PATTERN RECOGNITION BY HUMANS AND MACHINES

Volume 1

Speech Perception

Edited by

Eileen C. Schwab and Howard C. Nusbaum

ADEMIC PRESS SERIES IN
GNITION AND PERCEPTION

Pattern Recognition by Humans and Machines

Volume 1

Speech Perception

Edited by

EILEEN C. SCHWAB

AT&T Information Systems
Indianapolis, Indiana

HOWARD C. NUSBAUM

Speech Research Laboratory
Indiana University
Bloomington, Indiana

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Berkeley Boston
London Sydney Tokyo Toronto

ACADEMIC PRESS, INC.
Orlando, Florida 32887

ACADEMIC PRESS, INC.
1250 Sixth Avenue, San Diego, California 92101

LIBRARY OF CONGRESS CATALOGING-IN-PUBLICATION DATA

Main entry under title:

Pattern recognition by humans and machines.

(Academic Press series in cognition and perception)
Includes index.
Contents: v. 1. Speech perception — v. 2 Visual. perception.
1. Pattern perception. 2. Pattern recognition systems. I. Schwab, Eileen C. II. Nusbaum, Howard C. III. Series.
BF311.B316 1986 001.53'4 85-20110
ISBN 0–12–631401–2 (v. 1: hardcover) (alk. paper)
ISBN 0–12–631403–9 (v. 1: paperback) (alk. paper)

PRINTED IN THE UNITED STATES OF AMERICA

87 88 89 9 8 7 6 5 4 3 2

Pattern Recognition by Humans and Machines

Volume 1

Speech Perception

This is a volume in

ACADEMIC PRESS
SERIES IN COGNITION AND PERCEPTION

A Series of Monographs and Treatises

Preface

The basic problem of understanding how humans perceive information in the constant bombardment of sensory flux is, without question, one of the most difficult problems facing cognitive science. However, understanding perception is critical to developing a complete theoretical account of human information processing because perception serves to interface the physical world with the mental world. The patterns of energy impinging on sensory receptors must be transformed into representations that are canonical with the structural and dynamic properties of the physical world. At the same time, these representations must be compatible with the cognitive processes that mediate perception, comprehension, and memory. Theories of perception, therefore, need to take into account the physics of light and sound, the structure of objects and language, the neurophysiology of the sensory system, the limitations and capabilities of attention, learning, and memory, and the computational constraints of parallel and serial processing.

Perception will not be understood solely through better or more complete descriptions of the physical stimulus or the neural code that is used to represent the product of sensory transduction. Recognition simply does not occur in the sense organs. Similarly, new algorithms for pattern matching, or theories of memory or attention, or even new basic theories of computation will not, by themselves, solve the problem. Instead, the solution to understanding human perceptual information processing will depend on an interdisciplinary approach that integrates scientific knowledge about cognitive psychology, computation, physics, mathematics, and linguistics.

With this approach in mind, we decided to bring together some of the essential and yet diverse aspects of research on perception. Previous treatments of this subject have tended to consider perception from the

perspective of cognitive psychology alone, or artificial intelligence alone, or some perspective in isolation, but there have been few attempts to integrate work on human perception together with discussions of the basic computational issues surrounding the modeling of perceptual processes. Within the limitation of two volumes, it is impossible to deal with all of the interrelated and interdisciplinary issues that must be considered in the study of perception. Therefore, we chose to focus on several basic problems of pattern recognition in speech perception and visual form perception. Our aim in editing this book, then, was to assemble a set of chapters that would consider perception from the perspectives of cognitive psychology, artificial intelligence, and brain theory.

Certainly at a relatively abstract theoretical level, pattern recognition is, in its essence, quite similar for speech perception and scene perception. There are two theoretically distinguishable parts to the problem of pattern recognition: First, how does the perceptual system segment meaningful forms from the impinging array of sensory input? Second, how are these forms then recognized as linguistic or physical objects? It is our belief that in spite of the apparent differences in the processing and representation of information in speech and vision, there are many similar computational issues that arise across these modalities as well. Some of these cross-modality similarities may be captured in basic questions such as: What are perceptual features, and how are these features organized? What constitutes a perceptual unit, and how are these units segmented and identified? What is the role of attention in perception? How do knowledge and expectation affect the perceptual processing of sensory input? And what is the nature of the mechanisms and representations used in perception? It is this set of fundamental questions that we have chosen to cover in these two volumes.

In considering the theme of perception across the domains of speech and visual form, some questions are quite apparent: Why compare speech perception and visual form perception? Why not compare the perception of spoken and written language, or general audition with vision? The reason is that our intention in editing these volumes was to focus specifically on the perception of meaningful sensory forms in different modalities. Previous books on spoken and written language have emphasized the role of linguistics in perception and thus are concerned with the role of meaning in perception. However, the problems of pattern segmentation and recognition for spoken and written language are not directly comparable; printed words are clearly segmented on the page by orthographic convention, while there are no clear linguistic segmentation boundaries in spoken language. Similarly, with respect to the issues that arise in books comparing the perception of arbitrary patterns in audition and vision, the

emphasis is generally more on the psychophysical problems of transduction, coding, and detection than on the cognitive psychology of segmentation and recognition.

These chapters have been organized into two volumes—one focusing on speech perception and the other focusing on visual form perception. It is important to note that within each volume the theoretical issues touched on by the chapters are all quite distinct, while between volumes there are a number of similarities in the issues that are discussed. In Volume 1, some of the basic theoretical questions in speech perception are considered, including the perception of acoustic–phonetic structure and words, the role of attention in speech perception, and models of word and phoneme perception. In Volume 2, several fundamental questions concerning visual form perception are considered, including the perception of features and patterns, the role of eye movements in pattern processing, and models of segmentation and pattern recognition.

These volumes would certainly not have developed without the cooperation and contribution of the authors. In addition, we are grateful to a number of colleagues for their assistance. We would like to thank David Pisoni and Barry Lively for their active support and encouragement of our work on this project. We would also like to acknowledge Stephen Grossberg for his constant stimulation to bring this project to fruition despite several problems and setbacks. Finally, we conceived of this book while we were graduate students at the State University of New York at Buffalo, and it developed as a direct consequence of numerous discussions and arguments about perception with our colleagues and professors there. We thank Steve Greenspan, Jim Sawusch, Irv Biederman, Jim Pomerantz, Erwin Segal, Naomi Weisstein, and Mary Williams for providing the scientific climate from which this book could develop.

Contents of Volume 2

Visual Perception

Contents

CHAPTER 1

Speech Perception: Research, Theory, and the Principal Issues*

David B. Pisoni and Paul A. Luce

Department of Psychology, Indiana University, Bloomington, Indiana 47405

I. INTRODUCTION

The basic problems in speech perception are, in principle, no different from the basic problems in other areas of perceptual research. They involve a number of issues surrounding the internal representation of the speech signal and the perceptual constancy of this representation—the problem of acoustic–phonetic invariance and the phenomena associated with perceptual contrast to identical stimulation. When viewed from a fairly broad perspective that stresses the commonalities among sensory and perceptual systems, the problems in speech perception are obviously similar to those encountered in vision, hearing, and the tactile system. However, when viewed from a more narrow perspective that emphasizes the differences among sensory and perceptual systems, speech perception immediately becomes more distinctive and unique because of its role in language, thought, and communication among members of the species. Indeed, many researchers have suggested that seemingly unique biological specializations have developed to meet the demands imposed by the use of a vocal communication system such as speech.

The field of speech perception is an unusually diverse area of study involving researchers from a number of disciplines including psychology, linguistics, speech and hearing science, electrical engineering, and artificial intelligence. Despite the diversity of approaches to the study of

* Preparation of the chapter was supported in part by NIH Research Grant NS-12179 and NSF Grant BNS-83-05387 to Indiana University in Bloomington. We thank Beth Greene and Jan Charles-Luce for helpful comments and suggestions.

speech perception, a small number of basic questions can be identified as "core" problems in the field. For the psychologist, the fundamental problem in speech perception is to describe how a listener converts the continuously varying acoustic stimulus produced by a speaker into a sequence of discrete linguistic units, and how the intended message is recovered. This general problem can be approached by examining a number of more specific subquestions. For example, what stages of perceptual analysis intervene between the presentation of a speech signal and eventual understanding of the message? What types of processing operations occur at each of these stages? What specific types of mechanisms are involved in speech perception, and how do these interact in understanding spoken language?

Although the speech signal may often be of poor quality, with much of the speech slurred, distorted by noise, or at times even obliterated, the perceptual process generally appears to proceed quite smoothly. Indeed, to the naive observer, the perceptual process often appears to be carried out almost automatically, with little conscious effort or attention. Listeners are conscious of the words and sentences spoken to them. The speech sounds and the underlying linguistic organization and structure of the linguistic message appear transparent, and a good part of the perception process is normally unavailable for conscious inspection.

One important aspect of speech perception is that many components of the overall process appear to be only partially dependent on properties of the physical stimulus. The speech signal is highly structured and constrained in a number of principled ways. Even large distortions in the signal can be tolerated without significant loss of intelligibility. This appears to be so because the listener has several distinct sources of knowledge available for assigning a perceptual interpretation to the sensory input. As a speaker of a natural language, the listener has available a good deal of knowledge about the constraints on the structure of an utterance even before it is ever produced. On one hand, the listener knows something about the general situational or pragmatic context in which a particular utterance is produced. Knowledge of events, facts, and relations is used by listeners to generate hypotheses and draw inferences from fragmentary or impoverished sensory input. On the other hand, the listener also has available an extensive knowledge of language which includes detailed information about phonology, morphology, syntax, and semantics. This linguistic knowledge provides the principal means for constructing an internal representation of the sensory input and assigning a meaningful interpretation to any utterance produced by a speaker of the language.

In understanding spoken language, we assume that various types of information are computed by the speech processing mechanisms. Some

forms of information are transient, lasting for only a short period of time; others are more durable and interact with other sources of knowledge that the listener has stored in long-term memory. Auditory, phonetic, phonological, lexical, syntactic, and semantic codes represent information that is generally available to a listener. The nature of these perceptual codes and their potential interactions during ongoing speech perception have been two of the major concerns in the field over the last 10–15 years.

In this chapter we review what we see as the principal issues in the field of speech perception. Most of these issues have been discussed in the past by other researchers and continue to occupy a central role in speech perception research; others relate to new problems in the field that will undoubtedly be pursued in the future as the field of speech perception becomes broader in scope. Each of these problems could be elaborated in much greater depth, but we have tried to limit the exposition to highlight the "core" problems in the field.

II. THE PRINCIPAL ISSUES

II.A. Linearity, Lack of Acoustic–Phonetic Invariance, and the Segmentation Problem

As first discussed by Chomsky and Miller (1963), one of the most important and central problems in speech perception derives from the fact that the speech signal fails to meet the conditions of linearity and invariance. The *linearity condition* assumes that for each phoneme there must be a particular stretch of sound in the utterance; if phoneme X is to the left of phoneme Y in the phonemic representation, the stretch of sound associated with X must precede the stretch of sound associated with Y in the physical signal. The *invariance condition* assumes that for each phoneme X there must be a specific set of criterial acoustic attributes or features associated with it in all contexts. These features must be present whenever X or some variant of X occurs, and they must be absent whenever some other phoneme occurs in the representation. As a consequence of failing to satisfy these two conditions, the basic recognition problem can be seen as a substantially more complex task for humans to carry out. Although humans can perform it effortlessly, the recognition of fluent speech by machines has thus far proven to be a nearly intractable problem.

For more than 30 years, it has been extremely difficult to identify acoustic segments and features that uniquely match the perceived phonemes independently of the surrounding context. As a result of coarticulation in speech production, there is typically a great deal of contextual

variability in the acoustic signal correlated with any single phoneme. Often a single acoustic segment contains information about several neighboring linguistic segments (i.e., parallel transmission), and, conversely, the same linguistic segment is often represented acoustically in quite different ways depending on the surrounding phonetic context, the rate of speaking, and the talker (i.e., context-conditioned variation). In addition, the acoustic characteristics of individual speech sounds and words exhibit even greater variability in fluent speech because of the influence of the surrounding context than when speech sounds are produced in isolation.

The context-conditioned variability resulting from coarticulation also presents enormous problems for segmentation of the speech signal into phonemes or even words based only on an analysis of the physical signal, as shown in the spectrograms displayed in Figure 1.1. Because of the failure to meet the linearity and invariance conditions, it has been difficult to segment speech into acoustically defined units that are independent of adjacent segments or free from contextual effects when placed in sentence contexts. That is, it is still extremely difficult to determine strictly by simple physical criteria where one word ends and another begins in fluent speech. Although segmentation is possible according to strictly acoustic criteria (see Fant, 1962), the number of acoustic segments is typically greater than the number of linguistic segments (phonemes) in an utterance. Moreover, no simple invariant mapping has been found between acoustic attributes and perceived phonemes or individual words in sentences.

II.B. Internal Representation of Speech Signals

There has long been agreement among many investigators working on human speech perception that at some stage of perceptual processing, speech is represented internally as a sequence of discrete segments and features (see, e.g., Studdert-Kennedy, 1974, 1976). There has been much less agreement, however, about the exact description of these features. Arguments have been provided for feature systems based on distinctions in the acoustic domain or the articulatory domain, and for systems that combine both types of distinctions (Chomsky & Halle, 1968; Jakobson, Fant, & Halle, 1952; Wickelgren, 1969).

A number of researchers have come to view these traditional feature descriptions of speech sounds with some skepticism, particularly with regard to the role they play in ongoing speech perception (Ganong, 1979; Klatt, 1977, 1979; Parker, 1977). On reexamination, much of the original evidence cited in support of feature-based processing in perceptual experiments seems ambiguous and equally consistent with more parametric