PERCEPTION AND PRODUCTION OF FLUENT SPEECH

Edited by
Ronald A. Cole

PSYCHOLOGY LIBRARY EDITIONS:
COGNITIVE SCIENCE

ROUTLEDGE

PSYCHOLOGY LIBRARY EDITIONS:
COGNITIVE SCIENCE

Volume 4

PERCEPTION AND PRODUCTION OF FLUENT SPEECH

PERCEPTION AND PRODUCTION OF FLUENT SPEECH

Edited by
RONALD A. COLE

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published in 1980 by Lawrence Erlbaum Associates, Inc.

This edition first published in 2017
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-138-19163-1 (Set)
ISBN: 978-1-315-54401-4 (Set) (ebk)
ISBN: 978-1-138-19389-5 (Volume 4) (hbk)
ISBN: 978-1-315-63893-5 (Volume 4) (ebk)

Publisher's Note
The publisher has gone to great lengths to ensure the quality of this reprint but points out that some imperfections in the original copies may be apparent.

Disclaimer
The publisher has made every effort to trace copyright holders and would welcome correspondence from those they have been unable to trace.

PERCEPTION and PRODUCTION of FLUENT SPEECH

Edited by
RONALD A. COLE
Carnegie-Mellon University

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1980 Hillsdale, New Jersey

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

Library of Congress Cataloging in Publication Data

Main entry under title:

Perception and production of fluent speech.

Bibliography: p.
Includes indexes.
1. Speech perception. 2. Speech. I. Cole, Ronald Allen, 1944-
BF455.P388 153.6 78-25481
ISBN 0-89859-019-1

Printed in the United States of America

Contents

Preface

Perception and Production of Fluent Speech looks at the mental processes involved in producing and understanding spoken language. Although there have been several edited volumes on speech in the past ten years, this volume is unique in that it deals exclusively with perception and production of *fluent* speech. A little reflection reveals what a curious state of affairs this is. Why should a volume on speech be unique because it deals with natural continuous speech? The answer is that, in the past 25 years, the study of speech perception has dealt mainly with "low-level" processes—processes involved in recognizing individual speech sounds. These experiments have provided important information about how we categorize and discriminate among speech sounds, but have left unanswered the question of how people communicate using natural continuous speech.

Fortunately, the past few years have seen an increase in research and theory by psychologists, linguists, and computer scientists on perception and production of fluent speech. The chapters in this volume, contributed by distinguished scientists from each of these fields, deal with such questions as: How are ideas encoded into sound? How does a speaker plan an utterance? How are words recognized? What is the role of knowledge in speech perception? In short, how do people communicate with each other using speech?

I have arranged the chapters within the book into four sections. The chapters in Section 1, *The Patterns of Speech,* provide three opinions on the best way to look at speech. When reviewing these chapters, I was reminded of the parable of the three blind men who attempt to discover the nature of an elephant by examining its individual parts. One blind man examines the

elephant's ear, another its leg, and another its trunk, and each arrives at a vastly different conception of the structure and function of an elephant. So it is with speech. In order to examine its structure, it is necessary to *look* at the energy in sound, and any visual transformation of the energy in speech emphasizes some important features and leaves us blind to others. If we look at the energy in speech as it is displayed on an oscilloscope, we appreciate its rhythmic and temporal structure, but are blind to most of its spectral detail. If we look at a speech spectrogram, we see a detailed layout of spectral information, but are blind to its prosodic structure. Thus, the way we choose to look at speech turns out to be critically important in structuring our hypotheses about what information is important to its perception.

In Chapter 1, Cole, Rudnicky, Zue, and Reddy ask the question: How much of the phonemic information in natural continuous speech is displayed on a speech spectrogram? Their answer is that at least 90% of the phonemes in a spoken sentence can be identified from a spectrogram. They report the performance of an expert spectrogram reader who is able to identify between 85 and 90% of the phonemes in an unknown utterance from a spectrogram. Cole et al. conclude that there is a direct and learnable relationship between the visual patterns displayed on a spectrogram and the phonemes of a language, and discuss the implications of their research to theories of speech percpetion, machine recognition of speech, and speech therapy.

In Chapter 2, Scott argues that "the almost exclusive use of the spectrograph for speech analysis has confined our concept of the speech signal to a spatio-temporal display that is quite unlike the signal as it exists in space." As an alternative, Scott offers a Gestalt theory of speech perception that views speech as integrated patterns in time rather than spectral patterns laid out in space, and emphasizes the importance of relations among elements rather than individual spectral cues. The theory receives empirical support in studies that demonstrate the importance of relational cues in vowel perception. The theory has also led to two inventions: a tactile aid to speech reception that conveys relational information about time and frequency (and has been shown to produce immediate and substantial improvement in lip reading) and a speaker-independent pitch meter that follows the intonation contour of a speaker's voice.

In Chapter 3, Searle, Jacobson, and Kimberly provide us with a new look at speech. They argue that, if we want to understand how humans perceive speech, we should be looking at a visual display that models the human ear. That is, the speech scientist should look at the same information that is presented to the auditory cortex. A speech spectrogram does not provide this information. It acts as a constant bandwidth filter system, and therefore provides the same frequency–time resolution at all frequencies. Experiments in auditory psychophysics and physiology demonstrate that the human ear provides excellent frequency resolution at low frequencies, and excellent

temporal resolution at high frequencies. Based on this research, Searle et al. conclude that a visual display of the output of a one-third octave filter system accurately reflects the tradeoff between time and frequency found in the peripheral auditory system. A computer analysis of the parameters of their display produced excellent recognition of the stop consonants of English.

In his discussion of Chapters 1–3, Stevens raises specific problems with each approach, and offers a theoretical framework for considering the three rather disparate views of speech provided in the first three chapters. Because the comments offered by the discussants require a thorough reading of the chapters, I will not attempt to summarize them here. But I urge you to read them carefully—they are uniformly insightful and entertaining.

The chapters in Section 2, *Understanding Spoken Language,* examine the information-processing strategies that humans use to transform sound into meaning. In Chapter 5, Bond and Garnes provide both experimental and observational evidence that various sources of knowledge are used during speech perception. Their experiments show that the interpretation of an ambiguous stimulus—such as a word that can be heard as either "date" or "gate"—is consistent with the semantic structure of the sentence in which it occurs. At the observational level, Garnes and Bond have collected over 900 cases of misperceptions of conversational speech—"slips of the ear." Their sophisticated analysis of these errors provides convincing evidence that listeners interpret conversational speech in terms of their knowledge of the phonological, syntactic, and semantic structure of their language.

In Chapter 6, Cole and Jakimik offer a model of speech perception that describes the information-processing strategies that transform speech into an ordered series of words. The model consists of four assumptions about the way in which sound and knowledge are used to recognize words. The results of a series of experiments in which listeners monitor fluent speech for mispronounced words provide support for each of the assumptions of the model.

Foss, Harwood, and Blank in Chapter 7, argue that two independent codes are activated during word recognition—a phonetic code that is computed directly from the acoustic input, and a more abstract phonological code that is stored with each lexical item in memory and emerges as the item is retrieved. Foss et al. use the dual code hypothesis to explain why listeners are able to detect target phonemes faster in predictable words than unpredictable words, while no such difference exists between frequent and infrequent words. In Chapter 8, Posner and Hanson discuss the three chapters in Section 2 in terms of the similarities and differences between speech perception and reading.

The chapters in Section 3, *Machine-Motivated Models,* present models of speech recognition based on recent advances in computer science in the development of speech understanding systems. In Chapter 9, Reddy describes two of the speech understanding systems developed at Carnegie-Mellon

University—HEARSAY II and Harpy. These systems are able to recognize over 1,000 words in connected speech from a large (but finite) number of sentences. Although the two systems are similar in their performance statistics, they provide very different solutions to the problems involved in recognizing natural continuous speech. HEARSAY consists of independent knowledge sources that interact with each other to arrive at the best interpretation of a sentence. Harpy consists of a single precompiled network of sound sequences, and a sentence is recognized by finding the "best" path through this network. Reddy describes the architecture and design principles of the two systems, and illustrates the operation of each system by working through an example sentence. The chapter concludes with a discussion of the strengths and limitations of the two approaches to automatic speech recognition.

The HEARSAY system is a psychologist's dream. It provides a complete information-processing model of speech perception in which independent sources of knowledge dynamically interact to understand a spoken sentence. The Harpy system, on the other hand, is (at least at first glance) a brute force, engineering solution to the problem of connected speech recognition. It is somewhat surprising then, that both Klatt, in Chapter 10, and Newell, in Chapter 11, present models of speech perception that incorporate the basic features of Harpy, rather than HEARSAY. In Chapter 10, Klatt identifies eight problems that any speech recognition system must overcome in order to recognize words from fluent speech. The solution to these problems is offered in two new computer systems—SCRIBER and LAFS. Taken together, these systems provide a model of lexical access from acoustic data. Although the model is presented as an engineering solution to the problem of word recognition, it is intended to be taken seriously as a model of human speech perception.

In Chapter 11, Newell combines two artificial intelligence models to produce a theory of human speech perception. Newell first describes a production system architecture called HPSA77, "a proposed structure of the architecture within which human cognition occurs." The principles of the Harpy speech understanding system are then incorporated into the production system architecture. The result is PS.Harpy, "a theory of speech perception embedded in a theory of general cognition." After constructing the theory, Newell considers the way in which PS.Harpy handles various phenomena in speech perception, such as the predictive use of knowledge, categorical perception, and phonemic restorations. Newell's chapter is a fine example of *sufficiency analysis,* in which a system that is sufficient to perform a complex congnitive task is evaluated as a model of human cognition.

After Norman's discussion of the models by Klatt and Newell in Chapter 12 (in which he offers timely advice on "copycat science" to cognitive psychologists), we turn to the final section of the book, *Production of Fluent Speech.*

In Chapter 13, Sorensen and Cooper examine the relationship between the syntactic structure of a sentence and its physical realization. They find that reliable changes in fundamental frequency occur at those points in a sentence at which transformational rules have been applied. For example, in the sentence "The seamstress wove your hat and the maid your scarf," the verb "wove" has been deleted from the underlying structure "the maid wove your scarf." Sorensen and Cooper find a reliable change in fundamental frequency between "maid" and "your." This same change is not observed in the control sentence "The seamstress wove your hat and then made your scarf." The experiments provide impressive support for the operation of abstract grammatical operations during speech production.

In Chapter 14, Fay considers the best way to represent the grammar of a language in a model of speech production. His proposal is as intriguing as it is bold. Fay argues for a *direct realization* of transformational rules in models of speech production. He argues that the processing operations that map ideas into words preserve both the substance and the form of transformational rules. After describing the difference between direct and indirect realization models of speech production, Fay offers support for a direct realization model from errors observed in spontaneous speech.

In Chapters 15 and 16, Sternberg, Monsell, Knoll, and Wright report an extensive series of studies of rapid speech, which show how a motor program for an entire utterance, prepared in advance, controls the execution of its elements. In Chapter 15, they report their initial experiments on the latency and duration of rapid utterances, show how the phenomena generalize to typewriting (another domain of motor control), develop a model for the control of rapid movement sequences, and reject various competing explanations. In Chapter 16 they extend their findings in several directions that should be of special interest to students of speech, elaborate their model, defend it against a new contender, and test it further. Although Chapter 15 has been published elsewhere, it was included here because it serves as a necessary prelude to Chapter 16.

The book concludes—most appropriately, I believe—with Herb Simon's discussion of the chapters on speech production. Simon's chapter includes advice on how experimental psychologists can win at twenty questions with nature—a question that has bothered many of us in recent years.

I would like to thank a number of friends for their help in making this book a reality. Karen Locitzer provided invaluable editorial assistance on each chapter in this volume. Without her help, it is doubtful that this book would have appeared during my lifetime. Ed Seiger and Betty Boal typed many of the chapters, and did so extremely well. I am indebted to the staff of the Psychology Department at Carnegie-Mellon University—Janet Mazefsky, Lou Beckstrom, Lois Iannacchione, and Muriel Fleishman—for taking care of the numerous arrangements and details before, during, and after the Carnegie Symposium, from which this book emerged. Most of all, I thank my

wife, Loretta, and my children, David and Debbie, for their love, patience and encouragement while I was editing this book, and for allowing me to smoke large aromatic cigars in the house during this project.

Finally, I thank the participants in the Carnegie Symposium—the authors of this volume—for their excellent contributions. I forgive Al Newell, John Sorensen, and Bill Cooper for causing me to spend a week of my summer vacation reading their chapters rather than sunbathing on the Jersey shore. It was worth it.

RONALD A. COLE

THE PATTERNS OF SPEECH

1 Speech as Patterns on Paper

Ronald A. Cole

Alexander I. Rudnicky
Carnegie-Mellon University

Victor W. Zue
Massachusetts Institute of Technology

D. Raj Reddy
Carnegie-Mellon University

INTRODUCTION

Ever since the invention of the sound spectrograph some thirty years ago (Koenig, Dunn, & Lacey, 1946), the spectrogram has been the single most widely used form of display for speech. The popularity of the spectrogram is at least partly due to the fact that it is relatively easy to produce, and that it provides a visual display of the relevant temporal and spectral characteristics of speech sounds. To be sure, a speech spectrogram sometimes introduces distortions to the acoustic structure of speech and often does not provide adequate information on certain linguistically relevant cues, such as stress and intonation. Nevertheless, a speech spectrogram gives a good description of the segmental acoustic cues of speech, and it has been an invaluable tool in the development of our understanding of speech production and perception.

A speech spectrogram of the utterance, "The boy was there when the sun rose," is shown in Fig. 1.1. As Fig. 1.1 reveals, the speech spectrogram provides a display of the energy in the speech wave in terms of frequency—along the vertical axis; time—along the horizontal axis; and intensity—by the darkness of the markings.

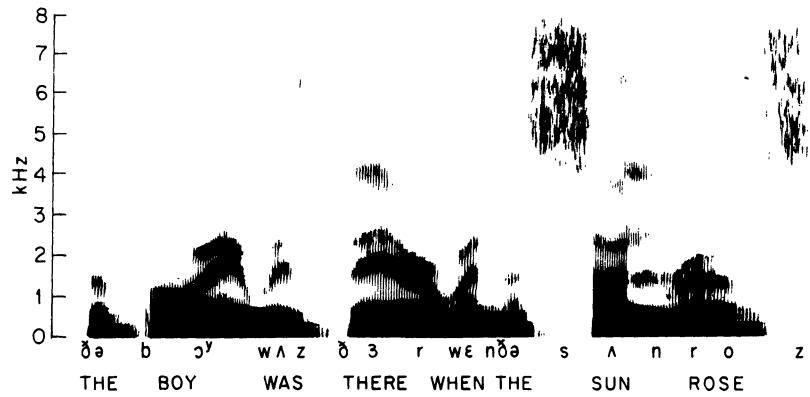

FIG. 1.1. Speech spectrogram of "the boy was there when the sun rose."

Is it possible to read a spectrogram? Can one (who is familiar with such a visual representation) examine a speech spectrogram of an unknown utterance and determine what was said? The common view is that speech spectrograms cannot be read (e.g., Fant, 1962; Liberman, 1970; Lindblom & Svensson, 1973). Fant (1962), for example, states that "I have not met one single researcher who has claimed he could read speech spectrograms fluently, and I am no exception myself [p. 4]."

It is obvious that Fant never met Lev Rubin, an inmate of a prison camp for scientists in Solzhenitsyn's *The First Circle* (1968). We are introduced to Lev Rubin as "the only person in the Soviet Union who can read visible speech." As part of a secret project suggested by Stalin, Rubin learned to read the output of

> A visible speech device—known as VIR—which turns out what is called a 'voice print'.... In these voice prints speech is measured three ways at once: frequency—across the tape; time—along the tape; and amplitude—by the density of the picture. Therefore each sound is depicted so uniquely that it can be recognized easily, and everything that has been said can be read on the tape [pp. 186–187].

A description is even provided of Rubin reading a voiceprint:

> "You see, certain sounds can be deciphered without the least difficulty, the accented or sonorous vowels, for example. In the second word the 'r' sound is distinctly visible twice. In the first word the accented sound of 'e' and in front of it a soft 'v'—for there can't be a hard sound there. Before that is the formant 'a,' but we mustn't forget that in the first, the secondary accented syllable 'o' is also pronounced like 'a'. But the vowel 'oo' or 'u' retains its individuality even when it's far from the accent—right here it has

> the characteristic low-frequency streak. The third sound of the first word is unquestionably 'u.' And after it follows a palatal explosive consonant, most likely 'k'—and so we have 'ukovi' or 'ukavi.' And here is a hard 'v'—it is clearly distinguished from the soft 'v' for it has no streak higher than 2,300 cycles. 'Vukovi'—and then there is a resounding hard stop and at the very end an attenuated vowel, and these together I can interpret as 'dy.' So we get 'vukovidy'—and we have to guess at the first sound, which is smeared. I could take it for an 's' if it weren't that the sense tells me it's a 'z.' And so the first word is—" and Rubin pronounced the word for "voice prints"—"'zvukovidy'" [p. 189].

Solzhenitsyn's report of an expert spectrogram reader has had little impact on the field of speech perception. It is widely believed that it is *not* possible to determine the content of an unknown utterance from a speech spectrogram. One reason for this belief is that research with synthetic speech has led various investigators to conclude that each sound of speech is *not,* in Solshenitsyn's words, "depicted so uniquely that it can be recognized easily." In an article entitled "Why are speech spectrograms hard to read?", Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1968) describe the problem:

> Visible speech is hard to read largely because speech is not a simple alphabet. Speech is, rather, a complex code in the sense that the phonemic message is quite drastically restructured at the level of sound. As a consequence, the acoustic signal corresponding to a particular phoneme is typically different in different phonetic environments. Worse yet, from the standpoint of one who is trying to read spectrograms, definable segments of sound do not correspond to segments at the phoneme level [p. 128]. Moreover, there are, in general, no acoustic segments corresponding to the phoneme segments. As a consequence, looking at a spectrogram does not readily reveal how a stretch of speech might be divided into segments corresponding to phonemes, or even how many phonemes it might contain [p. 130].

It is clearly the case that the consonant and vowel sounds that one hears when listening to speech do not exist as discrete non-overlapping acoustic segments in the speech wave. It is not possible, for example, to insert silent periods between segments of a naturally recorded utterance and still perceive all of the phonemes. But is the relationship between sound and phoneme so complex that it cannot be learned?

On theoretical grounds, Liberman and his colleagues have argued that spectrogram reading cannot be learned. In their view, the speech signal is such a complex code that phonemes can only be perceived through the working of a special decoder. Liberman et al. (1968) argue that:

> Encoded sounds can be efficient vehicles only if there is a decoder—a special device that processes the complex signal so as to recover the string of phonemes. There is such a device in human beings, but, unfortunately for those of us who would send speech through the eye, that decoder can be made to work only from an auditory input. It does not process speech signals (in spectrographic form, for example) that come in through the eye, and it cannot be made to do so by training [p. 128].

In this chapter, we offer an alternative view. We suggest, following Cole and Scott (1974), that there may be sufficient invariant information in the speech wave for humans to be able to perceive phonemes without recourse to a special decoder. Moreover, enough of this information is present in a spectrographic display so that phonemes can often be recognized reliably from a speech spectrogram. The experiment reported in this chapter demonstrates that given sufficient experience, it is possible to identify the phonetic content of an unknown utterance from a speech spectrogram and thereby determine what was said. In addition, protocols provided by the expert spectrogram reader demonstrate that spectrogram reading is a skill that is based on *explicit* knowledge—the identification of specifiable visual features and visual patterns and the application of well-defined rules.

PAST ATTEMPTS AT SPECTROGRAM READING

In order to understand why spectrograms have been considered difficult, if not impossible, to read, we must consider some of the past attempts at learning spectrogram reading and the reasons for their failure.

Potter, Kopp, and Green (1947)

The first attempt at teaching spectrogram reading was reported by Potter et al. in their classic book, *Visible Speech* (1947). The original purpose of the Potter et al. study was to develop a speech understanding aid for the deaf based on the newly constructed Direct Translator, a real-time spectrograph. The Direct Translator was a device based on the original spectrograph except that it produced a continuous dynamic spectrographic display instead of a static display on paper. The dynamic display was produced by having a phosphor belt move past a row of lights activated by a bank of 12 filters into which the speech signal was fed. The device functioned as a 12-channel vocoder, transmitting speech through a telephone-like bandwidth (300–3,000 Hz).

In order to assess the feasibility of such a device as an aid to the deaf,

an evaluation program was undertaken at Bell Laboratories using normal hearing subjects. It was reasoned that if normal hearing subjects could not master the task, there would be little point in attempting to train deaf subjects. Accordingly, an initial group of two women began a training program with the Direct Translator which included the following: speaking into the machine and becoming familiar with their own speech patterns, reading the voice of an instructor, and finally, carrying on conversations mediated by the Translator. The material taught to the subjects consisted of word patterns and short phrase patterns chosen from a basic vocabulary that was oriented towards simple conversation. In addition, the trainees learned phonetic principles and studied the acoustic correlates of phonemes from speech spectrograms produced on paper.

As the study progressed, two more women joined the original group, and these in turn were joined by a congenitally deaf Bell employee who became interested in the project. The experiment with the original group of normal-hearing subjects was terminated after 90 hours of training (30 hours for the late joiners). The deaf subject went on to train for a total of over 200 hours and acquired a vocabulary of about 800 words, a rather impressive achievement. Vocabulary acquisition appeared to be a steady linear function of training; a word took about 15–20 minutes of practice to acquire. In theory, the training procedure could have been extended indefinitely to provide a trainee with as large a vocabulary as needed.

During the experiment, pairs of subjects attempted to communicate with each other without sound by means of the Direct Translator. Subjects were seated in separate sound proof booths from which they could view the Direct Translator. Each booth was equipped with a telephone with a mouthpiece but no receiver. As each subject spoke, his or her speech was displayed in real time on the Direct Translator. Subjects attempted to communicate by reading each other's speech on the Direct Translator.

According to Potter et al.,

> The visible speech class members were able to converse satisfactorily among themselves by talking clearly and at a fairly slow rate. Within the limits of their vocabularies, they were able to carry on conversations with about the same facility as a similarly advanced class in some foreign language. When new and entirely unfamiliar words were displayed on the translator screen, the more experienced students usually were able to read the words after a few repetitions [p. 26].

Unfortunately, a number of factors temper a positive evaluation of the Potter et al. effort. There were no methodological details provided about the training procedures used in the study. It is therefore impossible to

determine whether the linear function of training on word acquisition reflects the participants' ability to learn spectrographic patterns, or reflects the fact that words were presented at a set pace. Another unfortunate omission is any mention of the test procedures used to determine learning. What were the criteria used? How often were items tested and how?

In addition to these problems in the evaluation of the program, there were a number of constraints, some noted by the authors, that limit the generality of the results. For example, the speech to which the trainees were exposed was of a particularly simple kind both in its content and in its form—spoken slowly and clearly. This is very much unlike the speech, rapid and distorted, to which one is exposed in normal conversation (Klatt & Stevens, 1973; Reddy, 1976). Also, as the authors point out, the particular representation chosen (which we described as a 12-channel vocoder) was in all probability not the optimal one, emphasizing irrelevant and hiding relevant features of the speech signal.

Even with all of these potential problems and qualifications, the Potter et al. study represents a substantial achievement. After all, the participants in the study did learn to read speech from a visual display. Moreover, they were able to do this in real time and with vocabularies as large as 800 words. The achievement is even more impressive when it is considered that it was not until a decade later that comprehensive accounts of the acoustics of speech became available (Fant, 1960; Stevens & House, 1961). The Bell Labs project represents a substantial achievement, even by our contemporary standards. It was a demonstration that, in principle, speech could be understood in an other-than-auditory modality.

Despite the potential shown by the Potter et al. study, the Direct Translator did not find much use outside speech therapy, most likely because of its impractical size and large cost. Since the Potter et al. study, there have been no further published attempts to teach real-time spectrogram reading.

Svensson (1974)

As part of a study of prosodic and grammatical influences on speech perception, Svensson (1974) had a group of subjects read spectrograms. Svensson's subjects consisted of workers at the Speech Transmission Laboratory of the Royal Institute of Technology in Stockholm and of students at the University of Stockholm who had participated in a spectrogram reading course. Thus, these subjects had a relatively high degree of sophisitcation in spectrogram interpretation. In the first part of the study, the subjects were all presented with spectrograms of nonsense utterances, consisting of phonologically permissible nonsense words spo-

ken with sentence intonation. The subjects were given written instructions on spectrogram interpretation, which included relevant acoustic data and a decision procedure for segment identification. A total of nine spectrograms were presented, an hour being allotted for reading each one. The results, according to Svensson, were disappointing, with performance ranging from a low of 22% segments identified to a high of 51%. The average level was 38%. In the second part of the study, spectrograms of meaningful utterances were used. Performance on this material was almost identical to that in the first part of the study, despite the fact that the subjects knew that they were dealing with meaningful utterances generated from a restricted grammar and a limited lexicon.

Klatt and Stevens (1973)

In the Klatt and Stevens study, the authors attempted to label phonetically a set of 19 spectrograms of unknown utterances spoken by five unfamiliar talkers. In order to minimize the possibility of recognizing words in the spectrogram, a mask was placed on each spectrogram, allowing only 300 msec of speech to be visible at one time. In addition, the reading was done in a single pass. Despite these constraints, 33% of all segments were correctly transcribed, with a further 40% given a correct partial specification. Given the limited opportunity to scan the spectrogram, this performance is quite good.

The Klatt and Stevens study is unfortunately limited, for our present purposes, in that the readers gave themselves only a single opportunity to scan the spectrogram. It is not clear whether they could have realized a substantial improvement in reading accuracy if they had used a less restricted procedure.

Summary of Past Attempts

Taken together, the results of these studies do not make one overly optimistic about the potential of spectrogram reading. While the Potter et al. study demonstrated that it is possible to learn to read spectrograms of carefully articulated speech in real time, its generality is limited, since the task did not reflect the complexity of conversational speech. On the other hand, the contemporary studies, despite the fact that they could benefit from over twenty years of intensive research into the acoustics of speech, revealed surprisingly low accuracies in attempts to identify phonetic segments from speech spectrograms.

What has been lacking is an approach that combines the better aspects of the previous work: the detailed knowledge gained from contemporary

acoustic phonetics and the systematic training used in the Potter et al. study. The expert spectrogram reader (whom we shall call VZ) described in this chapter has managed to combine these aspects successfully.

THE EXPERIMENT

The purpose of the experiment was to arrive at a better understanding of the process of spectrogram reading by examining in detail the methods used by VZ. While it was evident from casual observation that VZ was a highly skilled spectrogram reader, we did not have a clear idea of the extent of this skill. Thus, one of the goals of the study was to form a quantitative description of VZ's performance, both in terms of the accuracy of his transcriptions and in terms of the type of material he is able to interpret.

In addition to establishing the level of VZ's skill, we were also interested in learning something about the methods VZ uses to interpret spectrograms. For example, it has been suggested that high-level sources of knowledge, such as syntax and semantics, play an important role in speech understanding (Lindblom & Svensson, 1973; Miller & Isard, 1963). To what extent does VZ make use of such knowledge in interpreting the phonetic content of a spectrogram? Finally, a detailed examination of VZ's methods was undertaken to provide information that could be of use in the design of speech understanding systems.

The Subject

The expert reader VZ began his systematic study of spectrograms in 1971. At that time, he was taking part in the Advanced Research Projects Agency (ARPA) speech understanding project at Massachusetts Institute of Technology Lincoln Laboratories while completing his graduate studies at MIT. As an initial attempt to learn about the acoustic cues of speech sounds in continuous speech, VZ began to study the material in *Visible Speech*. He soon decided that this approach was inadequate for the study of real speech, since, among other things, the carefully articulated material in *Visible Speech* lacked many of the characteristics of continuous speech. He then began to collect his own data by preparing spectrograms from recordings made for him by various talkers, using such materials as nonsense consonant–vowel (CV) utterances and the Harvard Phonetically Balanced List of Sentences (Egan, 1944). During this time, he concentrated on trying to identify the relationships between the segmental features of individual speech sounds and their acoustic correlates. After two years of attention to segmental features, VZ became interested

in the transformations that an individual speech sound undergoes when it appears in the context of other sounds, and began to systematically study the acoustic–phonetic and phonological rules of American English. Since that time, VZ has made extensive use of spectrograms in his research and has maintained a consistent interest in spectrogram reading. Between 1971 and the present, VZ has devoted, on the average, between one-half and a full hour per day on spectrogram reading. He estimates that over the years he has spent between 2,000 and 2,500 hours reading spectrograms. It is this extensive amount of practice that enables him to perform at his present level of skill.

General Procedure

All formal spectrogram reading sessions took place at the Department of Psychology at Carnegie-Mellon University on two separate occasions, the first in October, 1977, the second in February, 1978. On both occasions, VZ came to Carnegie-Mellon for a period of two days, during which he participated in four experimental sessions. During each of these sessions, which lasted from two to three hours, VZ read the equivalent of about four to six spectrograms of full utterances. In order to preserve all details of the sessions for later analysis, all sessions were recorded on videotape.[1]

The spectrogram reading sessions were organized as follows. VZ sat at a table, together with a prompter who handed him spectrograms for reading. These spectrograms were chosen randomly from one of the sets available to the prompter (the spectrogram sets are described later in this text). Thus, the prompter, except under a few circumstances, was not aware of the correct identity of a spectrogram being read by VZ. The spectrogram was placed in a work area in view of the video camera. VZ then proceeded to interpret the spectrogram in his customary manner. VZ was asked to make his notations directly on the spectrogram in order to preserve a permanent record of his transcription.

Since one of the main purposes of this study was to analyze the methods used by VZ, he was instructed to verbalize the decisions he was making and to describe those features of the spectrogram he was attending to at any one time. The main task of the prompter was to ask VZ to elaborate interesting or unclear points. For example, if VZ said "This looks like an [l]," the prompter would say, "Why does that look like an

[1]A short film (15 min) based on the October, 1977 visit has been made and is available from any of the authors, or from the Computer Science Library at Carnegie-Mellon University. The film shows how a typical spectrogram is read, and describes in some detail the procedures used by VZ.

[l]?". If VZ replied "Because of the relationship of the second and third formants," the prompter would say, "What is the relationship?". By the end of our experiment, VZ was an extremely verbal and informative spectrogram reader.

Selection and Preparation of Spectrograms

Several sets of spectrograms were prepared for this study. The largest set consisted of normal English utterances; the other sets consisted of altered utterances, described in detail later in this text. One of our concerns in producing the spectrograms was the fact that VZ was accustomed to working with spectrograms made with a "Voice-Print" (Model 4691A) spectrograph, while the machine used to generate the spectrograms at C-MU was a "Kay Sona-Graph" (Model 6061B). While both spectrographs produce basically similar displays, a number of differences exist, most notably an expanded frequency display in the Voice-Print (a 7 kHz frequency range, compared to 8 kHz for the Kay Sona-Graph) and a low-frequency attenuation feature of the Voice-Print. In order to determine that VZ's performance was not due to the exact form of the representation, we asked William Cooper, then at MIT, to prepare a set of spectrograms of normal English sentences using the Voice-Print with which VZ was most familiar. All remaining spectrograms used in the experiment were produced on the Kay Sona-Graph at Carnegie-Mellon University.

Examples of spectrograms produced on each machine are shown in Fig. 1.2. These spectrograms show the utterances, "The soldiers knew the battle was won," produced on the Sona-Graph, and "Yesterday Bill saw the Goodyear blimp," produced on the Voice-Print. The transcriptions produced by VZ, including segmentation marks, are shown directly below each spectrogram. (For presentation purposes, we had the transcription copied professionally.) The transcription of the utterance "The soldiers knew the battle was won" represents VZ's first attempt to read a spectrogram made on a Kay Sona-Graph, from an utterance produced by an unfamiliar speaker. The transcription was accurate enough to enable a linguist to identify the sentence without hesitation.

The talker for the Carnegie-Mellon spectrograms was one of the authors, RC. The talker for the MIT spectrograms was John Sorensen. VZ had had no previous opportunity to read spectrograms made from utterances spoken by these talkers.

Table 1.1 displays the 23 normal and anomalous sentences that were presented to VZ on speech spectrograms during the experiment. Fifteen of the spectrograms—11 from the C-MU set and four from the MIT set—displayed normal English sentences. In addition to the normal

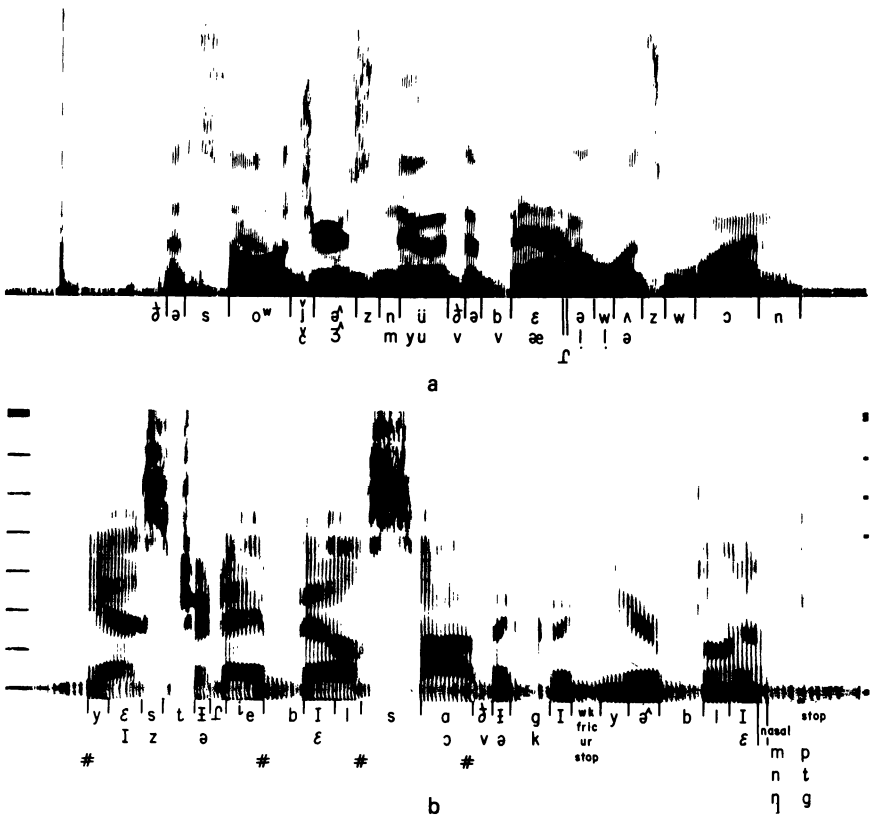

FIG. 1.2. Spectrograms: (a) "The soldiers knew the battle was won," produced on a Kay Sona-Graph. (b) "Yesterday Bill saw the Goodyear blimp," produced on a Voiceprint.

sentences, eight spectrograms were presented in an effort to determine what sources of knowledge VZ used to read spectrograms. If VZ is in the habit of using higher-level information to guide the interpretation of phonetic information, then giving him a spectrogram of some familiar utterance, such as a proverb, should make this apparent by producing an increase in the accuracy of his performance relative to less familiar utterances. In addition, we tried to assess the salience of the nonphonetic information by deliberately introducing mispronunciations into proverbs. If nonphonetic information carries a great deal of weight for VZ, then the mispronunciations should go undetected.

A set of phonetically anomalous sentences was used to see whether VZ made use of lexical information, i.e., whether he was able to detect word patterns in the spectrograms. Phonetically anomalous sentences were produced by including nonsense words along with normal words in En-

TABLE 1.1
Utterances Read by VZ from Spectograms

English Sentences
C-MU Set
1. The soldiers knew the battle was won.
2. Smoking is bad for your mind and body.
3. Basketball is fun to play.
4. Folk dancing makes me dizzy.
5. The cross-eyed seamstress couldn't mend straight.
6. Winning is never a pretty thing for him.
7. The baby gave its mother a kiss.
8. A shark may be dangerous when hurt.
9. Go paddle your own canoe.
10. After the sixth story, Alice fell asleep.
11. Bill knew his aardvark was smaller.
MIT Set
12. The plants in the office needed more light.
13. I left my house at nine o'clock.
14. Yesterday Bill saw the Goodyear blimp.
15. Tim read the novel last week.
Proverbs
16. Waste not want not.
17. An apple a *t*ay keeps the *b*octor away.
Phonetically Anomalous Utterances
18. What teelings are we day with?
19. Give mine yadiya of his hate raret.
20. Our young people leh nose today.
21. Good tar hand zaim to come by.
Syntactically Anomalous Utterance
22. Bears shoot work on the country.
Nonsense Utterance
23. Wake jungles gasoline sudden bright.

glish sentences. This manipulation should make it impossible to use word-level knowledge to interpret the phonetic content.

There were at least two other potential sources of knowledge available to VZ: syntactic constraints and semantic content. Knowledge of syntactic structure can be used to predict the class of an unknown word and thus

TABLE 1.2
Words Used as Stimuli in the Carrier Phrase "Say _____again"

baby	hatch	elephant	slow	yoyo
tribe	paper	breath	snowshoe	whorehouse
feather	toothache	clover	pleasure	move
smooth	rocket	spinner	stitches	drum
zebra	swinger	thread	garage	fuzzy
shallow	florist	this	wagon	thing
fresh	tower	dog	soap	vest
change	criminal	bladder	letter	school
jade	knife	glasses	cages	prize

facilitate its identification. Similarly, once several words of an utterance are known it becomes possible to guess the remainder on the basis of the semantic context, the "meaning" of the utterance. Two utterances taken from Miller and Isard (1963), one semantically anomalous and one both syntactically and semantically anomalous, were presented to VZ in order to assess the role of syntactic and semantic knowledge during spectrogram reading.

In addition to the spectrograms of normal and altered sentences, spectrograms of 45 words in the carrier phrase "Say _____ again" were prepared. The purpose of this set was to evaluate VZ's ability to interpret a phone string (i.e., a word) in connected speech when the word boundaries were known. The words in this set were balanced for phonetic content by having all consonants appear in all permissible positions within words: initially, medially, and finally. Table 1.2 shows the words included in this set.

The results are presented in two parts. The first part examines the subject's ability to segment and label speech spectrograms, and the second part examines the nature of the segmentation and labeling process.

PART I: PERFORMANCE

How Performance Was Measured

VZ's ability to label phonetic segments in spectrograms was measured against phonetic transcriptions produced by three phoneticians. Each phonetician (a) had received formal training in transcription phonetics; (b) was currently teaching (or had recently taught) a course in phonetic transcription or English phonology; and (c) used phonetic transcription as part of his or her research.

Transcribers were provided with a high-fidelity cassette recording (produced on a Dolby system) of the 23 utterances and the 45 words in the

carrier phrase "Say _____ again." Transcribers were warned that utterances might contain nonsense words or mispronunciations.

Each transcriber was provided with a copy of the ARPAbet symbol system shown in Table 1.3. As the table reveals, the ARPAbet contains a set of phonetic symbols, a corresponding set of orthographic symbols, and an example of a word containing the sound corresponding to each symbol. Transcribers were instructed to use only the phonetic symbols listed in the ARPAbet in their transcription. We explained that in order to minimize scoring problems based on differences in phonetic notation, it was essential that all transcribers use the same symbol system.

The ARPAbet system provides a broad, *phonetic* transcription that is nearly phonemic. For example, the same symbol—/t/—is used to indicate aspirated [t^h] (in a stressed CV syllable), unaspirated [t] (in an /st/ cluster) and unreleased (word-final) [$t^\neg$]. For the purposes of this study, the loss of phonetic detail was more than compensated for by the convenience of having a standardized transcription system.

TABLE 1.3
Phonetic Symbols[a] Used by VZ and the Three Transcribers

Phoneme	ARPAbet	Example	Phoneme	ARPAbet	Example
/i/	IY	b*ea*t	/ŋ/	NX	si*ng*
/ɪ/	IH	b*i*t	/p/	P	*p*et
/e/ (e^y)	EY	b*ai*t	/t/	T	*t*en
/ɛ/	EH	b*e*t	/k/	K	*k*it
/æ/	AE	b*a*t	/b/	B	*b*et
/ɑ/	AA	B*o*b	/d/	D	*d*ebt
/ʌ/	AH	b*u*t	/g/	G	*g*et
/ə/	AO	b*ough*t	/h/	HH	*h*at
/o/ (o^w)	OW	b*oa*t	/f/	F	*f*at
/u/	UH	b*oo*k	/ɵ/	TH	*th*ing
/u/	UW	b*oo*t	/s/	S	*s*at
/ə/	AX	*a*bout	/š/	SH	*sh*ut
/ɨ/	IX	rose*s*	/v/	V	*v*at
/ɜ/	ER	b*ir*d	/ð/	DH	*th*at
/ɚ/	AXR	butt*er*	/z/	Z	*z*oo
/ɑ^w/	AW	d*ow*n	/ž/	ZH	a*z*ure
/ɑ^y/	AY	b*uy*	/č/	CH	*ch*urch
/ɔ^y/	OY	b*oy*	/ǰ/	JH	*j*udge
/y/	Y	*y*ou	/ʍ/	WH	*wh*ich
/w/	W	*w*it	/l̩/	EL	batt*le*
/ɾ/	R	*r*ent	/m̩/	EM	bott*om*
/l/	L	*l*et	/n̩/	EN	butt*on*
/m/	M	*m*et	/ɾ̃/	DX	ba*tt*er
/n/	N	*n*et	/ʔ/	Q	(glottal stop)

[a]From Bolt, Beranek and Newman, Inc. Report No. 3438, Vol. II, p. 72.

Definition of a Segment

Fig. 1.3 displays the phonetic transcription produced by each transcriber for the utterance "The soldiers knew the battle was won," and the transcription produced by VZ while reading this spectrogram (shown in Fig. 1.1). As this figure suggests, transcribers almost never disagreed on the number of segments in an utterance. Moreover, there was unanimous agreement on 85% of all segment labels, so there was little difficulty in aligning the three transcriptions for each utterance.

We adopted a majority-vote criterion for the definition of a segment:

A segment was assumed to exist when two transcribers produced a segment label—even if they did not produce the same label.

The utterance depicted in Fig. 1.3, for example, contains 22 segments according to such a criterion.

In the 23 utterances shown in Table 1.1, there were only four cases where one transcriber produced a segment label and the other two did not. Two of these involved insertion of [ə] by a single transcriber, once before [r] and once before [n] (the other two transcribers indicated syllabic [n̩]). In a third case, one transcriber produced [ur], while the other two produced [ɚ]. In the fourth case, one transcriber inserted an [m] in "apple", producing "ample." In these four cases, we assumed that no segment existed.

In an additional six cases, two transcribers produced a segment label while the third did not; in these six cases a segment was assumed to exist. Finally, there were 493 cases where all three transcribers produced a segment label. There was thus a total of 499 phonetic segments, defined by the agreement of two or more transcribers on the existence of a phonetic segment. VZ's ability to segment and label speech spectrograms was based on these 499 segments.

By the same token, the 45 words shown in Table 1.2 contain 201 seg-

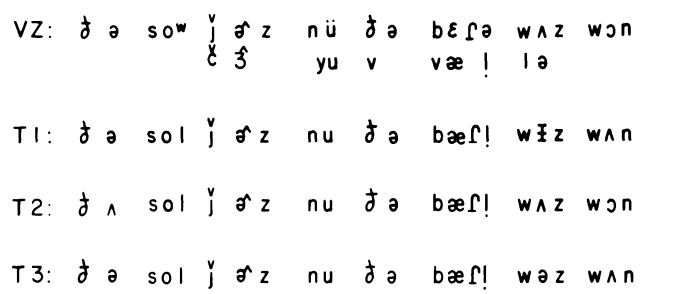

FIG. 1.3. Transcriptions produced by VZ from a speech spectrogram, and by three transcribers (Ts) who listened to the speech.

ments. There was only one case where all three transcribers did not provide a segment label: Two of them decided that the word "criminal" contained a final reduced vowel followed by [l], while the third indicated a syllabic [l̩].

Definition of a Label

As a working hypothesis, speech scientists typically assume that speech is composed of an ordered sequence of phonetic segments. But it is important to remember that a phonetic transcription is an *interpretation* imposed upon a continuously varying acoustic signal. In a typical utterance, the identity of one or more phonetic segments may be ambiguous due to imprecise articulation or masking caused by environmental noise (although the latter presumably was not a factor in the present experiment). In such cases, listeners are known to generate a phonetic interpretation of the input that is consistent with their expectations about what the speaker is likely to be saying (Garnes & Bond, 1977; Miller, 1956; Schubert & Parker, 1955; Warren, 1970).

The point of this discussion is that, because the acoustic signal is sometimes ambiguous and therefore open to interpretation, there is no single "correct" phonetic transcription of a spoken utterance. As we shall see, even under the most ideal conditions (unlimited time, high-fidelity recording, broad transcription, native language, familiar speaker) transcribers disagreed on about 15% of all segment labels. Therefore, in order to produce a fair picture of VZ's labeling ability, we used three different scoring measures.

Measure 1: VZ and All Ts. The first measure (M1) asks: When all three transcribers agree on the same segment label, how often does VZ produce the same label? M1 examines all cases where, according to the transcribers (Ts), segments are perceptually unambiguous. This measure has the advantage that we can be fairly confident about the identity of each segment, since all Ts provided the same label. The measure will inflate performance somewhat, since it examines only those segments that are perceptually unambiguous, and these segments may also be acoustically less ambiguous than the remaining segments.

For purposes of analysis, we divided segments into three broad classes: consonants, vowels, and a third class consisting of the liquids [r] and [l], the semivowels [w] and [y], the syllabic consonants [l̩] and [n̩], and the retroflexed vowels [ɝ] and [ɚ]. For convenience, we will call this class "others." There were no instances where the transcribers did not agree on the class of a segment, according to our rather broad classification.

According to the transcribers, the 499 segments consisted of 245 consonants, 171 vowels, and 83 others. In order to examine how well the

transcribers agreed on the specific segment labels in each class, we counted the number of instances in which all three transcribers produced the same segment label. For consonants, all Ts agreed on 235 of the 245 segments, or 96% agreement. It is interesting that six of the 10 cases where agreement was not unanimous involved a word-final stop. For vowels, the three transcribers agreed on 121 of the 171 segments, or 71% agreement (specific vowel disagreements are described below). Finally, for liquids, semivowels, and syllabic consonants, transcribers agreed on 68 of 83 segments, or 81% agreement. To summarize, the transcribers agreed unanimously on 85% (424 out of 499) of the segment labels. They were much more likely to agree on consonant labels than vowel labels, although a few acoustically (and perceptually) similar vowels accounted for most of the vowel disagreements.

Measure 2: VZ and Any T. M2 asks: How often does VZ produce the same segment label as *any* transcriber? This measure considers all 499 segments. At first glance, this measure may seem liberal, since VZ is given credit for a correct label if he agrees with *any* transcriber. However, the 75 segments on which the Ts disagree are probably acoustically more ambiguous than the segments on which agreement is unanimous. We favor this measure because it captures the variability inherent in phonetic transcription.

Measure 3: VZ and Each T. The third measure asks: How well does VZ agree with the transcription produced by each transcriber? This is the most severe measure, because VZ is penalized for each disagreement with each transcriber. For this analysis we computed, for each utterance, the proportion of segments on which VZ agreed with each transcriber, and then calculated the average agreement.

Performance on Utterances

Segmentation

In this subsection we examine VZ's ability to parse a speech spectrogram into units corresponding to phonetic segments. VZ indicates segments in two ways: (a) by the placement of "segment markers" directly under the spectrogram; and (b) by the placement of segment labels side by side between two markers. Such cases almost always involved sonorant–vowel or vowel–sonorant combinations. For example, the phrase "for your" was represented by VZ as [fˌɚ˞yɚ˞], with three labels sharing the same segment markers.

VZ produced both optional segments (indicated by longer segment markers or by an annulus around a segment label) and alternate segmenta-

tions. When scoring the data, credit was given for correct optional segments and correct alternate segmentations. We felt that in these cases, VZ provided information that could be used by a perceiver (or a computer) during the recognition process, and that no penalty was deserved. Finally, in measuring segmentation, we were concerned only with VZ's ability to identify the existence of phonetic segments; accurate segmentation did not necessarily correspond to accurate labeling.

VZ identified 485 of the 499 segments defined by the transcribers. Thus, slightly more than 97% of all segments were identified from the speech spectrogram. The 14 missed segments were: [d] in "land," "read," and "couldn't," [t] in "plants," [n] in "want," [ð] in "the" (twice), [h] in "his" (twice), [l] in "Bill" and "soldiers," [ə] in "gasoline," and both [y]s in "yadiya." Examination of the speech spectrograms revealed that the visual cues for these segments were either very weak or completely absent. For example, in Fig. 1.2, we can find no visual cues for the [l] in "soldiers," whereas all transcribers indicated the presence of this segment. We expect that for some of these segments, the transcribers' perception may have been influenced by their use of context. For example, the nonsense word "leh" in the phrase "leh nose today", was transcribed as "land" by all Ts. Similarly, "gasoline" was probably produced as "gas'line," since no formants were visible for the vowel [ə], although all Ts agreed on its existence.

VZ produced 20 alternate segmentations. An alternate segmentation was written below the original segmentation and was sometimes produced after the initial segmentation, during the labeling process, upon closer examination of the spectrogram. In 16 of the 20 cases, the alternate segmentation indicated two segments where a single segment had been originally proposed. In one case, three segments were proposed instead of two, and in three cases, a single segment was postulated where two were originally proposed. When an alternate segmentation was proposed, it was correct 13 of 20 times.

Labeling

VZ used a single label to identify a segment 52% of the time (254 out of 485 cases), two labels 35% of the time, and three labels 6% of the time (30 cases). On the remaining 33 segments, VZ produced 17 partial transcriptions, 13 optional segments, and provided no label in three cases. If we exclude optional labels from consideration, and count each partial transcription as three labels, then VZ produced an average of 1.53 labels to each segment. When more than one label was given, they were almost always rank ordered so that it was possible to score VZ's performance for first, second, and third choices. We arbitrarily decided to score all partial transcriptions as equivalent to a third choice.

TABLE 1.4
Agreement among VZ and all Ts on Segment Labels

	Consonants	*Vowels*	*Other*	*Total*
All Ts	235	121	68	424
VZ				
1st choice	165	78	48	291
2nd choice	26	18	10	54
3rd choice and partial	19		2	21
VZ/Ts	210/235	96/121	60/68	366/424
Percent agreement	89	79	88	86

Measure 1. Table 1.4 reveals the number of cases where VZ's segment label agreed with the label produced by all three transcribers. It can be seen that 210 of 235 (89%) consonants were correctly labeled, 96 of 121 (79%) vowels were correctly labeled, and 60 of 68 (88%) others were

TABLE 1.5
VZ's Agreement with all Ts for Each Consonant Segment in Word-Initial, Medial, and Final Position

	Initial		*Medial*		*Final*	
	Ts	*VZ*	*Ts*	*VZ*	*Ts*	*VZ*
/b/	13	12	2	1	0	0
/d/	4	4	8	5	5	4
/g/	6	5	1	1	0	0
/p/	5	5	3	2	2	2
/t/	7	6	10	8	14	12
/k/	7	6	6	6	5	4
/m/	9	7	4	4	4	4
/n/	9	9	10	9	10	8
/y/	0	0	2	2	5	5
/r/	0	0	2	2	0	0
/ɵ/	1	1	0	0	1	1
/ð/	11	8	1	0	0	0
/f/	5	5	3	3	0	0
/v/	0	0	3	3	3	3
/s/	1	1	9	8	11	10
/z/	1	1	1	1	11	10
/š/	2	2	0	0	0	0
/ǰ/	1	1	2	2	0	0
/h/	7	5	0	0	0	0
Total	97	89	67	57	71	63
Percent Correct	92%		85%		89%	

correctly labeled. Across the three categories, about 80% of all agreements occurred on the first (or only) choice, while most of the remaining agreement involved a second choice. For consonants, 17 of the 19 third-choice agreements were partial transcriptions. All but one partial transcription indicated an unreleased stop, a weak fricative (or weak voiced fricative), or both. Only one partial transcription (out of 20) was incorrect. To summarize, when transcribers provide the same label, VZ produced the same label, usually as a first chioce, 86% of the time. Agreement was better on consonants and others (88.5%) than on vowels (79%).

Table 1.5 displays the number of times that all transcribers produced each consonant label in word-initial, word-medial, and word-final position, and the number of times that VZ produced the same label, on any choice, as the three transcribers. Note that agreement between VZ and the three transcribers was slightly better in word-initial and word-final position (92% and 89% respectively) than in word-medial position (85%). It should be remembered, however, that agreement for the word-final stops (22 out of 26) includes credit for the partial transcription "unreleased stop" which was used by VZ most of the time.

Table 1.6 displays, for individual vowel segments and for [r], [l], [w],

TABLE 1.6
VZ's Agreement with All Ts for Each Vowel Segment and for /r/, /l/, /w/, /y/, /l̩/, /n̩/, /ɚ/, and /ɝ/ in Word-Initial, Medial, and Final Position

Vowels			*Other*						
				Initial		*Medial*		*Final*	
	Ts	*VZ*		*Ts*	*VZ*	*Ts*	*VZ*	*Ts*	*VZ*
/i/	16	13	/ɹ/	2	1	13	12	5	4
/ɪ/	15	10	/l/	2	2	10	9	4	3
/e/	15	14	/w/	12	11	1	1	0	0
/ɛ/	9	6	/y/	5	4	2	2	0	0
/æ/	8	6	/l̩/	0	0	1	1	5	4
/ə/	15	13	/n̩/	0	0	1	1	0	0
/ʌ/	6	4	/ɚ/	0	0	1	1	4	4
/ɑʸ/	9	9							
/ɑʷ/	2	2							
/ɑ/	13	11							
/ /	1	1							
/o/	6	5							
/u/	4	2							
/u/	2	0							
Total	121	96		21	18	29	27	18	15
Percent Correct	79%			86%		93%		83%	

TABLE 1.7
All Cases where VZ Disagreed, on any Choice, with the Labels Produced by All Three Transcribers

Consonants			*Vowels*			*Others*		
All Ts	*VZ*[a]	*Word*	*All Ts*	*VZ*[a]	*Word*	*All Ts*	*VZ*[a]	*Word*
/b/	/n/	*B*ill	/i/	/ɪ/	dancing	/r/	/ɚ/	goodyear
/d/	missed	couldn't	/i/	/ɪ/	keeps	/r/	/w, l/	read
/d/	missed	read	/ɪ/	/ɨ/	be	/r/	/ɜ, ɚ/	bears
/d/	/ɾ/	bo*d*y	/ɪ/	/i/	Tim	/l/	missed	so*l*diers
/d/	wk. vcd. fric.	su*dd*en	/ɪ/	/ə, ɚ/	pretty	/l/	missed	Bi*ll*
/d/	/k, g/	aar*d*vark	/i/	/ü/	give	/w/	/f, b/	*wh*en
/g/	/y/	*g*ive	/i/	/ə/	Alice	/y/	missed	*y*adiya
/p/	/b/	kee*p*s	/e/	/ɑʸ/	play	/l/	/l, m/	app*l*e
/t/	/s/	lef*t*	/ɛ/	/ʌ, ə/	fell			
/t/	/s/	*t*eelings	/ɛ/	/ɪ/	bears			
/t/	/m, n/	Baske*t*ball	/æ/	/ɛ, ɪᵊ/	*a*fter			
/t/	/k/	wan*t*	/æ/	/ɛᵊ/	*A*lice			
/t/	missed	Plan*t*s	/ə/	/ɪ, ʌ/	*a*			
/k/	/g/	wor*k*	/ə/	/ɚ/	moth*er*			
/n/	missed	wa*n*t	/ʌ/	/ɛ/	j*u*ngles			
/n/	missed	i*n*	/ʌ/	/ɑ, ɑʷ/	m*o*ther			
/n/	/m/	fu*n*	/ɑ/	/ɔ/	o'clock			
/ð/	missed	*th*e	/ɑ/	/ɔ/	n*o*vel			
/ð/	missed	*th*e	/o/	missed	gas*o*line			
/ð/	/f, ɵ/	*th*e	/u/	/ɛ, o/	kn*ew*			
/ð/	/ɾ/	mo*th*er	/u/	/æ, ɛ/	can*oe*			
/s/	/z/	it*s*	/u/	/ɪ, ɨ, ɛ/	c*ou*ldn't			
/s/	/z/	ba*s*ketball	/u/	/ɪ/	g*oo*dyear			
/z/	/s/	i*s*						
/h/	missed	*h*is						
/h/	missed	*h*is						

[a] When VZ made more than one choice, all are listed: /first, second, . . . /.

[y], [l̩], [n̩], [ɝ], and [ɚ], the number of times the three transcribers produced the same segment label, and the number of times that VZ agreed with the label produced by the transcribers. Since vowels occurred in medial position a vast majority of the time, data are combined for vowels in word-initial, word-medial, and word-final position.

To summarize, Tables 1.5 and 1.6 present, *for each segment*, (a) the number of times the three transcribers produced the same segment label, and (b) the number of times that VZ, on any choice, agreed with the transcribers. Table 1.7 presents all cases where VZ *disagreed*, on any choice, with the label produced by the three transcribers and displays the word or nonsense word in which the segment occurred. Taken together, these three tables present an exhaustive description of VZ's performance for cases in which the segment label is unambiguous.

Measure 2. Segment labels produced by VZ agreed with at least one transcriber on 424 of 499 segments, or 85% agreement. Table 1.8 summarizes VZ's agreement with at least one transcriber for segments in each sound class. As before, there was more agreement on consonants (87%) and others (86%) than vowels (81%).

Table 1.9 displays all cases where the three Ts disagreed on a vowel label, the word in which the vowel occurred, and the label(s) produced by VZ. This table reveals that half of all disagreements among the Ts involved [i]–[ɪ]–[ɨ], with the remainder involving mainly confusions among [ɛ]–[æ], [ʌ]–[ə], and [ɑ]–[ɔ]–[o]. On cases where the three transcribers did not agree, VZ agreed with at least one transcriber over 85% of the time.

There were only 10 cases in which the three transcribers disagreed on a consonant label, and these are shown in Table 1.10. Five of the ten disagreements occurred on a word-final stop consonant, and two occurred on a word-final fricative. On the remaining three disagreements, two transcribers indicated a medial /d/ in "paddle," "needed," and "yester-

TABLE 1.8
Agreement among VZ and any Ts on Segment Labels

	Consonants	*Vowels*	*Other*	*Total*
All Ts	245	171	83	499
VZ				
1st choice	169	109	57	335
2nd choice	26	27	11	64
3rd choice	19	3	4	26
VZ/Ts	214/245	139/171	72/83	425/499
Percent agreement	87	81	86	85

TABLE 1.9
Labels Provided by VZ and Each T for All Cases in Which the Three Transcribers Disagreed on a Vowel Label

Word	*VZ*[a]	*T1*	*T2*	*T3*
pretty	/e, i/	/i/	/i/	/ɪ/
thing	/ɪ/	/i/	/i/	/ɪ/
teelings	/ɪ/	/i/	/i/	/ɪ/
yadiya	/i/	/i/	/y/	/ɪ/
the	/i/	/i/	/ə/	/ɪ/
dangerous	/e, i/	/æ/	/e/	/e/
dancing	/ɛ, æ/	/æ/	/æ/	/e/
apple	/æ, ɛ, ɪ/	/æ/	/æ/	/ɛ/
paddle	/æ, ɛ^{ə}, ə/	/æ/	/æ/	/ɛ/
plants	/ɛ, æ/	/æ/	/æ/	/e/
is	/ɪ, i/	/ɨ/	/ɪ/	/ɪ/
basketball	/ɨ, ə/	/ɨ/	/ɪ/	/ɪ/
its	/ə, ɨ/	/ɨ/	/ɪ/	/ɪ/
in	/ə/	/ɨ/	/ɪ/	/ɪ/
his	/ʌ/	/ɨ/	/ɪ/	/ɪ/
body	/i, e/	/i/	/ɪ/	/ɪ/
winning	/ɪ/	/ɪ/	/ɪ/	/ɪ/
baby	/ɪ, i/	/i/	/ɪ/	/ɪ/
Goodyear	/ə/	/i/	/ɪ/	/ɪ/
office	/ʌ/	/ɨ/	/ɪ/	/i/
is	/ʌ, ə/	/ɨ/	/ɪ/	/i/
seamstress	/ɨ/	/ɨ/	/ə/	/ɨ/
needed	/ɪ, ɨ/	/ɪ/	/ə/	/ɨ/
dangerous	/ʌ/	/ɪ/	/ə/	/ɨ/
a	/ə, ɨ/	/ə/	/ʌ/	/ə/
at	/ə, ɨ/	/ɨ/	/ə/	/ə/
was	/ə, ʌ/	/ə/	/ʌ/	/ɨ/
today	/ɨ, ə/	/u/	/ə/	/ə/
the	/ə/	/ɪ/	/ə/	/ə/
the	/ə/	/ə/	/ʌ/	/ə/
the	/ə/	/ə/	/ʌ/	/ə/
the	/ə, ʌ/	/ɨ/	/ə/	/ə/
was	/ʌ, ə/	/ɨ/	/ʌ/	/ə/
sudden	/ɑ/	/ɑ/	/ʌ/	/ʌ/
of	/ə, ɔ, ʌ/	/ə/	/ʌ/	/ʌ/
won	/ɔ/	/ʌ/	/ɔ/	/ʌ/
office	/ɔ/	/ɑ/	/ɑ/	/ɔ/
smaller	/ɔ/	/ɔ/	/ɑ/	/ɔ/
saw	/ɑ, ɔ/	/ɑ/	/ɑ/	/ɔ/
cross	/ɑ/	/ɔ/	/ɑ/	/ɔ/
on	/ɛ, æ, ɑ/	/ɔ/	/ɑ/	/ɑ/
want	/ɑ, ɑ^{ɣ}, æ/	/ɔ/	/ɑ/	/ɑ/
yadiya	/ɛ, ɑ/	/ɑ/	/ə/	/ɑ/
what	/ɛ, ɪ, ɑ/	/ɑ/	/ɔ/	/ɑ/
more	/ɔ/	/ɑ/	/ɑ/	/ɔ/
story	/ɔ/	/ɑ/	/o/	/ɔ/
folk	[+back, −high]	/o/	/ɔ/	/o/
to	/ə, ɨ/	/u/	/u/	/ə/
your	/u/	/ɚ/	/ʊ/	/ɔ/

[a]When VZ made more than one choice, all are listed: /first, second, . . . /.

TABLE 1.10
Labels Produced by VZ and Each T for All Cases in Which the Three Transcribers Disagreed on a Consonant Label

Utterance	*Word(s)*	*T1*	*T2*	*T3*	*VZ*
4	fol*k* dancing	/k/	/ʔ/	/k/	no label
8	shar*k* may	/k/	/k/	/p/	/k/
9	pa*dd*le	/ɾ/	/d/	/d/	/ɾ/
12	nee*d*ed	/ɾ/	/d/	/d/	/ɾ, v, ð/
14	yester*d*ay	/ɾ/	/d/	/d/	/ɾ/
18	wi*th*	/f/	/ə/	/ə/	/ə/
20	lan*d* nose	no label	/d/	/d/	no label
21	han*d* zaim	/t/	/d/	/d/	/n, m/
22	bear*s* shoot	/ž/	/z/	no label	no label
23	wa*k*e jungles	/k/	/t/	/k/	/k/

day," while the third transcriber indicated a flap. It is interesting to note that VZ indicated a flap in each case. It is also of interest to note that five of the ten disagreements occurred in the anomalous utterances.

Measure 3. The average agreement between VZ and each transcriber for all segments was 81%. The average agreement among the three transcribers for all segments was 90%.

Use of Higher-Level Knowledge

All of the evidence suggests that labeling was performed without the use of syntactic or semantic knowledge. Labelling performance was *better* on the four utterances consisting of nonsense words interspersed with normal words. For the three measures just considered (agreement with all Ts, with any T, and with individual Ts), VZ averaged 93%, 92%, and 88%, respectively, on utterances containing nonsense words. This compares to 86%, 85%, and 81% agreement for the three measures on the entire set of utterances. VZ was therefore slightly more accurate labeling segments in nonsense syllables.

A number of tests were originally designed to determine the extent to which higher-order contextual information was used during labeling. Observation of the labeling process soon revealed, however, that syntactic and semantic information was rarely used during labeling (although certain common words, such as "the," were probably recognized on sight). The labeling process was typically not left-to-right, and labels were not consistently placed first at beginnings of words. Moreover, when labeling was completed on a particular utterance and we asked VZ to identify the sentence, it was obvious that even in utterances where all segments were correctly identified, VZ had not yet identified the words.

But Were They Read?

So far, we have only considered the identification of phonetic segments. The more interesting question is whether VZ's transcriptions are sufficiently accurate to determine what was actually said. To answer this question, we presented a linguist with the 15 transcriptions produced by VZ from spectrograms of the normal English utterances. If we exclude three confusions of "a" and "the," 10 of the transcriptions were read perfectly. Of these 10, four were read from left to right without hesitation. The remaining six were interpreted by first identifying individual words, and then "solving" the sentence like a puzzle. Of the five utterances in which all words were not identified, four involved an error on a single word: "Ella's" for "Alice," "leave" for "left," "square" for "folk" (a guess, which followed identification of "dancing") and "lack" for "want" (another guess, in "waste not, want not"). In the remaining utterance, the linguist identified "New knowledge aardvark was smaller" from the transcription of "Bill knew his aardvark was smaller." The linguist was extremely clever at interpreting the sentences from VZ's transcriptions and performed slightly better than VZ did when attempting to identify the utterances he had transcribed. Altogether, the linguist identified 92% of the words.

Performance on Words in a Carrier Phrase

Segmentation

The 45 words in carrier phrases contained 201 segments. VZ identified all 201 segments. Moreover, VZ did not propose any optional segments or alternate segmentations. VZ was apparently quite confident in his segmentation of words in the carrier phrase, a confidence justified by perfect performance.

Labeling

Measure 1. The 201 segments consisted of 102 consonants, 64 vowels, and 35 others. The three Ts produced the same segment label on 187 of the 201 segments. All but one of the 14 disagreements (the final [l] in "criminal") occurred on vowels, and eight of the 13 vowels involved confusion of [ɪ] or [ɨ], either with each other or with other vowels. To summarize, all Ts produced the same segment label on all 102 consonants, on 51 of 64 vowels, and 34 of 35 others.

For the 187 cases in which all Ts produced the same segment label, VZ produced the same label, on any choice, on 173 segments, or 92.5%. VZ agreed with all Ts on 99 of 102 consonants (97%), 42 of 51 vowels (82%),

and 32 of 34 others (94%). The specific disagreements are shown in Table 1.11.

Measure 2. Of the 14 cases where the three transcribers did not produce the same segment label, VZ agreed with at least one of them on all 14 segments. By this measure, the number of VZ's disagreements stays the same (14), but the total number of segments increases (from 187 to 201), so the proportion of agreements increases slightly, to 93% of all segments. Agreement was again 97% on consonants (99 of 102), rose to 86% for vowels (55 of 64), and stayed at 94% for others (33 of 35).

Why does VZ identify phonetic segments more accurately when a word is in a carrier phrase, rather than an unknown utterance? The major advantage provided by the carrier phrase was that it defined the beginning of the unknown word. VZ was able to use his knowledge of English phonotactics (permissible phoneme sequences) to identify segments. In natural continuous speech, virtually any sequence of segments can occur at a word boundary. Since VZ did not attempt to label spectrograms of unknown sentences word by word (or even left to right), word boundary information, and therefore phonotactic knowledge, was typically not used to identify segments from spectrograms of unknown sentences. The use of

TABLE 1.11
All Cases Where the Label Produced by VZ Disagreed with the Label Produced by all Three Ts, for Word in a Carrier Phrase

	Ts	*VZ*[a]	*Word*
Consonants	/ð/	/g/	smoo*th*
	/b/	/f/	*f*resh
	/ǰ/	/č/	ca*g*es
Vowels	/e/	/i/	c*a*ges
	/e/	/ʌ/	fr*e*sh
	/æ/	/i, ɪ, e/	w*a*gon
	/æ/	/ɪ, ɛ/	h*a*tch
	/æ/	/oʷ/	sh*a*llow
	/ʌ/	/ɑ/	dr*u*m
	/ɑ/	/ɛ, æ/	r*o*cket
	/o/	/w, l/	shall*ow*
	/u/	/ʌ/	sm*oo*th
Other	/l/	/l/	sha*ll*ow
	/r/	/ɚ/	who*r*ehouse

[a] When VZ made more than one choice, all are listed: /first, second, . . . /.

phonotactic knowledge for words in a known carrier phrase probably accounts for the better performance on these words.

Summary

To summarize, VZ correcly identified the existence of 485 of 499, or 97%, of all segments from speech spectrograms of normal and anomalous sentences. Depending upon the scoring method used, VZ agreed with a panel of phoneticians who listened to the sentences on between 81% and 86% of the segment labels. Performance on words in a known carrier phrase was substantially better. VZ identified the existence of all 201 segments identified by the panel of phoneticians, and agreed with the phoneticians on 93% of the segment labels.

PART II: PROCESS

The Segmentation Process

The initial step taken by VZ in reading a spectrogram was to segment the continuous speech wave into units that corresponded roughly to phones. Segmentation is often the necessary prerequisite for labeling, although in some cases a partial hypothesis of segment identity will aid the segmentation process. In this section, we will discuss the criteria that VZ used to locate segment boundaries and the strategies used for segmentation.

Boundary Placement Criteria

Conceptually, the criteria for boundary placement are quite simple, and VZ appeared to make use of only a few simple principles, as shown in the following protocol excerpt:

> I am marking at various places
> where it shows, you know,
> maximal spectral difference...
> I'm basically using the spectral change
> as a parameter for marking the boundaries...
> There is an intensity,
> a sharp intensity difference....

Spectral changes accompany changes in manner of articulation. Each phone has a characteristic acoustic form that is a function of the manner in which it is articulated. A succession of phones will produce successive changes in the form of the speech wave. VZ places boundaries at these points of change.

Spectral Discontinuities. The most striking change in the speech wave occurs in the transition between sonorants (vowels, nasals, and liquids) and obstruents (stops, fricatives, and affricates). Sonorants are often characterized by the presence of low-frequency energy, formant structure,and glottal striations. In contrast, obstruents usually have an aperiodic structure, and little or no energy in the low frequencies. Because of these differences, a boundary between a sonorant and an obstruent is usually easy to detect.

Within the sonorant category, there is a major acoustic difference between nasal and nonnasal segments similar in its distinctiveness to the difference between sonorants and obstruents. A transition from a nonnasal to a nasal is marked by a sharp amplitude drop and an abrupt change in the formant structure, while transitions between nonnasal sonorants are usually marked by smooth formant movements. Again, this usually allows a boundary to be easily detected.

Note, however, that a spectral discontinuity in itself does not constitute a segment boundary, since abrupt spectral changes can occur within single phonetic segments. For example, when a prestressed syllable-initial plosive follows a vowel, as in "the *c*ake," spectral discontinuities occur at the onset of the closure interval, the onset of the stop burst, and the onset of voicing of the following vowel. The discontinuity at the burst onset is (correctly) ignored and the closure and release are considered to be part of a single segment—[k]. Thus, a sharp discontinuity in the spectrograms is not *by itself* a sufficient cue to segmentation; the cue must be interpreted in light of acoustic–phonetic knowledge.

Together, sonorant/obstruent and nasal/nonnasal boundaries account for over 75% of all boundary types. Thus, 75% of all segment boundaries can be easily detected and are accurately marked by VZ. The remaining 25% of the boundaries involve transitions between acoustically similar segments—for example, between vowels and liquids, or between two nasals or two stops—and consequently are more difficult to detect.

Duration. Some portions of the speech wave can be segmented on the basis of duration cues. For example, two adjacent stops, as in "Fol*k* *d*ancing . . ." (sentence 4) can be identified as such by noting the duration of the closure interval between the two words. When compared to other (single stop) closure durations in the utterance, it is unusually long. Similarly, two adjacent nasals can be identified by the presence of an uncharacteristically long nasal segment. Duration also serves to indicate the presence of adjacent sonorant segments, although additional information is needed for accurate boundary placement.

Formant Movements. Boundary placement within sonorant sequences is difficult, since there are no discrete cues such as spectral

discontinuity to guide interpretation. Nevertheless, there is sufficient information in such cases to allow fairly accurate segmentation. Intervocalic glides and liquids can be identified by the dip they induce in the first formant frequency. Some examples of this phenomenon are shown in Fig. 1.4a. More generally, the presence of multiple sonorant segments can be identified by nonmonotonic formant movement within a sonorant stretch (excluding, of course, the transition movements that occur at boundaries between obstruent and nasal segments).

An additional cue is provided for glides in the drop in formant amplitude due to the close articulation of these sounds. Figure 1.4b shows a good example of this cue.

When dealing with adjacent sonorants, VZ usually found it easier not to place boundary markers at all, and instead marked off an aggregate segment. Quite often VZ would not place boundaries between liquids and

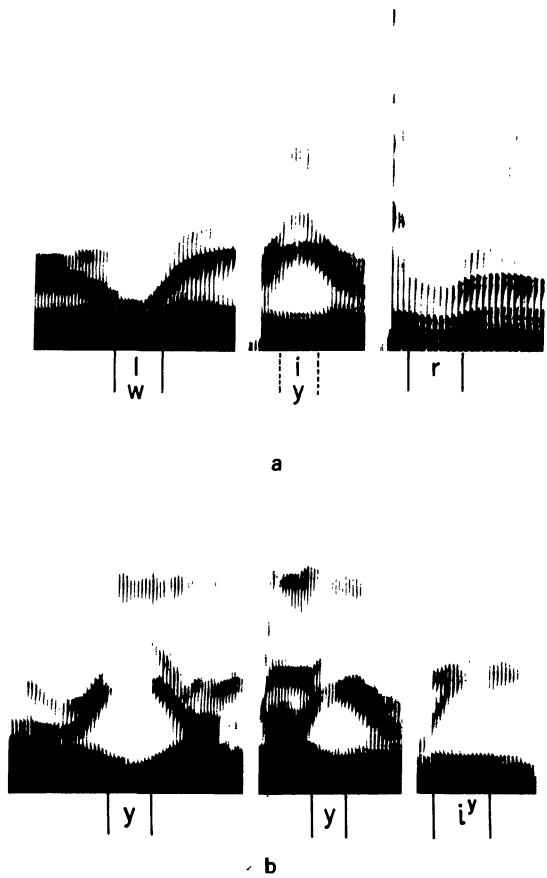

FIG. 1.4. Cues for segmentation. (a) First formant movement within sonorant stretches. (b) Drop in formant amplitude.

vowels and would consider the resulting segment as a single unit, although eventually two labels were placed between the segment markers. In the case of liquid–vowel sequences, this is probably the preferred solution, since liquids have a marked influence on adjacent segments and the two segments are difficult to consider in isolation.

Boundary Placement Strategies

We can define two basic segmentation strategies, left-to-right and nonsequential. VZ is apparently able to use either strategy. Some utterances were segmented in a single left-to-right pass, while others were segmented in an apparently random order. Some of the variation in segmentation can apparently be attributed to the task demands: At the beginning of the experiment, VZ confined himself to a strict left-to-right strategy, apparently believing that this was expected of him. As the session progressed, however, VZ began to use a nonsequential strategy, which seemed more natural to him.

The nonsequential strategy is not random. VZ typically marks the most distinctive boundaries first, and then proceeds to deal with more difficult boundaries:

> Um, I'm segmenting again
> where I consider sonorant stretches are...
> and then... I'll try to break it down
> primarily using spectral changes....

Table 1.12 shows the mean rank order of VZ's boundary placements for representative boundary types, calculated from a corpus containing 244 boundaries (17 of which were considered unclassifiable). The rank order of placement for a boundary is predicted quite well by the visual clarity of the spectral discontinuity at that boundary, as described previously. Thus, it seems that the order in which boundaries were marked probably reflects their actual discriminability for VZ.

Optional Segmentation

Under ideal circumstances, placing a boundary is a straightforward procedure: The appropriate acoustic cues are identified and a boundary is marked. Under actual circumstances, however, various factors will conspire to eliminate the information necessary to detect a boundary.

There are two sources of difficulty in segmentation: the limitations of the spectrographic representation and the nature of speech production. The spectrograph obscures information because of its limited dynamic range and occasionally poor frequency resolution. The information content of the spectrogram is also degraded by processes intrinsic to the

TABLE 1.12
Mean Rank Order of Boundary Placement for All Boundary Types of Frequency Greater Than 1 [a]

Boundary Type	*Mean Rank*	*Category Frequency*
fricative–vowel	7.8	39
stop–glide	8.3	3
stop–vowel	8.5	82
stop–liquid	9.4	5
nasal–stop	9.8	10
nasal–fricative	10.2	6
nasal–vowel	13.3	32
fricative–stop	13.5	13
glide–vowel	15.4	11
fricative–fricative	16	3
liquid–vowel	18.6	7
nasal–nasal	19.5	4
stop–stop	20.3	6
vowel–vowel	21	4

[a] Segment sequence is not considered separately, i.e., a stop–vowel boundary and a vowel–stop are considered to be the same for the purpose of this analysis.

nature of speech production. A good example of this is the drop in pitch and amplitude that normally occurs at the end of an utterance. This makes the detection of utterance-final segments difficult, or it can produce spurious cues (e.g., an amplitude drop could be natural, or it may be due to an utterance-final nasal segment).

Unambiguous interpretation of the spectrographic trace may also be difficult because of speaker characteristics. For example, a high-pitched voice will produce a choppy formant pattern that mimics certain boundary cues and makes boundary placement difficult.

In cases where insufficient information was available to unambiguously establish boundaries, optional segments were sometimes proposed. If we examine the identity of the optional segments proposed by VZ, we find that, with a few exceptions, they are either utterance-initial stops and weak fricatives or postvocalic liquids. Spectrographically, these segments are difficult to identify, either because of their weak energy, (e.g., [ð]) or because they produce only subtle changes in the signal (e.g., the [l] in "soldiers" in Fig. 1.2).

Optional segments were also postulated when duration alone was a potential cue to segmentation. In the utterance "Bears shoot...", the boundary between the first two words consisted of a long fricative seg-

ment which was initially marked as a single segment. Subsequently, this decision was reconsidered: "It is quite long ... it could in fact be two segments." Since duration, in most instances, is only a partial cue, segmentation must remain optional.

Summary

The first step in interpreting a spectrogram was to segment the speech wave into units corresponding to phones. This process was seen to be relatively straightforward, once relevant acoustic dimensions were identified: (a) spectral discontinuities; (b) duration; and (c) formant movement. An analysis of the segmentation strategy revealed that segmentation was essentially context-free and could be performed in a serial left-to-right manner. More commonly, easily distinguished boundaries were marked first, then the more ambiguous ones. Factors such as deficiencies in the spectrographic representation and the nature of speech production introduced difficulties, but despite these, segmentation was carried out with a high degree of accuracy; over 97% of all segment boundaries were identified.

The Labeling Process

One of the main goals of this study was to describe the nature of the methods that VZ used to identify the phonetic content of an utterance. To achieve this goal, we analyzed in great detail a set of twelve protocols chosen from those recorded during the first (October 1977) session. Table 1.13 shows the distribution of segments in this corpus by representative category. The categories were chosen to reflect similarities in the way their members were dealt with by VZ during labeling. Note that the categories represent sets of acoustically (or rather, visually) similar phones.

TABLE 1.13
Distributions of Segments in Corpus Used for Analysis

Segment	*Number*
Stops	60
Strong fricatives	30
Weak fricatives	11
Nasals	32
Liquids	14
Glides	10
Back vowels	12
Front vowels	43
Central vowels	17
Reduced vowels	19

Labeling Sequence

The first question we asked was whether it was possible to identify an order in which phone categories were labeled, perhaps similar to the order based on acoustic distinctiveness found for segmentation. The sequence in which labels were assigned was tabulated for each utterance and the mean rank for each category was calculated. Unlike the results obtained from the analysis of segmentation, no clear-cut pattern emerged. Closer examination of the results, however, suggested that there is some tendency for easily identifiable segments (such as strong fricatives) to be labeled first. Ease of identification depends on such factors as the acoustic distinctiveness of a pattern, its freedom from contextual influences and its clarity of realization. In this exchange, VZ elaborated a part of his strategy:

[AR: Why do you move around?
How do you select the spots you move to?]
ah, for example,
I'm going to ignore this one
because in order to make that decision...
I have to make a few decisions
before I can label it a vowel...
I'm trying to do the segmental labeling
independent of... phonetic context...
ah, then I try to do the other places...

Segment Identification

In this section, we consider in greater detail the methods that VZ used to label individual phones. These methods are of particular interest as they can provide useful insights into human perception and can serve as a guide for improving automatic speech-understanding systems.

For any one segment, VZ would verbalize only a small portion of the information he was using to come to a decision (despite our prompting). Since this meant that the information used to identify a particular phone was present only in a fragmented form, the labeling information was analyzed in two stages. First, VZ's remarks about each individual segment were recorded and summarized. Second, all remarks made about a given phone in the entire corpus were collected together. This allowed us to specify both the core of the procedure and also the variations induced by particular contexts.

The analysis of the protocols revealed that VZ approached the labeling task in one of three ways:

1. By far the greatest number of labeling decisions were based on the identification of *unique spectral patterns* characteristic of individual phones.

2. Pattern detection is augmented by an extensive *knowledge of coarticulatory effects* that distort spectral patterns.
3. In addition, VZ is able to make use of the *constraints imposed by English phonology and phonotactics* to narrow down possible interpretations.

Note that none of these procedures make use of the types of information thought necessary for speech perception. For example, VZ labeled individual segments without reference to the syntactic structure of the utterance or to its semantic content. The error analysis supports this interpretation of VZ's behavior.

Acoustic Patterns. Acoustic patterns, as we define them, consist of easily identified spectral configurations that are unique to a particular phone. Nasals provide one such pattern. Because of the manner in which nasals are articulated, a marked change occurs between a nasal and, say, an adjoining nonnasal sonorant. The regular formant structure is replaced by a steady-state pattern composed of several nasal formants; the overall amplitude drops markedly from adjoining segments. Information about place of articulation is usually available from adjacent sonorant segments or from other sources of knowledge (for example, phonological constraints, as discussed below). Once learned, the basic nasal pattern is almost always recognized correctly. The only difficulties that arise involve unusual circumstances, such as a very rapid speech rate or deficiencies in the representation. The identifying characteristics for a number of phones are listed in Table 1.14.

Apart from vowels, the most common segments in our corpus and in spoken language (Carterette & Jones, 1974; Fletcher, 1953) are the stop consonants. Stop consonants have been perhaps the most thoroughly studied consonants in speech perception, and this research has played an important role in theoretical approaches to speech perception (Cole & Scott, 1974; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Stevens & Blumstein, 1977). It is therefore of interest to examine the methods used by VZ to detect and label stop consonant segments.

A stop is easily detected by the presence of a closure interval. A somewhat more difficult problem is to specify the place of articulation and the voicing value for a stop. Previous work (see Liberman et al., 1967) has indicated that this may be a formidable task, because of the diversity of the acoustic realizations of stops. However, an analysis of VZ's protocols indicates that, at least in the case of spectrogram reading, the discriminations can be made with a high degree of accuracy. The reason for this is that it is possible to define a unique and distinctive pattern for each stop.

Bilabial stops have a characteristic short rising formant pattern which

TABLE 1.14
Some Descriptors Used by VZ in Reference to Phones

Vowels	height	varies inversely with F1
	frontness	varies directly with F1–F2 spread
		duration: /ɪ/ shorter than /i/
		offglides: /ɪ^{ə}/, /æ^{ə}/
	/ɑ/	F1 highest of all vowels
	reduced	short duration
		neutral formant pattern
	diphthongs	spreading formants (e.g., /ɑ^{ɣ}/)
		lowering F1 (e.g., /ɑ^{ʷ}/)
Strong fricatives	voicing	duration (voiced are shorter)
	/s/	aperiodic energy > 4kHz
	/š/	aperiodic energy < 4kHz
Nasals		energy below 300 Hz
		abrupt amplitude onset
		lower amplitude than vowels
		nasalization of adjacent vowels
Place of articulation		labial: all formants move down to closure
		velar: F2 and F3 merge at point of closure
Retroflex sounds		F3 dips below 2 kHz
		/ɚ/: F3 follows F2
		/ɹ/: F3 touches F2
Flaps		short duration < 20, 25 msec
Stops	closure	lack of energy
	burst	labial: little or none
		alveolar: high frequency
		velar: strong, occasionally double bursts
	voicing	voice-onset time (VOT) duration (longer for voiceless)
	transitions	labial: point down
		alveolar: F2 locus at 1800 Hz
		velar: F2 and F3 merge

is markedly different from the patterns observed for alveolar and velar stops. Bilabial stops also have weak release bursts, in contrast to the alveolar and velar stops which both have strong bursts with identifiable energy concentrations. Alveolar stops always have burst frequencies above 3.0 kHz (except when followed by liquids, which tend to lower the burst frequencies) and can be distinguished from velar bursts on the basis of frequency (which, in the case of the velars, is a function of the following vowel). The form of the burst also serves to distinguish the two

stops—the velar burst is usually longer, more intense, and is sometimes doubled. Additional information is obtained from the associated transitions, either from their implied "loci" or from special characteristics. (For example, front and central vowels will show a distinctive joining of the second and third formants at the point of velar articulation.)

If we examine labeling errors for stops, we find that most of them involve incomplete specifications due to missing acoustic information. Thus, for word-final stops produced without a release burst, the place judgment is unreliable and in many cases was not attempted. The remaining errors show that stops were misidentified because of missing major class cues (e.g., the stop closure) or because of inadequacies in the spectrographic representation.

VZ's labeling of vowels presents an interesting case, since proportionately the largest number of segment label errors were due to vowel misidentifications. The possible source of these errors will be considered in a later section. At present, we would like to examine the cues used by VZ to classify vowels.

The easiest distinction was between reduced and unreduced vowels. Reduced vowels are characterized by their short length, often as short as two glottal pulses, which sets them off from all other vowels in an utterance. Once a reduced vowel was identified, the high variant ([ɨ]) was distinguished from the low one ([ə)] by comparing the distance between the first and second formants and the distance between the second and third (F2 and F3 are closer for [ɨ]). Often, VZ did not distinguish between the two, as indeed it is unnecessary to do in natural speech.

To identify the remaining vowels, a variety of cues was used. Surprisingly, VZ rarely took advantage of his ability to directly measure formant frequencies with a template, and appeared to work directly from the formant patterns. In describing vowels, VZ often appeared to make use of the Jakobson, Fant, and Halle (1963) features of compact–diffuse and grave–acute. (This is not surprising, as the Jakobson et al. system was derived mainly from acoustic characteristics.) Front vowels were distinguished by their diffuseness (essentially the separation between F1 and F2), with different degrees of diffuseness indicating vowel height. Within the front series, finer discriminations were made on the basis of other cues. For example, duration was used to distinguish between [i] and [ɪ], which have similar formant patterns ([ɪ] is usually shorter). Offglides were also used to distinguish vowels. Thus, [e] will have a pronounced [y] offglide, in contrast to [ɪ] and [ɛ], while [æ] will often exhibit a schwa offglide, being realized as [æ^{ə}]. Similar statements can be made about the remaining vowels and indeed about all other speech sounds. That is, all segments can be classified into general categories, and then distinctive cues can be used to identify the phones within each category.

As the examples discussed thus far show, patterns can be composed of

either steady-state distributions of energy, as in the case of the fricatives, or of dynamic patterns, such as formant movements. The pattern for a given phone is not always confined to a single segment, but quite often extends to adjoining segments, most notably in the case of stops. That each phone has associated with it a characteristic set of acoustic features is not a novel proposal (see, for example, Fant, 1968). VZ's achievement is in having developed the ability to recognize the characteristic pattern for each phone, as it occurs under a large variety of conditions. Extensive exposure to spectrogram representations has allowed VZ to develop the appropriate prototypes for each English phone.

Phonetic Context. While many segments can be readily identified on the basis of their acoustic characteristics, there are cases in which coarticulatory effects disguise the identity of a segment. This is less of a problem with consonants, which, as we have seen, have essentially invariant cues, than with vowels, which tend to be highly influenced by surrounding segments. This is quite evident from the error scores for the two classes: Vowels were mislabeled more often than consonants. Figure 1.5 shows some examples of highly coarticulated vowels. In most cases, these are short vowels, surrounded by consonant segments that have very different places of articulation. In such cases, VZ is usually able to make a fairly accurate guess. The basis for these identifications is not always clear, but it appears that VZ is able to *compensate* for the coarticulation by computing appropriate formant displacements, arriving at a "noiseless" vowel.

Knowledge of coarticulation is an essential part of VZ's skill, as the following excerpt shows:

> Given that it's a /w/
> rather than an /l/
> Compensation of the second formant
> probably is not as much

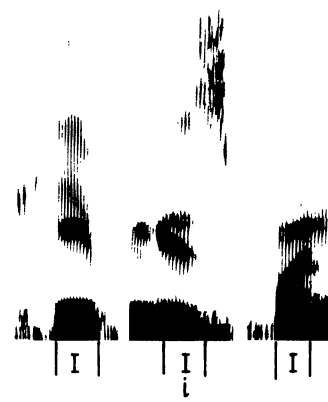

FIG. 1.5. Differences in formant structure of /ɪ/ in different phonetic environments.