CONTRIBUTORS

Arthur Freund  Joseph Schmuller
Joseph B. Hellige  Wayne Shebilske
Dominic W. Massaro  Kenneth B. Solberg
Kenneth R. Paap  Lucinda Wilder
Understanding Language

An Information-Processing Analysis of Speech Perception, Reading, and Psycholinguistics

Edited by

Dominic W. Massaro

Department of Psychology
University of Wisconsin—Madison
Madison, Wisconsin

ACADEMIC PRESS  New York  San Francisco  London

A Subsidiary of Harcourt Brace Jovanovich, Publishers
List of Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

Arthur Freund (357), Department of Psychology, University of Wisconsin—Madison, Madison, Wisconsin
Joseph B. Hellige (391), Department of Psychology, University of Southern California, Los Angeles, California
Dominic W. Massaro (3, 77, 125, 207, 241), Department of Psychology, University of Wisconsin—Madison, Madison, Wisconsin
Kenneth R. Paap¹ (151), Department of Psychology, University of Wisconsin—Madison, Madison, Wisconsin
Joseph Schmuller² (207), Department of Psychology, University of Wisconsin—Madison, Madison, Wisconsin
Wayne Shebilske (291), Department of Psychology, University of Virginia, Charlottesville, Virginia
Kenneth B. Solberg³ (315), Department of Psychology, University of Wisconsin—Madison, Madison, Wisconsin
Lucinda Wilder (31), Millhauser Laboratories, New York University Medical Center, New York, New York

¹ Present address: Department of Psychology, New Mexico State University, Las Cruces, New Mexico.
² Present address: Department of Psychology, Clark University, Worcester, Massachusetts.
³ Present address: Department of Psychology, St. Mary's College, Winona, Minnesota.
Preface

Three years ago we set out to understand how language is processed. The recent advances in experimental psychology in the information-processing area encouraged us to develop and utilize an information-processing approach to language processing. This task seemed appropriate because we aimed to describe how language is processed, not simply what the listener or reader must know to understand language. In the information-processing approach, language processing is viewed as a sequence of psychological (mental) stages that occur between the initial presentation of the language stimulus and the meaning in the mind of the language processor. Our goal was to define each of the processes and structures involved and to understand how each of them operates. This volume is intended to communicate what we have learned in this exciting and rewarding adventure.

We apply the latest advances in psychology and linguistics to the understanding of language processing. This volume articulates the current state of the art in speech perception, reading, and psycholinguistics, and it can serve as a basic text for any of these topics. The information-processing approach along with supporting evidence is described in the "Introduction." The section entitled "Speech Perception" covers the fundamentals of articulatory and acoustic characteristics of speech sounds, the acoustic features used in speech perception, the dynamic aspects of speech perception, and theories of speech perception. The visual
features used in reading, the dynamics of the reading process, reading eye movements, and theories of reading are covered in the section entitled "Reading." Part IV, "Psycholinguistics," treats the latest advances in linguistic theory, the dynamic aspects of word and phase recognition, and the role of syntactic and semantic structure in the processing of language. By limiting ourselves to the so-called earlier stages of language processing, we have focused on the dynamic psychological processes and structures involved in obtaining meaning from spoken and written sentences.

Charles Read, Richard L. Venezky, and Domenico Parisi read portions of an early draft of the volume. Their comments and helpful reactions are deeply appreciated. We would also like to thank Ronald A. Cole, William K. Estes, James Martin, and Edward E. Smith for copies of their work before it was published. In addition we are obliged to the scientists who gave their permission to reproduce findings from their research.

The writing and editing of the final version of the book took place while I was a John Simon Guggenheim Memorial Foundation Fellow. Their support and that of the Wisconsin Alumni Research Foundation is gratefully acknowledged.

D. W. M.
Language and Information Processing

Dominic W. Massaro

I. INTRODUCTION

This book attempts to apply the latest theoretical development in psychology and linguistics to language processing. In recent years psychologists have attempted to describe the mental or psychological processes that take a person from contact with a stimulus situation to some form of knowledge revealed in some observable response. For example, when presented with a letter string such as *cet* and asked if it spells an English word, the subject must resolve the shape of the letters, determine what letter each of these shapes represents in the English alphabet, and ask whether he knows the meaning of this particular letter string. Before he says no he may also ask if the letter string is a word he knows by sound but not by sight. In this case, the person may attempt to translate the letter pattern into a sound pattern and then determine if the sound pattern has any meaning. After failing to find meaning in the sound, the subject may state that the letter string is not in his vocabulary but could be a word. Given the letter string *cht*, the language user could reject it as an English word much more easily, since he could argue that it disobey the way words must be spelled. What this psychological experiment shows is that the language user knows certain things about the structure of his language and is able to apply them to the task at hand.
The most impressive implication of the results is that the language user's knowledge is in the form of rules or abstract principles rather than in terms of specific memories or experiences. It is unlikely that the subject had seen *cet* before, and he certainly had never been asked if it was a word in his vocabulary.

The psychologist is concerned with *how* the subject performs the task. His goal is to describe the sequence of psychological events that intervene between the original contact with the letter string and the yes or no answer. In this case, it is necessary to understand how the subject obtains, remembers, and utilizes the knowledge illustrated in his language behavior. To achieve the proper experimental control, the psychologist found it necessary to exorcize the subject's knowledge by creating experimental situations in which this knowledge was useless. Therefore, the psychologist studied the recognition of simple auditory and visual patterns, the learning and memory of random letter strings, and decision making in simple psychophysical situations. Out of this work developed methods for studying perceptual, memorial, decision, and response selection processes—those psychological processes central to understanding language.

The linguist, on the other hand, is concerned with formalizing a representation or description of the structure of the language. The structure includes stimulus-perceptual events such as speech sounds, abstract-cognitive events such as meanings, and a grammatical system that relates sounds and meanings. The description of the language provides a possible description or representation of what the language user knows. It may not be correct, since the linguist has been concerned mainly with linguists' judgments or intuitions about language under ideal conditions. In this case, the linguist may be influenced more by the elegance of the representation than by its psychological validity, that is, whether the representation actually describes the structure of the knowledge of the language user.

Although both linguistics and psychology contribute to our understanding of language, our approach is psychological rather than linguistic. One reason is that we are concerned with how language is understood as it is conveyed by speech or writing, not how it is produced by the speaker or writer. Although there is no logical reason for it, linguistics has concentrated on the production of language, whereas experimental research has focused on how it is understood. More important, we utilize a psychological approach because we aim to describe *how* language is processed, not simply what the listener or reader must know to understand language. As psychologists, we view the understanding of language as a sequence of psychological (mental) processes that occur between the initial pre-
sentation of the language stimulus and the meaning in the mind of the language processor. Our goal is to define each of the processes involved and to understand how each of them operates.

II. INFORMATION PROCESSING

This conceptualization of language processing can be formalized in an information-processing model. The information-processing model delineates each of the component processes or processing stages between the language stimulus and the meaning response. We utilize an information-processing approach because it makes possible an adequate explanation of how language is understood. Consider again our original question of whether or not cet is a word. The stimulus is the letter string cet, and the response is the answer no. No adequate explanation of how this task is performed can really ignore intervening psychological processes or operations that must occur between the presentation of the stimulus and the onset of the response. The information-processing approach allows the experimenter to be specific about what these processes are and how they occur. The information-processing approach not only provides a theoretical framework for describing language processing but also gives a precise methodology for performing and interpreting psychological experiments (Massaro, 1975).

Language processing is the abstraction of meaning from an acoustic signal or from printed text. To derive or arrive at meaning from the spoken (written) message requires a series of transformations of the acoustic (visual) stimuli arriving at the ears (eyes). We look upon language processing as a sequence of processing stages or operations that occur between stimulus and meaning. We understand language processing only to the extent that we understand each of these processing stages.

In this book we utilize a general information-processing model for a theoretical analysis of speech perception, reading, and psycholinguistics. The model is used heuristically to incorporate data and theory from a wide variety of studies. The model should be taken as an organizational structure for the state of the art in language processing. The information-processing framework allows us to evaluate results and theory from a number of different approaches to the study of language processing. The main advantage of the model is that it forces us to be consistent in our terminology, assumptions, and conclusions.

This book limits itself to the earlier stages of information processing. It allows us to disambiguate a sentence or two. It does not deal with
the additional processes responsible for deriving the plot of a mystery novel or the theme of a poem. However, the model does describe in detail the operations necessary to go from a physical stimulus to meaning at the sentence level. The book is both basic and yet advanced. It is basic because it covers such fundamentals as articulatory and acoustic phonetics and a discussion of linguistic theories. It is advanced because it also presents a critical review of current psychological theories and empirical studies of language processing.

The central assumption of our approach is that language processing should be described in the framework of an information-processing model. In this chapter our information-processing model will be presented and implications of the model for the processing of language discussed. Figure 1.1 presents a flow diagram of the temporal course of information processing. At each stage the system contains structural and functional components. The model distinguishes four processes or functional components (circles): feature detection, primary recognition, secondary recognition, and recoding. Corresponding to each process there is a structural component (boxes) that represents the information available to that stage of processing.

![Flow diagram of the temporal course of auditory and visual information processing.](image-url)
III. AUDITORY INFORMATION PROCESSING

The stimulus for deriving meaning from spoken language is the acoustic signal that arrives at the ear of the listener. This section will present our model by tracking the flow of auditory information through the processes of detection, recognition, and recoding, and introduce the topics to be considered in the chapters that follow.

A. Feature Detection

The feature detection process transforms the sound wave pattern into acoustic features in preperceptual auditory storage. The features are described as acoustic, since we assume that there is a direct relationship between the nature of the auditory signal and the information in preperceptual storage. This one-to-one relationship between the auditory signal and the information in preperceptual storage distinguishes the feature detection process from the following stages of information processing. There is no one-to-one relationship between the input and output of the following processing stages, since these later stages actively utilize information stored in long-time memory in the sequence of transformations. For this reason the passive transduction of feature detection contrasts with the active construction of the following processing stages.

Feature detection refers to the neural processing necessary for determining whether a feature was or was not present. Feature detection can be readily described by counter (McGill, 1963, 1967) or timer (Luce & Green, 1972) models. In the counter model, a feature is detected if the number of neural pulses along a given channel exceeds a criterion number in some period of time. In a simple timing model, a feature is detected if the time between successive pulses along a given channel is less than some minimal criterion time. Since the neural pulses are all-or-none, it is sufficient to process either the number of pulses or their interarrival times.

There are two points to be made about the detection process. First, different features require different amounts of time for feature detection. Therefore the features will not enter preperceptual storage simultaneously. Second, there is a certain amount of background noise in the feature detection process. Since this background noise fluctuates over time, a nonexistent feature may be erroneously detected or a feature actually represented in the stimulus may be missed.

In the information-processing approach our concern is with the acoustic features utilized in processing speech and with the nature of the feature
detection process. Given that the acoustic features are tied to the acoustic signal, it is first necessary to describe the acoustic characteristics of the speech stimulus. Since the acoustic properties in a speech stimulus follow directly from the properties of the human vocal apparatus that produces the speech, Chapter 2 begins with a detailed discussion of the production of speech sounds.

Chapter 2 then presents a description of the acoustic characteristics of the speech sounds of English. The visual representation of the sound patterns given by the sound spectrograph is used to characterize the acoustic properties of the sound patterns. The analysis shows that the phonemes and syllables of English can be distinguished by differences with respect to a number of acoustic characteristics described in the spectrographic analysis. The spectrograph still appears to be the best representation of the acoustic characteristics that are used to distinguish the sound patterns from each other.

Chapter 3 analyzes the psychological reality of the acoustic characteristics observed in the spectrographic analysis. The psychological studies presented in Chapter 3 ask what acoustic features or cues are sufficient for recognizing or distinguishing different speech sounds. Essentially these studies are concerned with whether the acoustic characteristics observed in the stimulus function as acoustic features. An acoustic characteristic will be called an acoustic feature when it is used to distinguish between different speech sounds.

The analyses in Chapters 2 and 3 make it clear that no small segment of the speech signal is sufficient for recognition of a speech sound. Rather, the signal must extend in time, since a complete sound pattern is necessary to provide enough information to distinguish it from other possible alternatives. The perception process cannot take place as the stimulus is arriving, since a complete sound pattern of some length is necessary for recognition to occur. Therefore the acoustic features in the signal must be stored until the sound pattern is complete. Our model assumes that this information is held in a structural component called preperceptual auditory storage, where it may remain for about 250 msec.

B. Preperceptual Auditory Storage

In the perception of speech or music, the preperceptual auditory image holds the first part of any auditory stimulus until the sound pattern is complete and primary recognition occurs. A second pattern does not usually occur until the first has been perceived. However, if the second pattern is presented soon enough, it should interfere with recognition of the first pattern. By varying the delay of the second pattern, we can deter-
mine the duration of the preperceptual auditory image and the temporal course of the recognition process. The experimental task is referred to as a **recognition-masking paradigm** (Massaro, 1972). In a typical experiment the test signals are two short (20 msec) tones differing in pitch quality, which the observer first learns to identify with the labels “high” and “low.” One of the test tones is presented and followed by a masking tone after a variable silent interval. The observer must report which of the test tones preceded the mask. The test and masking tones are presented at the same loudness so that the time between the onsets of the test and masking tones provides a true measure of the perceptual processing time available for the test tone.

Figure 1.2 presents the results of a typical experiment (Massaro, 1970b). With the exception of the initial rise in performance at the zero interval, the figure shows that for each of the three observers recognition performance improved with increases in the silent interval up to 250 msec. Further increases in the silent interval beyond 250 msec did not significantly facilitate recognition performance. These results provide information about the preperceptual auditory image of the test tone and the vulnerability of the auditory image to new inputs. Given that the test tone.

---

**Figure 1.2.** Percentages of correct identifications of the test tone for subjects AL, NS, and CB as a function of the silent interval. (Data from Massaro, 1970b.)
lasted only 20 msec, some preperceptual image must have remained available for processing so that recognition performance improved with increases in the silent intertone interval. This result indicates that the masking tone terminated perceptual processing of the image. Since recognition performance levels off at about 250 msec, a quarter of a second is an estimate of the useful life of the preperceptual image. The role of preperceptual auditory storage in speech processing is discussed in Chapter 4.

C. Primary Recognition

Primary recognition (perception) is the transformation of the features held in preperceptual auditory storage into a percept in synthesized auditory memory. The operations of the primary recognition process are discussed in Chapters 3 and 4. The minimal sound patterns in speech that can be recognized are referred to as **perceptual units of information**. Perceptual units correspond to those sound patterns that are uniquely represented in long-term memory by a list of acoustic features. The information in the perceptual unit can therefore be defined by a set of acoustic features that corresponds to a list of features in long-term memory. The primary recognition process finds a representation in long-term memory that matches the acoustic features held in preperceptual storage. Figure 1.3 provides a graphic description of a perceptual unit's representation in long-term memory. This representation is a **sign**, which is a combination of a feature list and a synthesis program. The feature list contains a description of the acoustic features in the perceptual unit. The synthesis program is an algorithm for synthesizing (saying or hearing) that particular sign. Chapters 3 and 4 provide a detailed discussion of the perceptual units employed in the recognition of speech.

It is important to note that these perceptual units are without meaning. We can repeat back a string of speech sounds, for example, nonsense syl-

---

**Figure 1.3.** Schematic drawing of representation of perceptual unit in long-term memory.
lables, without deriving meaning from the message. Although the first stage of primary recognition allows the observer to “shadow” speech, it does not enable him to paraphrase what has been said. This view of the recognition process agrees with most contemporary views of the speech perception problem (cf. Chapter 5). We agree with Cooper (1972), who believes that “speech perception can go on without reference to meaning [p. 54].” Accordingly, the first stage of speech perception is the transformation of the prereceptual representation of the pattern into a synthesized auditory percept.

The primary recognition process might utilize knowledge of sequential constraints in the message such as the phonological rules of the language. Phonological rules can be thought of as specifying permissible sequences of phonemes or phonetic features in the language. For example, if a stop consonant follows a nasal within the same syllable, they must be homorganic, that is, share the same place of articulation. Therefore we can have *ump* or *sunk* but not *sumk* or *sunp*. Assume that the primary recognition process was faced with partial acoustic information about the syllable /ump/. If it had enough information to recognize the vowel and the stop consonant but only enough to recognize the nasal as a nasal but not which one, the appropriate syllable could still be synthesized, because this phonological rule specifies the appropriate nasal. This example shows how the primary recognition process could actively construct a synthesized percept from the acoustic information in prereceptual auditory storage and phonological rules in long-term memory.

Chapter 5 presents the defining properties of contemporary theories of the perception process in speech. These models are evaluated in terms of our information-processing model and the empirical evidence presented there and in the previous chapters. The significant conclusion reached in Chapter 5 is that it is not necessary to reference the articulatory production machinery in our description of speech perception. A close analysis of the models reveals that they have more similarities than differences and that they can be incorporated readily into our general information-processing model.

D. Synthesized Auditory Memory

We use the term *synthesized auditory memory* because the primary recognition process is a synthesis of the information in prereceptual storage. Whereas prereceptual storage contains separate acoustic features, synthesized auditory memory holds a synthesized unit or gestalt. One experimental demonstration of synthesized auditory memory is a study of memory for voice quality by Cole, Coltheart, and Allard (1974). Sub-
jects, presented with a sequence of two spoken letters, reported as quickly as possible whether or not the second letter had the same name as the first. The independent variable of interest was whether the two letters were presented by same or different speakers. The results indicated that subjects could respond faster on both same and different name trials when the letters were in the same voice than when spoken by different voices. Also, this advantage was independent of the duration of silence separating the two letters (½ to 8 sec). These results indicate that subjects can synthesize what a speaker says, remember the characteristics of the speaker’s voice in synthesized auditory form, and use this information to enhance processing of a second signal. This result agrees with the observations of Ladefoged and Broadbent (1957), which show that a listener’s perception of a particular speech sound is influenced by the voice characteristics of the earlier speech input (see Chapter 3).

E. Secondary Recognition

The outcome of the primary recognition process corresponds to the phenomenal experience of hearing a particular speech sound. This percept is stored in synthesized auditory memory, and the listener’s task now involves an analysis for meaning. The goal of the secondary recognition process is to transform this perceptual information into conceptual information, that is, meaningful units in generated abstract memory. This conceptual stage of processing involves a lexicon in long-term memory, and possible utilization of syntactic rules of the language, contextual or situational knowledge, and abstract semantic knowledge.

The lexicon in long-term memory can be viewed as a multidimensional representation of words and a few common word phrases. The representation of a word has both perceptual and conceptual attributes. The perceptual code of wind might be the sound of the word wind, the look of the letters that spell wind, the sound of the wind blowing, and the pictorial representation of a windy scene. The conceptual code of wind would be the variety of properties that constitute the meaning of wind, such as air movement. The secondary recognition process looks for a match between the perceptual code of the sound pattern held in synthesized auditory memory and a representation of that code in long-term memory.

Every word, and possibly a few phrases, in the listener’s lexicon has a representation in long-term memory. This representation contains perceptual and conceptual dimensions. The auditory perceptual dimension contains the sequence of perceptual units in the word and their intonation pattern. The conceptual dimension contains the meaning of the word. The secondary recognition process tries to find the best match between
the sequence of perceptual units in synthesized auditory memory and a representation in long-term memory. The syntactic and semantic rules of the language and situational knowledge might also be utilized at this stage of information processing. The secondary recognition process transforms the synthesized sound pattern into a conceptual meaningful form in generated abstract memory. This memory is called generated abstract memory because the transformation is an active generation rather than a passive derivation and because meaning is abstract rather than modality specific.

We assume that the meaning of a word can be represented by a set of semantic features. Contact with the perceptual code of a word makes available the semantic features that correspond to that word. For example, consider the meaning of the word doctor. The meaning of this word contains specific properties or attributes with respect to a number of semantic dimensions. Some of the dimensions include sex, financial status, social class, place of work, color of working clothes, and so on. When the perceptual code corresponding to doctor is located in long-term memory, some of these conceptual properties are made available in generated abstract memory.

In our model secondary recognition logically follows primary recognition, but in fact the processes can overlap in time. Consider the case in which the primary recognition process is eliminating alternatives and the secondary recognition process is attempting to close off perceptual units into words. It is possible that certain contextual and situational knowledge could facilitate secondary recognition, which then could facilitate primary recognition. Consider the sentence Before going in the house clean the dirt off your shoes. It is possible that, on a perceptual level, the primary recognition process has not resolved whether the last word sounds like shoes or choose. On a semantic level, however, shoes is the only possible alternative. Therefore the secondary recognition process could feed back to the primary recognition process so that the word would actually be heard as shoes. Although the listener usually goes from percept to meaning, conceptual information might actually modify his perceptual experience.

F. Generated Abstract Memory

This storage structure corresponds to the primary, immediate, working memory, or the short-term memory of a number of previous investigators (Atkinson & Shiffrin, 1968; James, 1890; Miller, 1956; Waugh & Norman, 1965). In the present model, the same abstract structure is used to store the meaning of auditory and visual input. Generated abstract memory
has a limited capacity so that forgetting occurs to the extent that new information must be processed (recoded and rehearsed). Forgetting of an item in generated abstract memory is a direct function of the amount of processing of new items. Each new item interferes with memory of old items by an amount that is directly related to the processing of the new item. Therefore forgetting of an item should be a direct function of the number of new items and the amount of time these items are available for processing.

Waugh and Norman (1965) employed a probe recall study in which subjects were presented with a list of items followed by a test item and had to give the item that followed the test item in the preceding list. They varied the rate of presentation of the list and compared the forgetting functions under two different rates. The forgetting function was determined by systematically testing the subject for different items in the preceding list. A list of 15 digits was presented at a rate of 1 or 4 digits per second. Figure 1.4 presents performance as a function of the number of interpolated digits between a digit's original presentation and its test and the rate of presentation. The results show how quickly forgetting occurs at both rates of presentation. The curves through the points, however, show a systematic difference between the forgetting functions under the two rates of presentation.

The function describing forgetting at 4 items/sec starts out lower and ends up higher than the function of forgetting when the items are presented at 1/sec. The y intercept provides some measure of the original perception and storage of the digits, whereas the slope should provide an index of the rate of forgetting. According to this analysis, the items presented at 1/sec were better stored but forgotten faster than the items presented at 4/sec.

These results are compatible with two assumptions about processing in generated abstract memory (Massaro, 1970a). The first is that memory for an item is directly related to the amount of processing of that item. Accordingly, memory for an item will increase with increases in the time a subject has to recode and/or rehearse that item. The second assumption is that memory for an item is inversely related to the amount of processing of other items. A quantification of these assumptions gives the predicted lines in Figure 1.4. As can be seen in the figure, this "limited capacity" rule provides a good description of the acquisition and forgetting in Waugh and Norman's experiment.

We assume that retention of an item in abstract memory and processing of new items are inversely related. Waugh and Norman (1968) showed that a recently presented and redundant (predictable) item does not decrease the memory of earlier items, although new and unpredictable
Figure 1.4. Predicted and observed correct recall probabilities as a function of the number of interpolated items (including the probe) and the rate of presentation. (Data from Waugh & Norman, 1965; after Massaro, 1970a.)

items do interfere with memory. When subjects are able to predict the occurrence of an item, presentation of the item requires little, if any, processing for memory. Thus the lack of processing of predictable items preserves the integrity of earlier items in memory. These results show that forgetting is a direct function of processing new information during the forgetting interval.

G. Recoding and Rehearsal

The recoding process operates on the string of words in generated abstract memory to derive meaning from the entire phrase or sentence. This
operation does not change the nature of the information, allowing it to be recirculated through generated abstract memory. Miller (1956) demonstrated the advantages of a recoding process in remembering a sequence of binary digits. If sequences of 0's and 1's are recoded into sequences of octal digits (0 to 7), the memory span for binary digits can be increased by a factor of three. This follows from the fact that every three binary digits can be recoded into a single octal digit and the memory span has a fixed capacity in terms of number of units or chunks. Therefore, we can remember only $7 \pm 2$ binary digits unless these are recoded into larger chunks. A number of experiments since Miller's have supported the concept of recoding and a fixed memory span measured in number of units or chunks (Cruse & Clifton, 1973; Glanzer & Razel, 1974).

The recoding process has access to the lexicon, syntactic rules, semantic meaning, and whatever other knowledge the system has available. The recoding process can also operate in reverse. Given an abstract idea, it can transform this idea into a sequence of words in generated abstract memory or a perceptual representation in synthesized auditory or visual memory. (The synthesis of a visual representation of a spoken letter is described in more detail in Section IV, C.) Finally, the processing at this stage may be a simple regeneration or repetition of the information, in which case the operation is called rehearsal.

H. Speech Processing

Deriving meaning from the speech signals involves a sequence of successive transformations that give the system larger and larger chunks of information. The feature detection process inputs a short sound pattern made up of a set of acoustic features into preperceptual auditory storage. Preperceptual storage can hold only the most recently presented sound pattern. The primary recognition process transforms this sound pattern into a percept in synthesized auditory memory. The model allows the possibility that the primary recognition process utilizes the phonological rules of the language stored in long-term memory. The size of the perceptual unit read out of preperceptual auditory storage is on the order of a vowel (V), consonant–vowel (CV), or VC syllable. Synthesized auditory memory appears to have a finite capacity of $5 \pm 2$ perceptual units. The secondary recognition process transforms the sequence of syllables in synthesized auditory memory into words in generated abstract memory. The recoding process operates on the words in generated abstract memory to arrive at meaning at the phrase and sentence levels.
Besides the lexicon, the secondary recognition and recoding processes might utilize the contextual information that has already been processed, the syntactic rules of the language, semantic knowledge, and the expectancies of the listener derived from situational context.

These stages of processing can be clarified by a discussion of the necessary processing required for the example sentence *Do you know when it is due?* Assume for the moment that the syllable /du/ (pronounced *do*) is a perceptual unit. If *do* and *due* are homophones in the speaker’s dialect, the primary recognition process would reference the same sign in long-term memory for reading out *do* and *due*. Accordingly, the same percept would be in synthesized auditory memory for both words. However, the listener must be able to disambiguate the meaning of the first sound as *do* and the second as *due*. The percept /du/ would have to be held in synthesized memory until sufficient information about its meaning was obtained. Therefore the capacity of synthesized auditory memory must be large enough to hold the percept /du/ until its meaning is determined.

The secondary recognition process is faced with a string of perceptual units held in synthesized auditory memory. As each unit comes in it tries to close off the most recent units into a word. In our example sentence the first unit /du/ can function as the word *do, dew*, or *due* or the first syllable of the words *doing, doer*, and *duty*, etc. Obviously there is not enough information to decide on the basis of the first syllable alone; however, certain expectations can be formed so that the observer can begin operating with an analysis-by-synthesis routine in the readout of synthesized auditory memory. The readout of the second perceptual unit /ju/ (pronounced *you*) is sufficient to solve some of the ambiguity. In the English language the sequence of perceptual units /du-ju/ does not make up or begin a word. Therefore the first syllable /du/ can be closed off as a separate word before its meaning is determined. The word recognition process cannot yet decide on which alternative meaning of the syllable /du/ is intended. Perceiving the meaning of /du/ could, in fact, be held off until the meaning of the word beginning with the perceptual unit /ju/ is determined.

The analysis of /ju/ proceeds in the same way in that the secondary recognition process tries to close it off into a word. No decision can be made, however, until following perceptual units are analyzed. Although the perceptual unit /ju/ must be the first syllable of a word, the word could be *you, unison*, etc. Expectations can be built up now about both the meaning of the word corresponding to the percept /du/ and the word that begins with the percept /ju/. Already the observer can be said to be operating at the level of a grammatical phrase, since he is using syn-
tactical rules of the language and meaning to facilitate recognition of the meaning of the perceptual units in synthesized auditory memory. The perception of the third perceptual unit /no/ (pronounced no) provides more information. The sequence of syllables /ju-no/ does not constitute an English word. Therefore the perceptual unit /ju/ must be a word, and that word must be you. Therefore the syllable /du/ must mean do, since Do you is the only grammatical phrase conforming to the perceptual unit /du/ preceding the word you.

Some of the initial segmentation of the acoustical signal occurs at the level of preperceptual auditory storage. In Chapter 4 we discuss how the segmentation of the signal is critically dependent on the intensity changes over time of the acoustic pattern. A silent period or a significant change in intensity appears to segment the input and initiate the primary recognition process, the readout of preperceptual auditory store. The CV and VC syllables appear to function as perceptual units so that the readout occurs during the steady-state vowel of the CV syllable and the silent period after a VC syllable. In Chapters 3 and 4 we present a detailed discussion of a number of studies that support the hypothesis that V, CV, and VC syllables function as perceptual units in speech perception.

The perception of CVC syllables can also be interpreted in terms of our model. We assume that CVC syllables usually contain two perceptual units, the CV and VC portions. This is because there appear to be two readouts of preperceptual auditory storage, the CV during the steady-state vowel and the VC after the syllable presentation. Although there are two transformations and transfers to synthesized auditory memory, synthesized auditory memory also holds information of rhythm and intonation that can specify the relationship between adjacent perceptual units. This information allows the word recognition process to distinguish the difference between different sequences of the same perceptual units. For example, the sequence of perceptual units corresponding to green-house and green house are the same. However, the derivation of different meanings for the two representations is facilitated by their different stress patterns (Chomsky & Halle, 1968; Chapter 10, this volume).

The meaning of words can be derived almost simultaneously with the demarcation of word boundaries. Words are represented in a lexicon in long-term memory and contain acoustic, visual, syntactic, and semantic information (Brown & McNeill, 1966). Brown and McNeill’s subjects were given definitions of uncommon words. When the subjects did not know the word defined, they sometimes entered a “tip of the tongue” state. In this state of agony they were able to supply some information about the correct word. Subjects knew the number of syllables of the
word about half the time and also had information about stress location. Since they were sometimes able to give some of the letters in the word, they also had information about the sound of part of the word.

Word recognition appears to occur by a content-addressable lookup rather than a serial search of the words in memory. We assume that a sequence of perceptual units corresponding to the sound of a word has direct access to an analogous code in the lexicon. Stored with this code in long-term memory is the conceptual code or the meaning of the word. Both the content-addressable property and the neighboring storage of perceptual and conceptual word codes is supported by the Stroop color word phenomenon. Stroop (1935) showed that naming the color of colored symbols is disrupted when these symbols spell color names. In this case, the appropriate sequence of written letters is sufficient to bring to mind the name of a color even though it is irrelevant to the task. This phenomenon also supports our idea that word recognition can occur preattentively.

Word meaning alone is not sufficient to decode language, since the same word has different meanings in different contexts. The recoding process tries to close off the words held in generated abstract memory into a meaningful form at the phrase or sentence levels. The overall meaning of a sentence clarifies the meaning of words in the sentence in the same way that the meaning of words clarifies the nature of perceptual units in synthesized auditory memory. Word meaning is modified, changed, and even ignored in the interpretation of a sentence, depending on the situational context. The sentences The chair is comfortable to sit in, The chair recognizes the Senator from Utah, and The chair at Oxford is vacant require different meanings for the word chair. However, since we assume that meaning is determined by a bottom-up rather than a top-down process, the word chair is first recognized and contact is made with those conceptual and perceptual properties that signify chairness. The later sentential context enables the listener to choose the appropriate interpretation of chair so that the meaning of the message is conveyed. If the relevant context precedes the word chair, as in At Oxford, the chair is vacant, the common referent meaning of chair (Olson, 1970) may not come to mind.

The meaning derived from secondary recognition and recoding is stored in generated abstract memory. In the present model, the same structure is used to store the conceptual information derived from both speech and reading. One structure is sufficient, since the information in abstract memory is not modality specific but is in an abstract meaningful form. The recoding and rehearsal process provide the attentive control of the infor-
mation in generated abstract memory. Forgetting occurs in this memory to the extent that new information requires the recoding and rehearsal process (Massaro, 1970a).

As indicated in the previous discussion, syntax and semantics play a critical role in information processing at the level of synthesized and abstract memory. Accordingly, in Chapter 9 a detailed description of contemporary theories of linguistics is presented. Both the syntactically and semantically based theories have important consequences for information processing at the level of synthesized and abstract memory. Chapters 10 and 11 provide a review of psychological studies carried out in the framework of these linguistic theories.

Chapter 10 focuses on the transformation of information from synthesized to generated abstract memory. The problem faced is how the string of perceptual units in synthesized auditory memory is analyzed for meaning. If the listener did not know the language or its grammar, the information processing would stop at this point. How does this knowledge make this important leap to meaning possible? Evidence is presented that supports our assumption of word-by-word processing from synthesized to generated abstract memory. The lexicon as well as semantic and syntactic rules in long-term memory appear to be utilized by the secondary recognition and recoding processes. Since synthesized auditory memory corresponds to that part of the speaker's message that is currently being heard, another critical dimension is the intonation pattern of the perceptual units. These auditory features and others such as rhythm and stress allow the listener to separate syllables, words, and phrases and to determine the syntactic structure of the sentence. The utilization of these auditory features by the secondary recognition process is discussed in Chapter 10.

Most of the other studies of the transformation of perceptual to conceptual information have attempted to keep the auditory information neutral while varying the syntactical structure of the sentence. In these tasks, subjects are asked to locate an extraneous click in an auditory sentence. The assumption is that a click will not be located within the decision units that are functional during the process of deriving meaning from synthesized auditory memory. It follows that click location errors should reveal the functional units at this level. Chapter 10 presents a critical review and interpretation of these studies in the framework of our information-processing model.

Chapter 11 focuses on the role of recoding and generated abstract memory in sentence processing. We assume that abstract memory is limited in capacity and can hold only a small number of discrete or independent units of chunks of information. The syntactic constraints in the message
will therefore affect the storage capacity and processing in generated abstract memory. Accordingly, psycholinguistic studies of syntax and grammar will be discussed in terms of our information-processing model.

IV. VISUAL INFORMATION PROCESSING

Figure 1.1 also presents a flow diagram of the temporal course of processing visual stimuli. Analogous to auditory language processing, feature detection transforms the visual pattern into a set of visual features in preperceptual visual storage. The theoretical mechanisms postulated for auditory feature detection can also handle visual feature detection. This initial transduction of the physical signal into a neurological code places a set of visual features in preperceptual visual storage. The visual perception stage involves a readout of the information in the preperceptual visual image. This process requires an analysis and synthesis of the visual features available in the image. Since the formation of the image and the analysis takes time, it is assumed that the perceptual visual image holds the information in a steady-state form for the primary recognition process.

A. Preperceptual Visual Storage

The visual perception stage involves a readout of the information in the preperceptual visual image. This process requires an analysis of the visual information available in the image. This analysis takes time, and it is assumed that the preperceptual visual image holds the information for recognition to take place. A number of experiments have measured the temporal course of recognition in a visual-recognition-masking paradigm. If a first short visual stimulus produces a preperceptual image that outlasts the stimulus presentation, a second stimulus should interfere with this image and thus interfere with the recognition process.

Figure 1.5 presents the results of word recognition in a backward-masking paradigm (Massaro, 1973). The test word chosen from one of four words was presented for 1 msec followed by a blank interval followed by a visual noise mask over the location of the test word. The independent variable was the duration of the blank interval between the test word and the visual noise mask. The results showed that correct recognition of the test word improved significantly with increases in the blank interval.

These results indicate that some preperceptual visual image of the test letter must have lasted after the stimulus presentation to improve recog-
nition performance with increases in the blank interstimulus interval. The masking noise appears to have terminated perceptual processing of the visual image. Given that recognition performance levels off at about 200 msec, the image may have decayed within this period. Other studies (Averbach & Coriell, 1961; Eriksen & Eriksen, 1971) and evidence presented in Chapter 6 show that information in preperceptual visual storage lasts on the order of 200 to 300 msec.

B. Primary Recognition

Chapter 6 analyzes the perception stage of visual information processing in reading. Since we assume that perception involves the readout of the visual characteristics of letters and words, we first ask what visual features are used in the primary recognition process. Current theories of letter recognition are also evaluated against the available data and in terms of our information-processing model. Given that recognition takes time, we discuss a few demonstrations that a short visual stimulus produces a preperceptual visual image that can outlast the stimulus presentation. These studies also make it possible to determine the temporal course of the perception process. Finally, some evidence is presented that
defines how much of the visual field can be processed in a single eye fixation in reading.

One problem is central to understanding the perception process in reading. What is the perceptual unit of reading; that is, what information makes contact with a sign in long-term memory? Analogous to speech perception, we assume that every perceptual unit in reading has a corresponding sign in long-term memory. The sign contains a description of the distinctive visual features of the perceptual unit and a synthesis program that contains a set of rules or a schema to synthesize the perceptual unit, that is, make it available to synthesized visual memory. Using this model, Chapter 7 directs itself to studies of the perceptual unit employed in reading. We also discuss recent experiments that have shown that the recognition of a letter is improved when it is embedded in a sequence of letters that conform to the rules of English orthography. These studies show that the spelling rules of a language can influence what is synthesized (seen) in a visual display.

Our eyes do not move continuously across a page of text; rather, we make discrete eye fixations at a rate of three or four a second. The movements between eye fixations occur very rapidly so that most of the time the eye is fixated on a portion of the text. It is assumed that the perception process involves a readout of the information in the preperceptual image during the fixation between saccadic eye movements. The next saccadic movement erases the information in this image. The primary recognition process transforms the preperceptual image into a form that is not disrupted by the next eye movement. This transformation presents the system with a synthesized visual representation that can be integrated or combined with the synthesized information from the preceding recognitions from preperceptual visual store.

C. Synthesized Visual Memory

Posner and his colleagues (Posner, Boies, Eichelman, & Taylor, 1969; Posner & Keele, 1967) have studied the contributions of visual and name codes in a same–different reaction time (RT) task. In this task, subjects are presented with a sequence of two letters and asked to report whether the second letter has the same name as the first. The independent variables are the interval separating the two letters and whether the second letter is physically identical to the first letter or simply has the same name. Of course the letters have different names on half the trials in order to keep the subjects honest.

In one experiment (Posner & Keele, 1967) an upper-case letter was followed by either a letter that had the same name, but was either upper-
case or lower-case, or a letter with a different name. For “same” trials the physical matches were 80 msec faster than the name matches when the second letter followed immediately, and this advantage decreased with increases in the interstimulus interval. With a 1.5-sec interval the same RTs did not differ on physical match and name match trials. In another study Posner et al. (1969) showed that the physical match advantage was not peculiar to matching upper-case letters, since lower-case physical matches facilitated RT in the same way.

These results show that presenting two letters with the same name in the same case can decrease the time it takes the observer to determine that they have the same name. Since this advantage disappears very quickly with increases in the interletter interval, subjects probably make their comparison on a strictly name basis at longer interletter intervals. If the second letter is in the same case, does it also facilitate comparison on different trials? Recall, from the Cole et al. (1974) study, that having a second letter in the same voice facilitates both same and different name matches. This result does not obtain in visual letter matches, which indicates that the synthesized visual memory for letter case is much more letter specific than the auditory memory for a speaker’s voice. It is possible that same or different type fonts in reading would be more comparable to same or different voices in speech perception. In this case, presenting two letters in the same font should facilitate performance on both same and different trials.

To what extent is it necessary for the subject to see the first letter to facilitate physical matches? According to our model, visual information can be synthesized without a visual stimulus but through the recoding and rehearsal process. We can all visualize the differences between upper- and lower-case letters. Posner et al. (1969) and Beller (1971) have shown that an auditory cue signifying the case of the letter to be presented can be sufficient to facilitate letter matches. Subjects in the letter comparison task, then, can direct their attention to either the visual or name dimensions, whichever seems to be the best strategy in the particular task. The advantage of the physical matches probably decreases with increases in the interletter interval simply because it is not optimal to operate solely on the basis of visual information, since half of the same trials will be physically different even though they have the same name.

D. Secondary Recognition and Recoding

In reading, integration of synthesized visual information across successive eye movements allows the reader to see more than the information available from one eye fixation. We do not notice the discrete eye move-
ments when reading; rather, the page appears stable and the words seem to appear continuously rather than discretely. This integration allows us to process words and even phrases as apparent wholes. The secondary recognition and the recoding process operate on the information to determine the meaning of the message. Analogous to auditory language processing, syntax and meaning make a critical contribution at this stage of information processing.

E. Reading

The sequence of processes between stimulus and meaning in reading are assumed to be exactly analogous to the processing of speech. The feature detection process transmits visual features into preperceptual visual storage in the form of a preperceptual visual image. The primary recognition process entails a readout of the visual features of the letters in central vision. We assume that the letter is the perceptual unit of analysis and that the recognition process utilizes the spelling rules of the language. Since English orthography is redundant, only partial visual information is necessary to recognize some of the letters. The perception process transforms the preperceptual visual image into a string of letters and spaces in synthesized visual memory. Information from the last couple of fixations can be held in synthesized visual memory so that the page of text appears stable rather than jumping with the discrete jerks of the saccadic eye movements. The information in synthesized visual memory corresponds to what we are seeing at the moment; it provides us with our phenomenological experience of visual perception.

The secondary recognition process operates on the information in synthesized memory to transform the string of letters into a sequence of words. In speech processing, the secondary recognition process utilizes acoustic pauses or intensity changes to facilitate the word segmentation process. In reading, blank spaces and punctuation play this role. The secondary recognition process also has available syntactic rules, a lexicon, semantic knowledge, and expectations generated from the situational context.

The closing off of a string of letters into words is much easier in reading than in speech, since the blank spaces are usually an infallible cue. As mentioned earlier, the lookup in the lexicon appears to have direct-access properties. In this case, the perception of the sequence of letters takes the secondary recognition process directly to the location of the word in memory. This location contains both the perceptual and conceptual properties of the word. It should also be made clear that the word recognition process can be generating expectancies and accessing the lexicon for
a string of letters before the end of the word is reached by the primary recognition process. This is especially true for long words whose letters must be perceived across two or three eye fixations. It has not been demonstrated whether information processed at the secondary recognition stage can facilitate the primary recognition of preperceptual visual storage. The experiments discussed in Chapter 7 suggest that it can.

The successive stages of visual information processing can be seen in processing the example sentence *The cook did not cook today*. Assuming that letters are perceptual units, the same signs would be referenced in perceiving both versions of the word *cook*. Therefore the same information would be available in synthesized visual memory, and the reader would see *cook* in both representations. However, the surrounding context allows the reader to determine the different meanings for the two identical visual representations. In the example sentence *The cook did not cook today*, syntax can be utilized by the word recognition process to disambiguate two different meanings of the word *cook*. The first time *cook* appears it follows the article *the*, which increases the probability of the noun form and eliminates the verb form. The opposite is the case for the second appearance of the word, since the noun form does not follow *did not*. Although the meaning of the word is accessed first in this word-by-word recognition process, its meaning can be modified or changed to agree with the overall context of the phrase, sentence, or situational context. At this point, the sequence of operations becomes exactly the same to those discussed in speech processing, since they occur at the level of generated abstract memory.

Chapter 7 presents and evaluates current theories of reading in terms of our information-processing model and the empirical studies discussed there. The issues are whether phonological mediation is necessary for reading, the role of orthographic rules in recognition, and how semantic/syntactic context facilitates reading. A distinction we make is whether the stimulus to meaning process is mediated or nonmediated. Mediated models, like our own, assume that a sequence of processing stages occurs between stimulus and response. Nonmediated models assume that meaning can be derived from the stimulus directly. The data base developed in Chapter 7 is also used to evaluate these theories.

We believe that an understanding of eye movements during reading can make transparent some of the operations of reading. To this end, Chapter 8 presents a detailed analysis of the characteristics of reading eye movements. These properties of eye movements can be employed to evaluate the nature of information processing within and across successive eye movements. For example, the interfixation distance between eye movements and its variability have direct consequences for the utilization
of peripheral information in reading. Finally, the discussion shows that reading models cannot ignore the oculomotor control of eye movements.

V. CONCLUSION

We assume that there are distinct processing stages between stimulus and meaning in the processing of language. Corresponding to each psychological stage of processing, we have hypothesized a structural representation of the information at that stage of processing. We have presented evidence that supports the assumption of the distinct structures in our model. These studies also illuminate some of the operations of each of the stages of information processing. A more complete discussion of support for the structures and processes in the model can be found in Massaro (1975).

REFERENCES


Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychology Review*, 1956, 63, 81–97.


Articulatory and Acoustic Characteristics of Speech Sounds

Lucinda Wilder

I. INTRODUCTION

When two speakers of a language engage in a normal conversation, communication appears to take place efficiently and automatically. As participants in this situation we are concerned with understanding the ideas our friend is trying to convey without undue attention to the sounds that make up the message. However, it is the speech sounds themselves that travel between speaker and listener. The process of understanding the intent of the speaker begins with the decoding of the continuous stream of acoustic information reaching the ear of the listener. The study of speech perception is concerned with how listeners perceive the information present in the sound system of their language to ultimately arrive at the meaning of an utterance. The first goal of this research is to answer the basic question of what speech sounds are functional in speech perception. Before we can address ourselves to the possible perceptual units of speech and the process by which these units are decoded, we must consider the characteristics of the sound system itself.

The human vocal apparatus is capable of producing an almost infinite variety of speech sounds that can be perceptually distinguished from one another. The sound system of a language contains a subset of the possible perceptual contrasts. All speakers of English will agree that the sound
of the word *pig* is "different" from the sounds of the words *big, peg,* and *pit.* Each word appears to be composed of three different sounds that can be substituted for one another. The stream of speech can be segmented into a series of perceptual contrasts, contrasts that serve to distinguish one utterance from another. Therefore the sound system of a language can be described in terms of perceptual contrasts that have functional significance in that language.

The sound system of a language can also be described with reference to how the speech sounds are produced. In this type of descriptive system, a speech sound is defined in terms of the configuration of the vocal tract that produced it. Different sounds are differentiated by different vocal tract configurations. A third alternative for describing the sound system of a language is in terms of the acoustic properties of speech sounds. A unit that has one set of acoustic characteristics is defined as a single speech sound and can be differentiated from other sounds on the basis of these acoustic properties. In theory, the descriptive systems based on production and acoustic properties of speech sounds need not be structured in terms of the perceptual contrasts present in the sound system of a language. In practice, however, both these types of descriptive systems seek to delineate the production or acoustic correlates of sounds that are defined in terms of perceptual contrasts.

Perceptual contrasts influence the structure of these descriptive systems, yet they are not descriptions of speech perception. Rather, they are a means of classifying the speech sounds to reveal systematic relationships among the various sounds. These descriptive systems may take many forms, and the goal, of course, is to arrive at a system that provides the best description of their properties. At present there is no one completely satisfactory descriptive system of the production or acoustic correlates of speech sounds that can thoroughly and economically account for the contrastive relationships among the sounds.

While a description of the sound system of a language does not provide us with a theory of speech perception, understanding the characteristics of the speech sounds themselves seems a necessary prerequisite to discovering how they are perceived. Perception proceeds from the rich array of acoustical information that reaches the listener, and it is presumably on the basis of this information or some portion of this information that the listener begins to segment, decode, and attach meaning to the continuous information arriving over time. Adequate description of the acoustic characteristics of the speech signal must precede the search for the specific acoustic cues that enable the listener to perceive the various speech sounds.

This chapter is intended to provide the reader with an understanding
of the acoustic characteristics of the various English speech sounds. Because the acoustic characteristics of the speech signal are a consequence of the way the component sounds are produced by the vocal apparatus, this chapter will also discuss production of the speech sounds. This chapter differs from those that follow in that it does not treat the way in which information is processed by the language user but, rather, describes the production and characteristics of the stimulus information itself. The chapter was deemed necessary because the stimulus information available in spoken language is not as generally familiar nor intuitively obvious as, for example, the stimulus information present in a printed page of text. Those readers who already possess a basic knowledge of the production and acoustic characteristics of the speech signal may comfortably proceed to the following chapter, which deals with the acoustic cues that are used to recognize individual speech sounds.

We will first consider the basic mechanisms by which sound is produced with the human vocal apparatus and the general acoustic characteristics of speech sounds. Then a more detailed description of the production of discrete speech sounds will be presented. This will be followed by a brief discussion of the occurrence of speech sounds in natural language. The individual sounds of English will then be described in terms of their articulatory and acoustic characteristics. A brief discussion of the feature analytic approach to the description of speech sounds and a consideration of individual differences among speakers are then presented.

II. PRODUCTION OF SPEECH SOUNDS

The production of speech sounds, like the production of all sounds, depends on three factors: (1) a source of energy, (2) a vibrating body, and (3) a resonator. All these components are present in the group of body parts known collectively as the speech organs. This system, shown in Figure 2.1, consists of the lungs; the trachea; the larynx, which contains the vocal folds; the pharynx; the mouth or oral cavity; and the nasal cavity. This system will be discussed only briefly; a more complete description of the anatomy and physiology of the various speech organs can be found in Zemlin (1968). Although the relationship between respiration and phonation is complex (cf. Hixon, 1972), an oversimplification views the outgoing breath stream from the lungs as the source of energy for the production of speech sounds. The vocal folds can act as vibrators when set in motion by the breath stream. The ensuing vocal sound is then modified by the resonance chambers of the pharynx, oral cavity, and nasal cavity.
The **vocal folds** are two elastic bands of tissue attached to the side walls of the larynx and stretched from front to back. The back wall consists of movable cartilages (the arytenoid cartilages); when the position
of the cartilages is changed, the vocal folds are drawn apart or pulled together. The opening between the vocal folds is called the glottis. Three positions of the vocal folds are shown in Figure 2.2. Part a shows the position of the vocal folds in normal breathing. The glottis is maximally open, allowing air to pass unimpeded and inaudibly. Part b shows the glottis completely closed. In this position the air stream is obstructed. Part c shows the position of the vocal folds during voicing. The glottis is partially open, allowing air from the lungs to escape, but the outgoing breath causes the folds to vibrate.

The frequency of vibration is controlled by the degree of tension in the vocal folds and their mass. The subjective pitch quality of the voice is determined by the frequency of vocal fold vibration. The larger mass of the male vocal folds results in a lower frequency of vibration than the smaller mass of the female vocal folds. An individual can alter the pitch of her voice by varying the tension of the vocal folds. Increased tension leads to increased frequency of vibration, and decreased tension leads to lower frequency of vibration. The sound pattern produced by raw vocal fold vibration is then modified by the resonance chambers. The size and shape of the combined pharyngeal, oral, and nasal cavities can be altered, and it is these alterations that determine the acoustic characteristics of the speech sounds. The general process by which these cavities are altered is called articulation. The articulatory process will be considered in detail when production of the various speech sounds is discussed.

Vibration of the vocal folds is one way in which the energy from the breath stream can be used to produce an audible sound which is modified by the resonance chambers. However, there are several other means of producing sound with the vocal organs. One way is to force the breath stream through a narrow passage in the vocal tract above the larynx. This constriction creates turbulence in the airstream, and the turbulence produces an audible hissing sound. Another way is to completely obstruct the vocal tract, momentarily allowing pressure from the breath stream
to build up behind this point. Sudden release of the pressure creates a small explosion of sound. Like the sound produced by vocal fold vibration, that produced by constriction or obstruction of the vocal tract can be modified by the resonance chambers. Vocal fold vibration may accompany constriction or obstruction, since the operation of the periodic and noise sound sources is semiindependent.

III. GENERAL ACOUSTIC PROPERTIES OF SPEECH SOUNDS

When the vocal folds are set in motion by the outgoing breath stream, they vibrate with a certain frequency. The frequency of vibration in normal speech varies from about 60 to 350 Hz. The way in which the vocal folds vibrate causes pressure variations in the breath stream. If the amplitude of these pressure variations were plotted over time, we could observe the shape of the wave creating the pressure changes. The waveshape resulting from vocal fold vibration is described as complex and periodic. It is labeled periodic because the pattern of pressure changes repeats itself over time. The form of the pattern is not sinusoidal in appearance and therefore is labeled complex. However, it can be shown by Fourier analysis that a complex periodic wave is composed of a number of sinusoidal components of different frequencies, amplitudes, and phase. An amplitude spectrum of such a wave presents the component sinusoidal frequencies and their amplitudes in graphic form without regard to the phase of the individual components. The spectrum of vocal fold vibration would show a series of frequency components (harmonics), each of which was an integer multiple of the lowest frequency component (fundamental frequency). The fundamental frequency is the frequency of vocal fold vibration. The second harmonic is the component whose frequency is twice the fundamental frequency; the third harmonic is the component whose frequency is three times that of the fundamental, and so on. The vocal cord pulses and the corresponding amplitude spectrum are shown in the left-hand part of Figure 2.3. The duration in seconds ($T_o$) of one period of vocal cord pulses (time from opening to opening of the glottis) is the inverse of the fundamental frequency $F_o = 1/T_o$. The amplitude of the various harmonics present in the source spectrum $S(f)$ drops at the rate of 12 dB/octave.

The complex periodic wave resulting from vocal fold vibration is then modified by the resonance chambers of the upper vocal tract. A general property of resonators is that they respond differentially to vibrations of different frequencies. The amplitude of frequencies that are at or near the preferred (natural) frequency or frequencies of the resonator are rein-
forced, while the remaining frequencies are damped. The frequency response curve for the vocal tract configuration that produces the vowel “ah” is shown in the box marked “vocal transmission” in Figure 2.3. The peaks represent the natural frequencies. When the sound source from the vocal folds is passed through the vocal tract resonator, the amplitude of each of the harmonics $S(f)$ is multiplied by the value of the transfer function $T(f)$ at that frequency. The product is the spectrum of the radiated sound, which is shown at the right of the figure. The radiated wave itself is also shown.

The natural resonances of the vocal tract are called formants. As the size and shape of the resonance cavities are altered during speech production, these formant frequencies are also changed so that every configuration of the vocal tract has its characteristic formant frequencies. The peaks (points of highest energy concentration) present in the spectrum of a speech sound are thus a function the formant frequencies of the upper vocal tract and not the frequency of vocal fold vibration. These spectral peaks are not coincident with any specific harmonic present in the original vocal fold vibration. The independence of the harmonics and formant frequencies is illustrated in Figure 2.4. The figure shows the wave shapes and corresponding spectra of the vowel “ah” pronounced with the frequency of vocal fold vibration equal to 90 Hz (panel a) or 150 Hz (panel b). The fundamental frequency and the harmonics are indicated by the vertical lines at each component frequency. The spectral peaks are determined by the formant frequencies. The peak of lowest frequency (first formant) is approximately 750 Hz for both speech waves, even

Figure 2.3. Schematic representation of the relationships among the sound source, the vocal tract resonances, and the radiated wave. (From Fant, G. Analysis and synthesis of speech processes. In B. Malmberg (Ed.), Manual of phonetics. The Hague: Mouton & Co., 1960.)
though this is accomplished by relative enhancement of the eighth and fifth harmonics in panels a and b, respectively. The occurrence of spectral peaks of successively higher frequencies that correspond to the second and third formants are similarly coincident in the two panels. The spectra of speech waves are thus characterized by concentrations of energy in frequency regions that correspond to the formant frequencies or natural resonances of the vocal tract.

The preceding discussion has treated the acoustic characteristics of voiced speech sounds, i.e., those sounds produced with vocal fold vibration. However, speech sounds can be produced without vocal fold vibration by creating turbulence in the breath stream as it passes through the vocal tract. For example, the sound "sh" is produced by forcing the outgoing airstream through a constriction formed by the tongue and the roof of the mouth. Vibration produced in this manner can be described as a complex aperiodic wave. It is labeled aperiodic because the pattern
of pressure variations does not repeat over time. A characteristic of such waves is that they can have energy components at all frequencies rather than only at multiples of the fundamental frequency. Assume that the energy level is the same across the entire frequency range. Although the “ah” sound is produced relatively close to the opening of the vocal tract (lips), the small part of the vocal tract anterior to the place of production can and does act as a resonator. Thus the amplitude of the frequency components at the natural resonance(s) are enhanced. The spectral peaks are determined by the natural resonance(s) of the vocal tract for aperiodic as well as periodic sound sources.

Although spectral peaks in the radiated wave can arise as a function of extreme irregularities in the source spectrum, the spectral peaks of speech waves are generally equated with the formant frequencies of the vocal tract resonator. Given knowledge of the exact cross-sectional dimensions of the vocal tract resonator and the waveform produced at the sound source, it would be mathematically possible to predict the spectra of the resultant speech wave and to differentiate the spectral peaks produced by the formants from those produced by the source spectrum. In practice the spectral peaks are often assumed to correspond to the formants, and much of our knowledge of the acoustic characteristics of speech sounds comes from observation of the spectra that result from different vocal tract configurations rather than specifying how the spectra were determined by the configurations. A more sophisticated discussion of the acoustic properties of speech sounds may be found in Fant (1968).

The spectrum of the “ah” sound shown in Figure 2.4 presents the average energy present at each of the frequency components over some period of time. When the vowel “ah” is pronounced in isolation and sustained, there is little, if any, change in the vocal tract configuration and the resultant sound wave over the course of its duration. Therefore the average amplitude spectrum provides an adequate representation of the acoustic characteristics of the isolated vowel. When we want to observe the acoustic characteristics of individual sounds as they occur in normal speech, the simple amplitude spectrum is no longer adequate, for it will display the average energy at each frequency of all the sounds in the sample. What is needed is a display of the instantaneous spectral changes in the sample. A special machine called a sound spectrograph has been developed for this purpose. A diagram of the spectrograph is shown in Figure 2.5.

A short sample of speech up to 2.4 sec long is recorded on magnetic tape, which is then fed through an adjustable band pass filter. The filter allows energy in the frequency region between its upper and lower cutoff points to pass unimpeded, while energy in other frequency regions does
not pass through the filter. As the tape loop is repeated, the energy concentration in the passed frequency region is recorded as a function of time since the beginning of the sample. The cutoff points of the filter are then readjusted, and the process is repeated until the entire frequency range between 0 and 3500 Hz has been analyzed. The result is a spectrogram that displays the amplitude of the energy (intensity) present in each frequency band as a function of time. The variable filter may be adjusted to have a narrow bandwidth of approximately 45 Hz or a wide bandwidth of about 300 Hz. In both cases the lower cutoff point of the band is shifted upward approximately 15 Hz for each repetition of the speech sample so that 200 repetitions of the tape loop are required to analyze the whole frequency range. The size of the upward shift in Hz varies with different frequency ranges.

Examples of narrow- and wide-band spectrograms are shown in Figure 2.6. Time is represented on the horizontal axis, frequency on the vertical axis, and intensity by the shade of darkness. The narrow-band spectrogram shows fine details such as the harmonics associated with vocal fold vibration. The broad-band spectrogram provides a grosser picture of the acoustic characteristics of the complex wave. However, the formants or resonance bars are more readily distinguished on the wide-band spectrogram. These regions of high energy concentration are analogous to the spectral peaks observed in the average amplitude spectrum of the vowel "ah." Because of the interest in the relationship between vocal tract configuration and formant frequencies, wide-band spectrograms are typically used to study the acoustic characteristics of speech sounds. Further de-
Figure 2.6. Sound spectrograms of the words *Speech we may see*. The top panel was produced with a narrow-band analyzing filter (45 Hz) to portray harmonic structure. The bottom panel was produced with a wide-band analyzing filter (300 Hz) to emphasize vocal resonances. (From *Visible speech* by R. K. Potter, G. A. Kopp, and H. G. Kopp. Dover Publications, Inc., New York, 1966. Reprinted through the permission of the publisher.)

tails of the spectrographic pattern will be discussed when we consider the individual speech sounds.

The analysis of speech with the spectrograph provides acoustic information about the various speech sounds. The synthesis of speech sounds provides information on the relationship between the acoustic pattern and the perception of speech sounds. The investigator is often interested in determining what acoustic information is necessary for perception of a given speech sound. One way to study this relationship is to synthesize
speech with a machine developed at Haskins Laboratories called a **pattern playback**. A schematic diagram of the playback is shown in Figure 2.7. A light source passes through a tone wheel that “produces 50 bands of light, modulated at harmonically related frequencies that range from a fundamental of 120 Hz through the fiftieth harmonic at 6000 Hz [Liberman, Delattre, Cooper, & Gerstman, 1954].” The light bands are arranged to match the frequency scale of the spectrogram. When a hand-painted spectrogram like that shown in Figure 2.7 is fed under the lights, the painted frequency bands reflect light of the corresponding frequency and this energy is converted into sound. By varying the spectrographic pattern the investigator can determine which frequency combinations are required to produce a speech sound. Figure 2.7 shows that the syllables *di*, *da*, and *do* can be synthesized from steady-state concentrations of energy at frequency regions that correspond to the first and second formants when the steady state portion is preceded by an abrupt formant transition. The steady-state energy concentrations are sufficient to produce the vowel sounds of these syllables and would produce these vowel sounds if they were not preceded by the abrupt transition. The part of

*Figure 2.7.* Schematic diagram of the pattern playback and spectrographic patterns that produce the syllables *di*, *da*, and *do*. (From Liberman, A. M., Delattre, P., & Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, 65, 437.)
the painted pattern responsible for producing the consonant /d/ cannot be as easily specified. Where the transition precedes the steady state portion, the entire syllable is synthesized. When an isolated transition pattern is played, a rising or falling whistle is produced. The use of the pattern playback as a tool to determine what kind of acoustic cues are necessary in the perception of speech sounds will be discussed further in Chapter 3. Haskins Laboratory has now developed a computer-controlled speech synthesizer that is much faster and more flexible than the pattern playback device (Cooper & Mattingly, 1969; Mattingly, 1968).

IV. ARTICULATION OF SPEECH SOUNDS

Articulation is the process by which the configuration of the vocal tract is modified to produce the various speech sounds. In describing the articulatory process it is helpful to distinguish between vowels and consonants. In the articulation of vowels the tongue assumes one of a large variety of positions in the mouth, the lips are opened, and the arytenoid cartilages of the larynx are adducted so as to produce vocal fold vibration when the outgoing breath stream passes through the vocal tract. Although the position of the tongue actually varies along several dimensions, tongue position is described with reference to the location and height of its highest part. This point can occur in the front, central, or back parts of the mouth. Within these three regions the highest point of the tongue can be in the high, mid, or low part of the mouth.

The articulation of consonants can best be described in terms of their place of articulation, their manner of articulation, and whether they are accompanied by vocal fold vibration. The place of articulation may be defined as the point of maximum closure in the vocal tract while the speech sound is produced. Closure is effected as an articulator approaches or makes contact with a point of articulation. Articulators are movable parts of the oral cavity; there are only two of them—the lower lip and the tongue. The tongue is by far the more versatile because of the many positions it can assume. To more precisely describe the part of the tongue involved in an articulation, five regions of the tongue are distinguished; these are shown in Figure 2.8. The five regions are the tip or apex, the front, the center, the dorsum, and the root.

Points of articulation are immovable parts of the oral cavity; they are shown in Figure 2.9. These parts are the upper lip, the upper teeth, the alveolar ridge, the palate, the velum, the uvula, and the lower teeth. The point of maximum closure is described by naming both the articulator and the point of articulation. Some examples are given in Table 2.1.
Figure 2.8. Regions of the tongue: tip or apex, front, center, dorsum, root.

Figure 2.9. Points of articulation: upper lip, upper teeth, alveolar ridge, palate, velum, uvula, lower teeth.

<table>
<thead>
<tr>
<th>Name</th>
<th>Articulator</th>
<th>Point of articulation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>bilabial</td>
<td>lower lip</td>
<td>upper lip</td>
<td>pin</td>
</tr>
<tr>
<td>labiodental</td>
<td>lower lip</td>
<td>upper teeth</td>
<td>fin</td>
</tr>
<tr>
<td>interdental</td>
<td>tongue apex</td>
<td>upper teeth</td>
<td>thin</td>
</tr>
<tr>
<td>(apico–dental)</td>
<td>tongue apex</td>
<td>alveolar ridge</td>
<td>tin</td>
</tr>
<tr>
<td>apico–alveolar</td>
<td>tongue dorsum</td>
<td>velum</td>
<td>kin</td>
</tr>
<tr>
<td>dorso–velar</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The manner of articulation describes how a speech sound is produced. All vowel sounds are produced in the same way; i.e., the outgoing breath stream has relatively free passage through the vocal tract and the sound resulting from vocal fold vibration is modified by the resonance cavities. Consonant speech sounds may be produced in several different ways. In the articulation of a stop, the escape of the breath stream is completely impeded at the place of articulation. Pressure momentarily builds up behind this point, and then the pressure is released with a small explosion of sound. For this reason, stop consonants are often called plosives. The point of occlusion of the vocal tract can be any of the points of articulation described earlier.

Another way in which consonant sounds can be produced is to force the escaping breath stream through a small passage or constriction in the vocal tract. A sound produced in this manner is characterized by audible friction and is called a fricative. The point of constriction can occur at any of the points of articulation described earlier. In the production of stops and fricatives, audible sound is created other than by vocal fold vibration. Because of the absence of vocal fold activity, these sounds are labeled unvoiced. However, the production of both stops and fricatives may be accompanied by vocal fold vibration; i.e., they may be voiced. In the case of voiced stops and fricatives, there are two sound sources: the sound produced by the vocal folds and that produced by occlusion or constriction of the vocal tract. In both unvoiced and voiced stops and fricatives, the resulting sound is resonated exclusively in the oral cavity.

Sounds may also be resonated in the nasal cavity by allowing the breath stream to pass through the nasal cavity as well as the oral cavity. Passage of the breath stream into the nasal cavity is controlled by the velum. When the velum is lowered, the escaping air resonates in both the oral and nasal cavities. The positions of the velum are shown in Figure 2.10. Vowels are most often resonated in the oral cavity. When they are produced with the velum lowered, they are referred to as nasalized vowels.

One group of consonant sounds is always produced with the velum lowered, and their manner of articulation is thus described as nasal. Nasals are produced by occluding the oral cavity at some point of articulation and allowing the breath stream to escape through the nasal cavity. Nasal consonants in English are always voiced.

The final manner in which consonant sounds may be articulated is by changing the place of articulation during the course of their production. Sounds with a varying place of articulation are called glides. The sound source for glides is vocal fold vibration, and the configuration of the vocal