at relatively long delays greater than 600 msec. The primary dependent measure in the task was the percentage of fluent restorations, that is, the proportion of times the shadowers repeated what should have been said rather than what was said. About half of the mispronounced words were restored; the restorations were made on-line with an average latency, and the shadowing was not disrupted. (When the mispronunciation was repeated exactly, i.e., not restored, shadowing was disrupted and response times increased.)

The change in the percentage of restorations as a function of the three independent variables in Marslen-Wilson and Welsh's study can illuminate how acoustic information and high-order context are integrated by the listener in language processing. Figure 2 presents the observed results in terms of the percentage of fluent restorations. All three variables influenced the likelihood of a restoration: shadowers were more likely to restore a one-feature than a three-feature change, a change in the third than in the first syllable, and a change in a highly predictable than in an unpredictable word.

Marslen-Wilson and his colleagues interpret this series of experiments as evidence against serial theories of language processing, which assume that "varying degrees of delay before information at any one level of analysis can interact with information at a higher level" (Marslen-Wilson and Tyler, 1975, p. 784). However, the results *do* show exactly such a delay. Restorations seldom occur when the first syllable is mispronounced by three features even though the word is relatively probable given the preceding context. This means that some low-level perceptual analyses of the word occurred regardless of the higher-order constraints available and then

Fig. 2 Predicted and observed percentage of fluent restorations as a function of the amount of feature change, the syllable, and the contextual constraint of the mispronunciation (observed data from Marslen-Wilson and Welsh, in press).

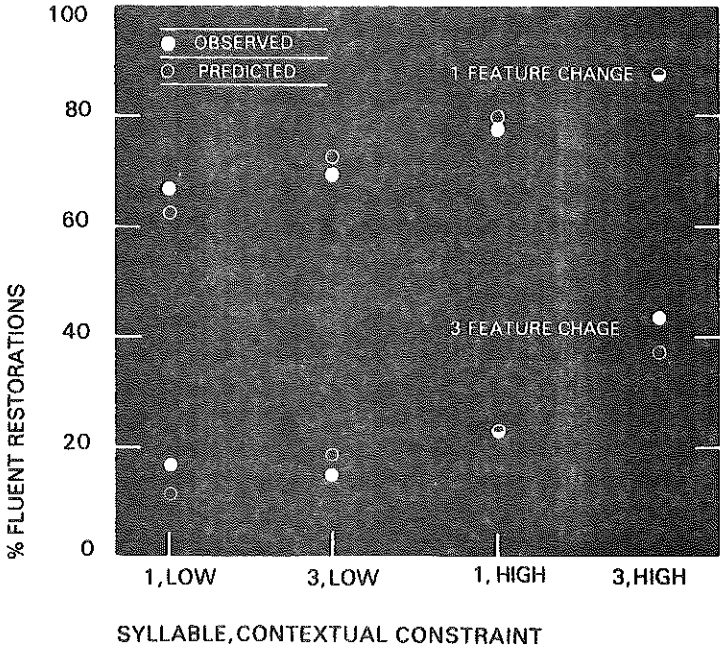
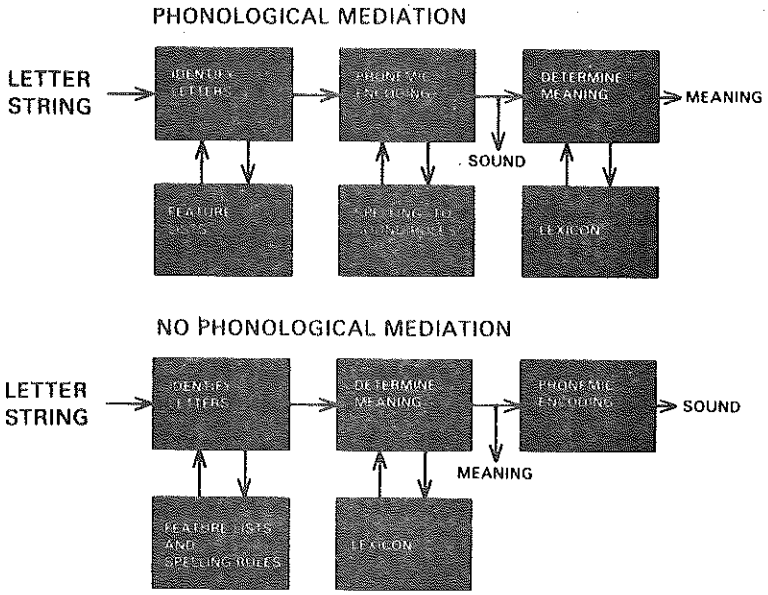


Fig. 3 Two stage models of the role of phonological mediation in reading.



the outcomes of these analyses were combined with the higher-order constraints. The fact that higher-order constraints in the passage influence shadowing does not mean that some analyses do not begin before others. More importantly, their view might be interpreted to mean that higher-order analyses modify the output of lower-level analyses. However, a quantitative model that assumes that both levels of analyses are functionally independent can accurately describe the results of their experiments. Figure 2 also presents the predictions of a quantitative formulation of the independence model (see Massaro, 1977, for the exact form of the model). The model assumes that the information passed on by the feature detection and evaluation process is equivalent regardless of the higher order constraints in the message. Therefore, it is not necessary to assume that higher-order constraints allow the subject to selectively attend to or selectively process certain acoustic properties of the speech input. In this model higher-order constraints do not modify the nature of low-level perceptual analyses performed on the input data.

B. Phonological mediation in reading

A persistent question in reading-related research is the extent to which the reader translates print into some kind of speech code before meaning is accessed. A similar but not identical question is the extent to which the speech code is *necessary* for the derivation of meaning. Figure 3 presents two extreme answers to the phonological mediation question. In the first model letters are identified and mapped into a speech code using spelling-to-sound rules, and meaning is determined on the basis of the derived speech code. In the second model meaning is determined from the letter resolution, and a speech code is not made available until after meaning has been accessed.

Gough and Cosky (1976) believe that they have accumulated some new data in support of the phonological-mediation view of Gough (1972). Subjects were asked to read aloud as rapidly as possible words that violated or obeyed spelling-to-sound rules. If phonological mediation occurs, regular words which conform to spelling-to-sound rules should be converted to a speech code faster than exception words which violate the rules. Accordingly, the time to comprehend the word and name it aloud should take longer for the words that violate spelling-to-sound rules. In support of their hypothesis, the pronunciation times for exception words averaged 27 msec longer than the pronunciation times for regular words. However, there is no assurance that differences in pronunciation time result from differences in word recognition time. The differences in reaction time could also have resulted from differences in the time for response selection and programming after the word had already been identified (see Massaro, 1975b, p. 262).

In order to provide evidence that differences in naming times directly reflect differences in word recognition time, it is necessary to perform a stage analysis of the naming task and to include a number of other independent variables known to influence specific stages in the task (Massaro, 1975a; Sternberg, 1969). Consider this task in terms of the model depicted in Figure 1. Naming a written pattern requires feature detection, primary recognition, and secondary recognition processes and also response selection and response programming operations. In terms of this analysis the reaction time (RT) between the onset of the written pattern and the onset of the spoken response is a composite of 5 component times:

$$RT(name) = FD + PR + SR + RS + RP \quad (2)$$

where FD, PR, SR, RS, and RP represent the times for the five respective processes.

The critical components in the naming task include the time to detect the visual features and resolve the shape of the pattern (that is, to see the pattern), attach a name (speech code) to the seen pattern, to select the appropriate articulatory program for the speech code, and finally to program the articulators to execute the response. Response execution time would not contribute to the actual RT since the RT is measured at the onset of the naming response. One simplification of the analysis would be to divide up the RT into input and output operations. In this case the first three operations would entail stimulus processing, whereas the last two would represent response operations.

$$RT = SP + RO \quad (3)$$

where SP and RO equal the times for stimulus processing and response operations, respectively.

The present concern for localizing naming-time differences at a particular stage of processing is not new; in fact, James McKeen Cattell (1888) provided exactly this analysis almost 100 years ago. Many contemporary investigators have cited Cattell's finding that a short word can be named in less time than a single letter, with the implication that words are perceived in less time than are single letters. Cattell realized, however, that the naming task included both perception time and "will-time," as he called the time to choose a response. Relative differences in perception time were determined by using a Donders Type C reaction, in which observer was required to make a simple response such as lifting a finger off a key in response to just one of many possible alternative stimuli. In different tests the subject was asked to respond to just one of many possible letters, words, colors, and so on. In contrast to the naming task the results showed shorter reaction times for single letter than word alternatives. Cattell's interpretation appears to be still valid today. To quote from his *Popular Science Monthly* article in 1888, "The time it takes to think," "A letter can be seen more quickly than a word, but we are so used to reading

aloud that the process has become quite automatic, and a word can be *read* with greater ease and in less time than a letter can be named [p. 23, my italics]." In a similar set of studies, Cattell found that it takes less time to perceive a color than a letter or word but much longer to name the color relative to naming the letter or word. Huey (1908) also believed that letter recognition mediated word recognition, but that the context set for words "drafts to itself the energy which would have been given to the letters, [p. 113]." This is analogous to the Stroop color-word phenomenon.

That naming RTs depend on response selection time is apparent in a series of experiments showing that the RTs increase with increases in the number of syllables of the word to be named (Eriksen, Pollack, and Montague, 1970; Klapp, 1974). Klapp, Anderson, and Berrian (1973) attributed the naming differences to response selection and preparation since similar tasks requiring recognition without pronunciation showed no effects of syllable length. In addition, picture naming was found to be syllable dependent in the same way as words. Frederiksen and Kroll (1976) found large effects of word length and syllable structure on naming RTs but no effects of these variables on the time to make lexical decisions. These results make it apparent that response processes must be accounted for in naming studies of word recognition.

Theios and Muise (1976) compared pronunciation times for real words and pseudowords, while attempting to control for the articulation response. For each real word there was a yoked pseudoword which was homophonic to the word, but spelled differently. Given that articulation differences between words and their corresponding pseudowords would be eliminated, the pronunciation times might provide a more direct index of recognition times. Naming times were 20 msec longer for pseudowords than for real words. Frederiksen (1976) carried out a similar experiment, but created his pseudowords by changing a single vowel in each of the real words. However, rather than randomizing words and pseudowords in the same session as in the Theios and Muise study, Frederiksen had words tested in one session and pseudowords in another. The pseudowords took 142 and 49 msec longer to pronounce for poor and good high-school readers, respectively.

The advantage of words over pseudowords can be interpreted to mean that subjects can more quickly recognize a word and retrieve the appropriate articulatory program stored in lexical memory than they can map a nonword letter string into a phonological code based on spelling-to-sound rules of the language. If derivation of a phonological code always preceded lexical access, as assumed by Gough (1972), then naming times should have been equivalent for words and pseudowords. Given that words and nonwords were equated on phonology in the Theios and Muise study and closely matched on orthography in the Frederiksen study, it seems likely that the word advantage over pseudowords reflects processes that depend on lexical access for words

but not pseudowords. Frederiksen may have found much larger differences than Theios because blocking the words in a session would encourage the subjects to pronounce the words via lexical access. Randomizing words and pseudowords in the Theios and Muise study might have encouraged pronouncing some of the words by way of spelling-to-sound rules rather than by way of lexical access. In agreement with this interpretation, Frederiksen and Kroll (1976) found a larger effect of word frequency on naming RTs when only words were presented in a block of trials relative to a random mixture of word and pseudoword trials.

Green and Shallice (1976) asked subjects to judge whether two words rhymed or whether they belonged to the same broad semantic category. Mispelling the words as homophones produced a much larger decrement in the semantic than the rhyming task. If lexical access occurs via phonological coding, there is no reason that the semantic task should have been slowed more by misspelling than the rhyming task was. The fact that the rhyming task was performed about twice as fast as the semantic task shows that lexical access was not necessary in the former task although it was in the latter. Spelling-to-sound rules would have been sufficient to perform the rhyming task, and misspelling should have very little effect on this process. In support of this, misspelling the words increased reaction times by only 11 percent. Lexical access should be drastically influenced by misspelling however, if it occurs via a visual code. Reaction times were slowed by 58 percent in the semantic task, arguing against the idea of phonological or speech recoding in lexical access and derivation of meaning. The results support other negative findings on the necessity of phonemic encoding in processing written language for meaning (see Massaro, 1975a).

V. Rehearsal and Recoding

In the present model, the same abstract structure stores the meaning of both listening and reading. Generated abstract memory (GAM) in our memory corresponds to the working memory of contemporary information processing theory. Rehearsal and recoding processes operate at this stage to maintain and build semantic/syntactic structures. There is good evidence that this memory has a limited capacity, holding about 5 ± 2 chunks of information. For a more detailed discussion of processing at this stage, see Massaro (1975a, Chapter 27).

Although GAM is assumed to be abstract relative to SAM and SVM, the nature of the information appears to be tied to the surface structure of the language rather than in terms of underlying meaning that is language independent. Some relevant research comes from work experiments carried out with bilingual subjects (Dornic, 1975, provides an excellent review). Recall from immediate memory (supposedly tapping GAM) does not differ for unilingual and bilingual lists, whereas recall of items as-

sumed to be no longer in GAM is poorer in bilingual than unilingual lists (Tulving & Colotla, 1970). Similarly, Kintsch and Kintsch (1969) showed that the semantic relationship between the words in different languages did not influence immediate memory, but did affect recall of items no longer active in GAM. Saegert, Hamayan, and Ahmar (1975) showed that multilingual subjects remembered the specific language of words in a mixed language list of unrelated words, but this information was forgotten when the words were presented in sentence contexts. Dornic (1975) points out that surface structure and item information are integrally related in immediate memory; subjects seldom report translations for the words. If the items are remembered, so are the appropriate surface structure forms.

In our model, GAM has a "limited capacity" and the learning and memory for information is a direct function of rehearsal and recoding processes. Memory of an item will increase with the time spent operating on that item, and will decrease with the time spent operating on other "unrelated" items. This "limited capacity" rule has provided a reasonable description of the acquisition and forgetting of information in GAM (cf. Massaro, 1975a, Chapter 27). A critical question for the recoding operation centers around the size of the units that are recoded. It seems unlikely that recoding occurs word by word given that many words are ambiguous until later context disambiguates their meaning.

VI. Conclusion

It seems valuable to attack reading and listening with similar methodological and theoretical forces in the framework of an information-processing model. Our concern is with *how* the reader and listener perform, and with the dynamics of this performance. Although the surface structure of written text and speech present questions unique to each skill, the apparent similarities in deep structure offer the hope of a single framework for understanding both reading and listening.

VII. Preview of Contributions on Reading and Listening

One reason that speech has been considered primary and reading and writing secondary is the supposedly uniqueness of certain speech perception phenomena. At the top of the list has been the categorical perception of speech sounds. Categorical perception refers to a basic perceptual limitation in the perception of speech sounds. Certain sounds cannot be discriminated from one another unless they are, in fact, categorized differently. For example, the two sounds /ba/ and /pa/ can be synthesized electronically so that they differ only along a single dimension called voice onset time (VOT, the time between the onset of the stop release and the onset of vocal cord vibration in real speech). If two of the sounds differ by a VOT of 10 msec, they will not be discriminated from each other if

they are normally categorized as the same (for example, /ba/) but will be discriminated if they are categorized differently (for example, /ba/ and /pa/). In this example a 10-msec difference could be perceived (discriminated) when the two sounds are from different phonemic categories but not when the two sounds are from the same phonemic category.

The phenomenon of categorical perception has attracted renewed interest in the last five years and it would take more than a special issue to give it adequate coverage. The bulk of the empirical and theoretical work, however, is easily summarized. It is now generally accepted that categorical perception does not necessarily reflect a perceptual limitation in the processing of certain speech sounds, and in addition, analogous phenomena have been demonstrated with nonspeech sounds. Working within this framework, Pastore develops the idea of a reference point as one important aspect in the establishment of categorical perception phenomena. Rather then studying the processing of speech sounds, however, he shows that this idea is equally applicable to the processing of alphabetic symbols.

Given an alphabetic writing system, it is only natural to expect the relationship between spelling and sound to be exploited. There is now sufficient evidence against the idea that the lexical access of printed words occurs via their sound. However, there are many other processing stages and tasks in which utilization of spelling-to-sound correspondence might be important. One such task is spelling. In many situations the spelling of a word can be at least partially disambiguated by spelling-to-sound correspondences. Consider the noun and verb forms of an opinion, counsel, or recommendation. The noun pronounced with a final *voiceless* fricative must be *advice* or *advise* whereas the verb with a final voiced fricative must be *advise* or *advize*. Although sound does not provide a complete disambiguation in this example, at least one incorrect alternative has been eliminated in both cases. The voiceless fricative can not be spelled with a *z* and the voiced fricative can not be spelled with a *c*. In a nice series of studies, Frith shows that good readers who are also good spellers, have mastered both the spelling-to-meaning and spelling-to-sound correspondences of the language. In contrast, good readers who are poor spellers show a deficit in their knowledge of spelling-to-sound correspondences but not in spelling-to-meaning correspondences. This result shows that reading and writing utilize different processes and that excellence in one does not insure excellence in the other.

Continuing the exploration into spelling-to-sound correspondences, Baron and Hodge attempt to distinguish among four processes underlying the learning of associations between printed and spoken words. The nature of the learning process has implications for both theories of reading and educational practice. In a series of experiments the authors are not only able to clarify the theoretical viewpoints; they show that some similarity

relations between the printed words and the spoken responses are critical. Readers learned spelling-to-sound correspondences only when it was the case that similar stimuli had similar responses.

Martin, Meltzer, and Mills explore one of the differences in processing spoken and written language. The spoken sentence paces listening, and the prosodic cues in the passage may contribute significantly to comprehension. In reading, the sentence is presented in static form and there are no external guides to pace reading. Working within the framework of Martin's rhythmic theory of the temporal organization of speech, the authors devise a situation in which a written sentence is presented dynamically in synchrony with a spoken version. In this rhythmic presentation, each written syllable appears on the screen simultaneously with the onset of the spoken syllable. High school students learning Spanish were trained on either rhythmic or nonrhythmic presentations and the results showed that the rhythmic training facilitated the reading fluency of new sentences. These results open up a series of questions on the relationship between reading and listening and how it should be implemented in pedagogical practice.

Levy's contribution clarifies the locus of the speech suppression effect which refers to interfering with memory of a written passage by requiring the readers to count rapidly or to shadow spoken material during reading. The interference supposedly occurs because speaking during reading prevents some kind of subvocal speech recoding of the visually-presented material. In the previous studies, however, the memory tests for the written material required the subjects to maintain the exact surface form of the sentences and discouraged thematic processing. To test whether speech suppression will also occur if the memory task does not require exact wording information, Levy asked subjects to perform a paraphrase detection task. In this task, none of the test sentences were identical to the original ones and the subjects had to indicate whether or not the paraphrase altered the meaning of one of the original sentences. Speech suppression did not affect performance in this condition supporting the idea that a speech code is not necessary for imposing meaning on printed text but is critical when exact wording information must be maintained. This result is consistent with our idea that GAM which maintains the surface form of a sentence is at least partially organized along a speech code dimension.

References

Blakemore, C.
The baffled brain. In R. L. Gregory and E. H. Gombrich (Eds.), *Illusion in nature and art.*
London: Duckworth, 1973.

Blakemore, C.
Mechanics of the mind.
New York: Cambridge University Press, 1977.

Blesser, B.; Shillman, R.; Kuklinski, T.; Cox, C.; Eden, M., & Ventura, J.
A theoretical approach for character recognition based on phenomenalogical attributes.
International Journal of Man-Machine Studies,
1974, 6, 701-714.

Cattell, J. M.
The time it takes to think.
Popular Science Monthly, 1888, 32, 488-491. Reprinted in James McKeen Cattell, *Man of science* Vol. 2, addresses and formal papers, Lancaster, PA: The science press, 1947.

Chomsky, N., & Halle, M.
The sound pattern of English.
New York: Harper & Row, 1968.

Cole, R. A.
Listening for mispronunciations: a measure of what we hear during speech.
Perception & Psychophysics,
1973, 13, 153-156.

Cole, R. A. & Jakimik, J.
Understanding speech: how words are heard.
Technical report,
Department of Psychology, Carnegie-Mellon University, 1977.

Cosky, M. J.
The role of letter recognition in word recognition.
Memory & Cognition,
1976, 4, 207-214.

Dornic, S.
Human information processing and bilingualism.
Report from the Institute of Applied Psychology, the University of Stockholm, No. 67, 1975.

Eriksen, C. W.; Pollack, M.D., and Montague, W. E.
Implicit speech: mechanism in perceptual encoding.
Journal of Experiment Psychology,
1970, 84, 502-507.

Frederiksen, J. R.
Decoding skills and lexical retrieval.
Paper presented at Psychonomic Society, St. Louis, November, 1976.

Frederiksen, J. R., and Kroll, J. F.
Spelling and sound: approaches to the internal lexicon.
Journal of Experimental Psychology: Human Perception and Performance,
1976, 2, 361-379.

Gough, P. B.
One second of reading. In J. F. Kavanagh & I. G. Mattingly (Eds.)
Language by ear and by eye,
Cambridge, Mass., MIT Press, 1972.

Gough, P. B., & Cosky, M. J.
One second of reading again. In N. J. Castillan, Jr.; D. B. Pisoni, and G. R. Potts, (Eds.).
Cognitive Theory, Vol. 2, Hillsdale, N.J., Erlbaum, 1976.

Green, D. W., & Shallice, T.
Direct visual access in reading for meaning.
Memory and Cognition,
1976, 4, 753-758.

Hubel, D. N., & Wiesel, T. N.
Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex.
Journal of Physiology,
1962, 160, 106-154.

Huey, E. B.
The psychology and pedagogy of reading.
New York: MacMillan, 1908 (Reprinted Cambridge: MIT press, 1968).

Jakobson, R.; Fant, C. G. M., & Halle, M.
Preliminaries to speech analysis: the distinctive features and their correlates.
Cambridge, Mass.: MIT Press, 1961.

Keyser, S. J., & Halle, M.
What we do when we speak. In P. A. Kolers & M. Eden (Eds.)
Recognizing Patterns.
Cambridge, Mass.: MIT Press, 1968.

Kintsch, W., & Kintsch, E.
Interlingual interference and memory processes.
Journal of Verbal Learning and Verbal Behavior,
1969, 8, 16-19.

Klapp, S. T.
Syllable-dependent pronunciation latencies in number naming: a replication.
Journal of Experimental Psychology.
1974, 102, 1138-1140.

Klapp, S. T.; Anderson, W. G., & Berrian, R. W.
Implicit speech in reading, reconsidered.
Journal of Experimental Psychology,
1973, 100, 368-374.

Ladefoged, P.
A course in phonetics.
New York: Harcourt, Brace & Jovanovich, 1975.

Lane, H.; Boyes-Braem, P., & Bellugi, U.
Preliminaries to a distinctive feature analyses of handshapes in American Sign Language.
Cognitive Psychology,
1976, 8, 263-289.

Lieberman, P.
Some effects of semantic and grammatical context on the production and perception of speech.
Language and Speech,
1968, 6, 172-187.

Lieberman, P.
Intonation, perception, and language.
Cambridge, Mass.: MIT Press, 1967.

Lindsay, P. H., & Norman, D. A.
Human information processing.
New York: Academic Press, 1977.

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

Marslen-Wilson, W. D.
Sentence perception as an interactive parallel process.
Science,
1975, 189, 226-228.

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

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

Massaro, D. W.
Understanding language: an information-processing model of speech perception, reading, and psycholinguistics.
New York: Academic Press, 1975(b).

Massaro, D. W.
Reading and listening.
Technical Report No. 423.,
Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin, Madison, Wisconsin, 1977.

Massaro, D. W., & Cohen, M. M.
The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction.
Journal of the Acoustical Society of America,
1976, 60, 704-717.

Massaro, D. W., & Cohen, M. M.
Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction.
Perception and Psychophysics,
1977, 22, 373-382.

Mayzner, M. S., & Tresselt, M. E.
Tables of single-letter and diagram frequency counts for various word-length and letter-position combinations.
Psychonomic Monograph Supplements,
1965, *1*, 13-32.

Naus, M. S., & Shillman, R. J.
Why a Y is not a V: a new look at the distinctive features of letters.
Journal of Experimental Psychology: Human Perception and Performance,
1976, 2, 394-400.

Oden, G. C., & Massaro, D. W.
Integration of place and voicing information in identifying synthetic stop-consonant syllables.
WHIPP Report #1,
Wisconsin Human Information Processing Program, July, 1977.
[Also *Psychological Review,* 1978, 85, 172-191.]

Reicher, G. M.
Perceptual recognition as a function of meaningfulness of stimulus material.
Journal of Experimental Psychology,
1969, 81, 275-280.

Rosenberg, S., & Lambert, W. E.
Contextual constraints and perception of speech.
Journal of Experimental Psychology,
1974, 102, 178-180.

Ross, J. R.
Parallels in phonological and semantic organization. In J. F. Kavanagh & J. E. Cutting (Eds.)
The role of speech in language,
Cambridge, Mass.: MIT Press, 1975.

Saegert, J.; Hamayan, E., & Ahmar, H.
Memory for language of input in polyglots.
Journal of Experimental Psychology: Human Learning and Memory,
1975, 1, 607-613.

Shillman, R. J.; Cox, C.; Kuklinski, T.; Ventura, J.; Blesser, B., & Eden, M.
A bibliography in character recognition. Techniques for describing characters.
Visible Language,
1974, 8, 151-166.

Spencer, H.
The visible word.
New York, Hastings, 1969.

Sternberg, S.
The discovery of processing stages: extensions of Donders' method.
Acta Psychologica,
1969, 30, 276-315.

Theios, J., & Muise, J. G.
The word identification process in reading. In N. J. Castellan, Jr.; D. B. Pisoni, & G. R. Potts (Eds.)
Cognitive Theory,
Vol. 2, Hillsdale, N.J., Erlbaum, 1976.

Treisman, A. M.
Verbal responses and contextual constraints in language.
Journal of Verbal Learning and Verbal Behavior,
1965, 4, 118-128.

Tulving, E., & Colotla, V. A.
Free recall of bilingual lists.
Cognitive Psychology,
1970, 1, 86-98.

Tweney, R. D.; Heiman, G. W., & Hoemann, H. W.
Psychological processing of sign language: effects of visual description on sign intelligibility.
Journal of Experimental Psychology: General,
1977, 106, 255-268.

Umeda, N.
Consonant duration in American English.
Journal of Acoustical Society of America,
1977, 61, 846-858.

Wrolstad, M. E.
A manifesto for visible language.
Visible Language, 1976, 10, 5-40.

Zadeh, L. A.
Quantitative fuzzy semantics.
Information Sciences,
1971, 3, 159-176.