Information Processing and Speech Perception*

David B. Pisoni

Indiana University

Bloomington, Indiana 47401

This paper discusses a number of important concepts within the framework of human information processing and their relevance to speech sound perception. A rough outline of a model of speech perception is described which incorporates the distinctions between sensory, short and long term memory and hierarchical stages of processing. Central to this approach is the continuity of processing and the interrelations between stages of analysis.

* This paper was prepared for the Speech Communication Seminar, Stockholm, August 1 - 3, 1974 and will appear in the proceedings which are to be published by Almqvist & Wiksell and John Wiley & Sons. This work was supported in part by USPHS NIMH Research Grant MH-24027-01 and in part by a Faculty Fellowship from the office of Research & Advanced Studies, Indiana University.

Information Processing and Speech Perception

David B. Pisoni
Indiana University
Bloomington, Indiana 47401 U.S.A.

Introduction

Current theories of speech perception have been quite general and vague, and for the most part, not terribly well developed (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Stevens & House, 1972). Indeed, it is probably fair to say that most of the current theoretical approaches to speech perception are only preliminary "guesses" at what a possible model of the speech perception process might entail. A few quotes should make this point clear. For example, in 1964 Liberman and his colleagues stated at the Symposium on Models for the Perception of Speech and Visual Form that:

> "Since this symposium is concerned with models, we should say at the outset that we do not have a model in the strict sence, though we are in search of one." (Liberman, Cooper, Harris, MacNeilage & Studdert-Kennedy, 1967, p. 68).

At the same meeting, Fant stated that:

> "Any attempt to propose a model for the perception of speech is deemed to become highly speculative in character and the present contribution is no exception." (Fant, 1967, p. 111).

And even more recently, Stevens & House stated in their chapter on Speech Perception that:

> "Since we are still far from an understanding of the neurophysiological processes involved, any model that can be proposed must be a functional model, and one can only speculate on the relation between components of the functional model and the neural events at the periphery of the auditory system and in the central nervous system." (Stevens & House, 1972, p. 47).

Although most investigators agree that the perception of speech sounds may involve processes and mechanisms that are in some way basically different from those employed in the perception of other sounds, very little work has been directed at specifying these differences. A large body of experimental work obtained over the last twenty years suggests that when listeners are presented with speech stimuli their ability to identify and discriminate these sounds on an auditory basis alone is limited to a very substantial degree by their linguistic knowledge. Differences in the perception of speech and non-speech stimuli and differences in perception among various classes of speech sounds have led numerous investigators to propose a special "speech perception mode" (Stevens & Halle, 1967; Liberman, 1970a,b). Other findings employing dichotic listening techniques have suggested that a specialized perceptual mechanism--"a special speech decoder" may exist as a distinct physiological entity for the processing of speech sounds (Studdert-Kennedy & Shankweiler, 1970). Other evidence has been accumulated to suggest that speech perception may involve some sort of active mediation of motor centers associated with speech production (Liberman, Cooper, Harris & MacNeilage, 1963).

In this paper, I consider some of the perceptual processes involved in speech recognition and then describe a rough model for speech sound perception based on recent work in human information processing.

Information Processing Approach

In recent years the study of speech perception has begun to adapt the aims and methods of human information processing models which have been employed quite successfully in the study of visual and auditory perception (Neisser, 1967; Haber, 1969; Massaro, 1972; Reed, 1973). This approach

views perception as a hierarchically organized sequence of events involving stages of storage and transformations of information over time. As Neisser points out, during these stages information is "transformed, reduced, elaborated, stored, recovered and used." A major assumption of this approach to perception is the continuity of different levels of processing. Sensation, perception, memory and thought are considered to be on a continuum of cognitive activity. These stages are thought to be mutually inter-dependent. Furthermore, it is argued that one can only understand perception, especially recognition, identification and perceptual memory by attempting to understand the whole range of these cognitive processes.

Since the information processing approach is fundamental to our approach to speech perception, I will first describe some of the major concepts involved. Then I will describe some of the stages of processing speech perception and then provide some of the details of the current model.

There are three basic assumptions in current information processing models. First, perception is not immediate but is the outcome of distinct operations distributed over time. One goal of information processing models is to attempt to specify the operations which occur from the presentation of a stimulus to the overt response of the observer. The various stages which lie between input and output are typically represented by a flow chart with block design. Much of the recent work on backward masking has been concerned with this question (Pisoni, 1972, 1974; Massaro, 1974).

The second assumption is that there are "capacity limitations" at various stages of processing. Because the nervous system cannot maintain all aspects of sensory stimulations and must integrate energy over time, limits on the

capacity to store and process sensory data occur which require that information be recoded into a different more abstract form. One goal of research in this area has been to identify the locus of these capacity limitations. For example, the recent work by Shiffrin and myself has been specifically directed at this problem (see Shiffrin, Pisoni & Castenada-Mendez, 1974).

The third assumption is that perception necessarily involves various types of memorial processes since recoding and retention of information will occur at all stages of information processing. Hence, the study of speech perception necessarily entails the study of perceptual memory. The work of Fujisaki and myself has shown the importance of short-term memory in speech perception (Fujisaki & Kawashima, 1970; Pisoni, 1971, 1973).

Sensory Memory, Short-term Memory and Long-term Memory

Central to information processing analyses is the motion of an iconic, echoic, or pre-perceptual memory store. This is typically thought of as a very temporary storage medium which preserves all of the stimulus information in a literal or veridical form for several hundred milliseconds. During this time period, the information is converted into a more persistent and abstract form for representation in short-term storage. Short-term (STS) or "working memory" is thought to have a very limited capacity from which information is rapidly lost unless active rehearsal or control processes are operating. Long-term store (LTS) on the other hand, is assumed to be the permanent repository for information. It has an unlimited capacity. Long-term store receives information from short-term store. The process of rehearsal of information in short-term store first regenerates the rapidly decaying memory traces and also causes information in short-term store to be transferred to long-term store.

Recognition is assumed to be a process whereby the sensory input or some derived version of it "makes contact" with a stored representation in long-term memory or some type of representation that has been constructed or generated by rules in long-term memory. Thus, recognition is assumed to take place in short-term memory. The information present in short-term memory is thought to consist of a combination of information from both the sensory input and information from long-term memory.

Sensory information is not simply transferred to short-term store but is "recoded" while still being maintained at the earliest stage of processing. It is generally assumed that the earliest stages of the recognition process occur "automatically" and without conscious control by the subject (see for example, Shiffrin & Geisler, 1973). A good part of the information from the earliest stages of processing is lost by decay and only a relatively abstract representation of the input is maintained in short-term memory.

Stages of Processing in Speech Perception

A number of recent accounts of speech perception have begun to emphasize process and to divide this process into a hierarchy of stages: auditory, phonetic, phonological etc. (see for example, Liberman, 1970; Studdert-Kennedy, 1974a,b; Studdert-Kennedy, Shankweiler & Pisoni, 1972; Wood, 1973). Figure 1 shows some of the processes which are assumed to take place between the initial

---
Insert Figure 1 about here
---

acoustic signal and its final conceptual representation. According to this view, the speech signal undergoes a series of successive transformations whereby information is recoded into more and more abstract forms of representation

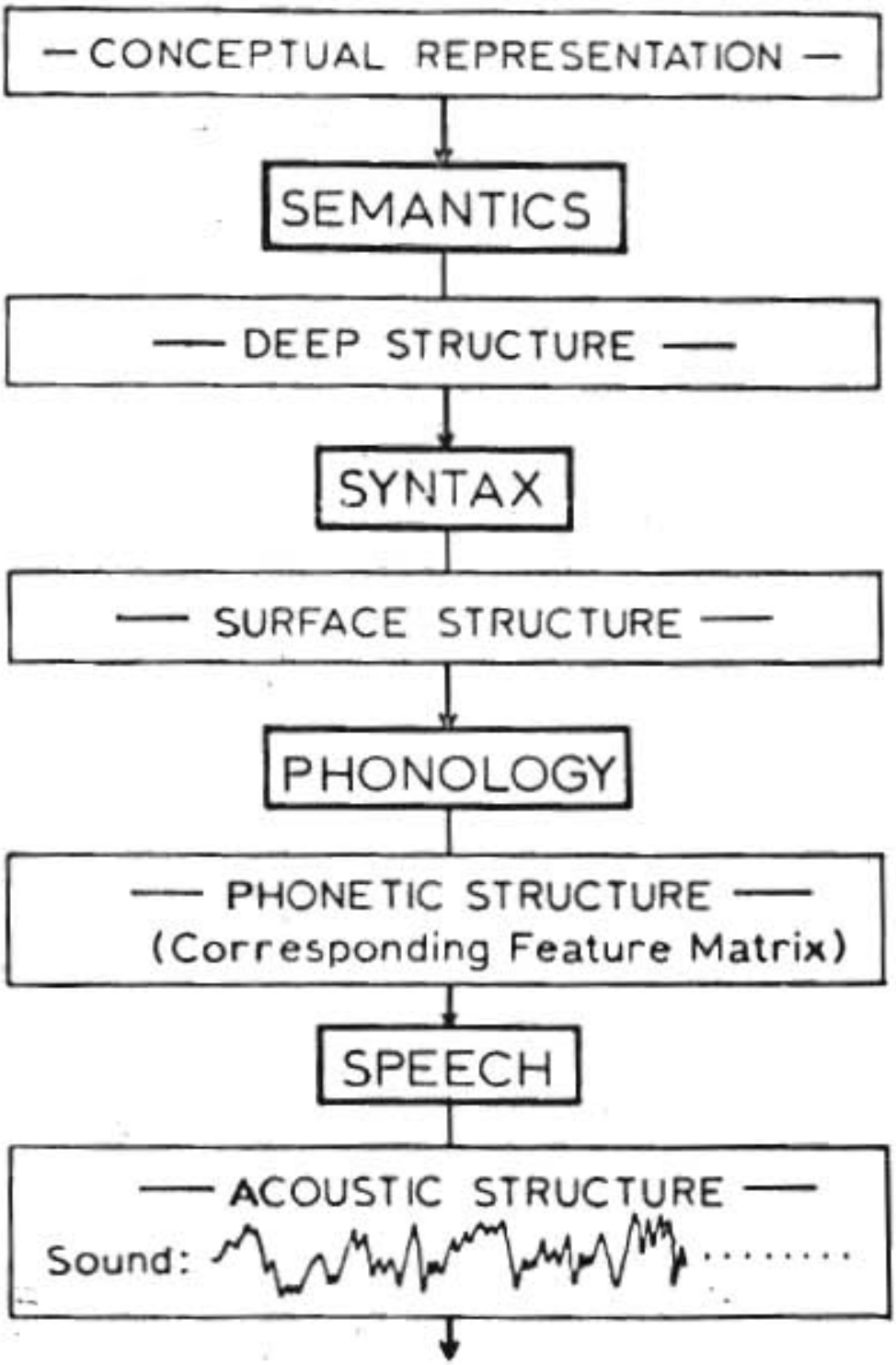
— CONCEPTUAL REPRESENTATION —
SEMANTICS
— DEEP STRUCTURE —
SYNTAX
— SURFACE STRUCTURE —
PHONOLOGY
— PHONETIC STRUCTURE —
(Corresponding Feature Matrix)
SPEECH
— ACOUSTIC STRUCTURE —
Sound:

Figure 1

(see also Liberman, 1970). The stages are thought to be partially successive since spoken language is inherently a temporal phenomena. However, decisions at various stages must also take place in parallel to permit information from higher levels to be employed in processes at lower levels.

Although the distinction between phonetic structure and higher levels of analysis is commonly accepted in linguistic theory, the distinction between auditory (i.e., acoustic structure) and phonetic levels of analysis has not been widely recognized. The auditory stage may be thought of as the first level of analysis. At this stage the acoustic waveform is transformed or recoded into some "time-varying" neurological pattern of events in the auditory system. Acoustic information such as spectral structure, fundamental frequency, intensity and duration is extracted by the auditory system. All subsequent stages of analysis beyond the auditory stage are thought to be abstract and based on these "auditory features." The phonetic stage is closely related to auditory analysis. Here, segments and features necessary for phonetic classification are abstracted or derived from the auditory features of the acoustic signal. At the output of this stage, the continuously varying acoustic stimulus has been transformed into a sequence of discrete phonetic segments. Information about the feature specification of these phonetic segments is then passed on to higher levels of processing for phonological and syntactic analysis.

Thus, the auditory level may be characterized as that portion of the speech perception process which is "non-linguistic," and therefore includes processes and mechanisms that operate on speech and non-speech signals alike. On the other hand, processes and mechanisms at the phonetic level are assumed

to perform a linguistic abstraction process whereby a particular phonetic feature is identified or recognized from some configuration of auditory features.

An Information Processing Model

In this section, I will briefly sketch the structure of the information processing model. Figure 2 shows a block diagram of the components. Auditory

---

Insert Figure 2 about here

---

input enters the system and is processed in progressive stages. The output of Preliminary Auditory Analysis is assumed to be some type of spectral display in terms of frequency, time, and intensity. Sensory input is processed automatically through several levels of analysis without the operation of conscious selective attention. Sensory information is maintained in a relatively gross unanalyzed form in the Sensory Information Store (SIS). Information is further processed by a "recognition device" which is shown as four distinct stages in this figure. Information from any or all of these stages of processing is placed in short-term store where the subject can selectively rehearse, encode or make decisions about it. It is assumed that information in long-term store is employed in the recognition process.

Automatic processing by the recognition device is assumed to take place as follows. In Stage 1, Acoustic Feature Analysis, we assume that auditory features of the speech signal are recognized by a system of individual auditory feature detectors (Stevens, 1973). For example, in the case of a simple CV syllable, we assume that specialized auditory detectors will respond selectively to at least some of the following types of information: (a) presence or absence

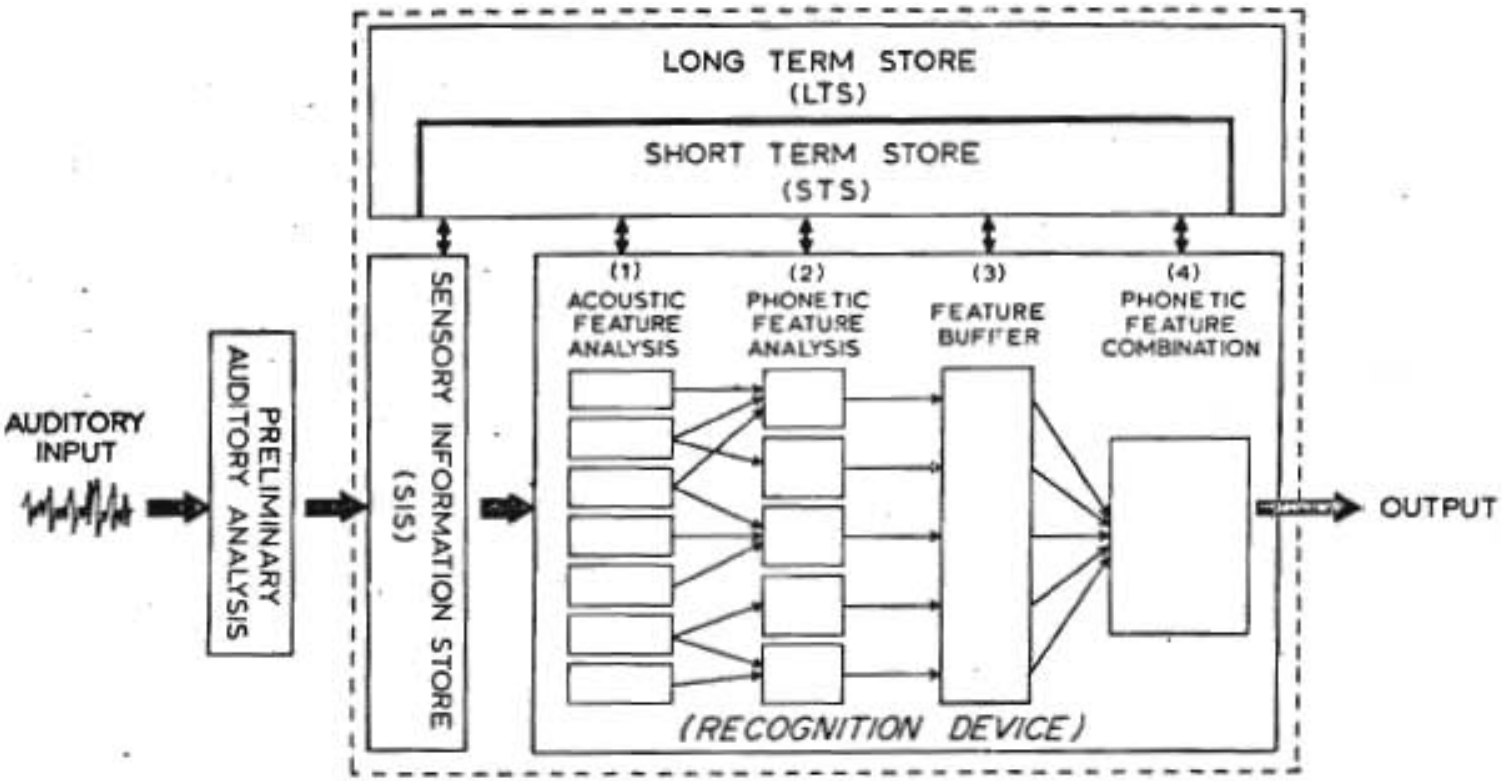
LONG TERM STORE
(LTS)
SHORT TERM STORE
(STS)
AUDITORY
INPUT
PRELIMINARY
AUDITORY ANALYSIS
SENSORY INFORMATION STORE
(SIS)
(1)
ACOUSTIC
FEATURE
ANALYSIS
(2)
PHONETIC
FEATURE
ANALYSIS
(3)
FEATURE
BUFFER
(4)
PHONETIC
FEATURE
COMBINATION
OUTPUT
(RECOGNITION DEVICE)

Figure 2

of a rapid change in the spectrum; (b) direction, extent, and duration of a change in the spectrum; (c) duration and intensity of noise, etc. The output of Acoustic Feature Analysis is some set of acoustic cues or auditory features which forms the input to the next stage of processing.

In Stage 2, Phonetic Feature Analysis, we assume that a set of decision rules is employed to map multiple auditory features into phonetic features. It is assumed that this is a many-to-one mapping where several different auditory features provide information about a particular phonetic feature. The output of phonetic feature analysis is a set of abstract phonetic features. These decision rules can be thought of as having knowledge of articulatory constraints in production although it is not essential for the model in its present form.

These features are subsequently maintained in Stage 3, the Feature Buffer. This may be thought of as simply a holding mechanism which maintains decisions about the feature composition of a particular syllable. There are two reasons for postulating a feature buffer. First, not all phonetic features are assumed to be processed (i.e., recognized) at the same rate. Secondly, some memorial process is needed to preserve and maintain phonetic feature information more-or-less independently for subsequent stages of linguistic processing (e.g., phonological).

Feature information is then used in Stage 4, Phonetic Feature Combination, where individual features are recombined to form discrete phonetic segments. The output of Stage 4 is a phonetic segment, where the feature specification is, for example, some form of an abstract distinctive feature matrix. This information is then passed on to higher levels of processing for phonological and syntactic analysis.

The model as I have described it thus far is still preliminary and a number of changes and revisions will obviously be required. However, I think it has much to offer as a framework for dealing with past research and providing a basis for future work. It combines the virtues of recent information processing models with their emphasis on stages of processing, memory and recoding of information. It also incorporates the recent distinctions between auditory and phonetic stages of processing in speech perception. Finally, I think such a model can be used to generate new and important questions about speech perception that can be tested empirically. For example, what is the general organization of auditory and phonetic stages of processing and the nature of the interaction between them? What is the locus or stage of processing at which processing peculiar to speech is initiated and what is the nature of these perceptual operations?

In a more global sense, the model can be used to specify the ways in which speech sounds may require specialized neural mechanisms for perceptual processing and the ways it may conform to more general principles of human information processing common to other modalities.

References

Fant, G. Auditory patterns of speech. In W. Wathen-Dunn (Ed.) Models for the Perception of Speech and Visual Form. Cambridge, Mass.: M.I.T. Press, 1967.

Fujisaki, H. and Kawashima, T. Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute, Vol. 29, Faculty of Engineering, University of Tokyo, Tokyo, 1970, 207-214.

Haber, R.N. Information-Processing Approaches to Visual Perception. New York: Holt, Rinehart and Winston, 1969.

Liberman, A.M. The grammars of speech and language. Cognitive Psychology, 1970, 1, 301-323.

Liberman, A.M. Some characteristics of perception in the speech mode. In D.A. Hamburg (Ed.) Perception and Its Disorders, Proceedings of A.R.N.M.D. Baltimore: Williams and Wilking Co., 1970. Pp. 238-254.

Liberman, A.M., Cooper, F.S., Harris, K.S., and MacNeilage, P.F. A motor theory of speech perception. In C.G.M. Fant (Ed.), Proceedings of the Speech Communication Seminar, Stockholm, 1962. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, 1963.

Liberman, A.M., Cooper, F.S., Harris, K.S., MacNeilage, P.F., and Studdert-Kennedy, M. Some observations on a model for speech perception. In W. Wathen-Dunn (Ed.), Models for the Perception of Speech and Visual Form. Cambridge: M.I.T. Press, 1967.

Liberman, A.M., Cooper, F.S., Shankweiler, D.S., and Studdert-Kennedy, M. Perception of the Speech Code. Psychological Review, 1967, 74, 431-461.

Massaro, D.W. Preperceptual Images Processing Time, and Perceptual Units in Auditory Perception. Psychological Review, 1972, 79, 2, 124-145.

Massaro, D.W. Perceptual units in speech recognition. Journal of Experimental Psychology, 1974, 102, 2, 199-208.

Neisser, U. Cognitive Psychology. New York: Appleton, 1967.

Pisoni, D.B. On the Nature of Categorical Perception of Speech Sounds. Status Report on Speech Research (SR-27), Haskins Laboratories, New Haven, 1971, 101.

Pisoni, D.B. Perceptual processing time for consonants and vowels. Paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., December, 1972.

Pisoni, D.B. Auditory and Phonetic Memory Codes in the Discrimination of Consonants and Vowels. Perception and Psychophysics, 1973, 13, 2, 253-260.

Pisoni, D.B. Dichotic Listening and Processing Phonetic Features. In F. Restle, R.M. Shiffrin, N.J. Castellan, H. Lindman, and D.B. Pisoni (Eds.), Cognitive Theory: Volume I. Potomac, Maryland: Erlbaum Associates (1974, In Press).

Reed, S.K. Psychological Processes in Pattern Recognition. New York: Academic Press, 1973.

Shiffrin, R.M. & Geisler, W.S. Visual Recognition in a Theory of Information Processing. In R. Solso (Ed.), The Loyola Symposium: Contemporary Viewpoints in Cognitive Psychology. Washington: Winston, 1973.

Shiffrin, R.M., Pisoni, D.B. and Castaneda-Mendez, K. Is attention shared between the ears? Cognitive Psychology, 1974, 6, 2, 190-215.

Stevens, K.N. The potential role of property detectors in the perception of consonants. Paper presented at the Symposium on Auditory Analysis and Perception of Speech, Leningrad, USSR, August, 1973.

Stevens, K.N. and Halle, M. Remarks on analysis by synthesis and distinctive features. In Wathen-Dunn, W. (Ed.), Models for the Perception of Speech and Visual Form, Cambridge: M.I.T. Press, 1967.

Stevens, K.N. and House, A.S. Speech Perception. In J. Tobias (Ed.) Foundations of modern auditory theory: Volume II. New York: Academic Press, 1972, 1-62.

Studdert-Kennedy, M. The Perception of Speech. In T.A. Sebeok (Ed.), Current trends in linguistics, Volume XII, The Hague: Mouton, 1974(a).

Studdert-Kennedy, M. Speech Perception. In Lass, N.J. (Ed.), Contemporary Issues in Experimental Phonetics, Springfield, Illinois: C.C. Thomas, 1974(b).

Studdert-Kennedy, M. and Shankweiler, D.P. Hemispheric Specialization for Speech Perception. Journal of the Acoustical Society of America, 1970, 48, 2, 579-594.

Studdert-Kennedy, M., Shankweiler, D., and Pisoni, D.B. Auditory and Phonetic Processes in Speech Perception: Evidence from a Dichotic Study. Cognitive Psychology, 1972, 3, 455-466.

Wood, C.C. Levels of Processing in Speech Perception: Neurophysiological and Information-processing Analyses. Unpublished doctoral dissertation. Yale University, 1973. (Also appears in Status Report on Speech Research, SR-35/36, Haskins Laboratories, New Haven.)