MODULARITY
and
THE MOTOR THEORY OF SPEECH PERCEPTION

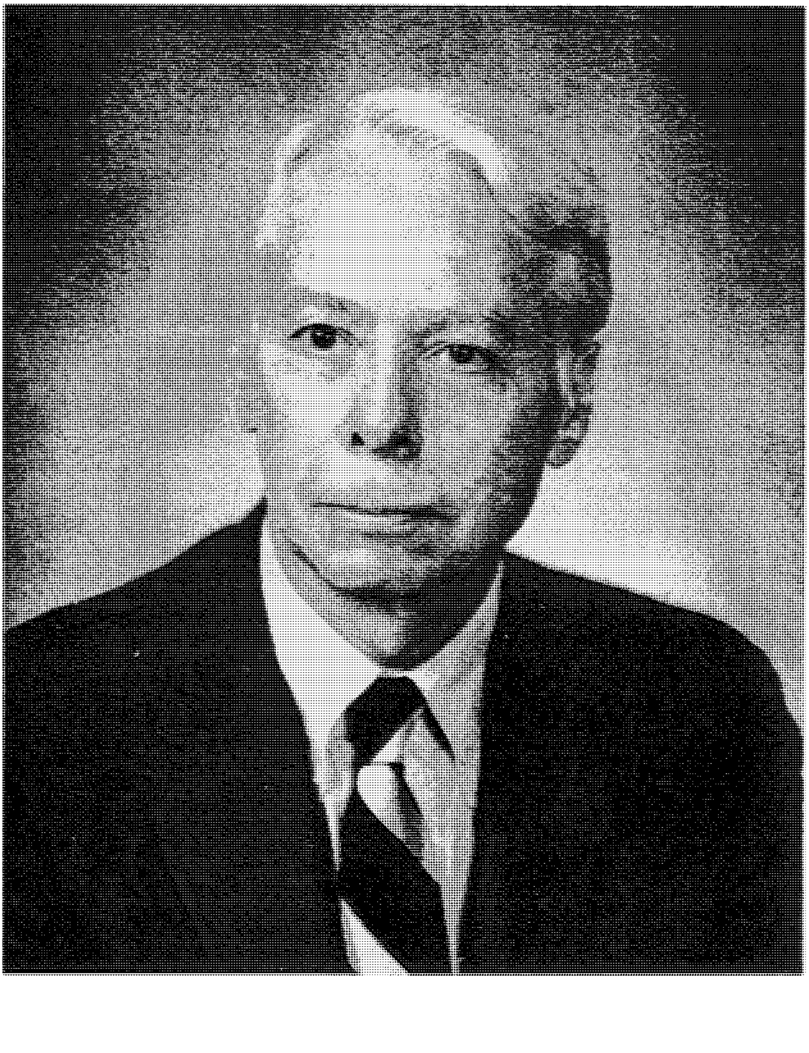
Proceedings of a Conference
to Honor Alvin M. Liberman

Psychology Press

Edited by
Ignatius G. Mattingly
Michael Studdert-Kennedy

MODULARITY AND THE MOTOR THEORY OF SPEECH PERCEPTION

Proceedings of a Conference
to Honor Alvin M. Liberman



MODULARITY AND THE MOTOR THEORY OF SPEECH PERCEPTION

Proceedings of a Conference
to Honor Alvin M. Liberman

Edited by
Ignatius G. Mattingly
Michael Studdert-Kennedy
Haskins Laboratories
New Haven, Connecticut

Psychology Press
Taylor & Francis Group
NEW YORK AND LONDON

First Published 1991 by Lawrence Erlbaum Associates, Inc.

This edition published 2014 by Psychology Press
711 Third Avenue, New York, NY 10017

and by Psychology Press
27 Church Road, Hove, East Sussex, BN3 2FA

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

Library of Congress Cataloging-in-Publication Data

Modularity and the motor theory of speech perception / edited by Ignatius G. Mattingly,
Michael Studdert-Kennedy.
p. cm.
"Proceedings of a conference to honor Alvin M. Liberman."
Includes bibliographical references and index.
ISBN 0-8058-0331-9
1. Speech perception—Congresses. 2. Liberman, Alvin M. (Alvin Meyer)—
Congresses. I. Mattingly, Ignatius G. II. Studdert—Kennedy, Michael.
III. Liberman, Alvin M. (Alvin Meyer)
QP399.M63 1991
612.7'8—dc20 90-44718
CIP

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 may be apparent.

ISBN-13: 978-1-315-80794-2 (ebk)

Contents

Page Intentionally Left Blank

Contributors and Participants

Ursula Bellugi
The Salk Institute
San Diego, California

Paul Bertelson
Laboratoire de Psychologie
expérimentale
Université Libre de Bruxelles
Brussels, Belgium

Sheila Blumstein
Department of Linguistics
Brown University
Providence, Rhode Island

Albert Bregman
Department of Psychology
McGill University
Montreal, Quebec

Catherine Browman
Haskins Laboratories
New Haven, Connecticut

J. C. Catford
Department of Linguistics
University of Michigan
Ann Arbor, Michigan

Franklin S. Cooper
Haskins Laboratories
New Haven, Connecticut

Stephen Crain
Haskins Laboratories
New Haven, Connecticut

C. J. Darwin
Laboratory of Experimental
Psychology
University of Sussex
Brighton, England

Beatrice de Gelder
Laboratoire de Psychologie
expérimentale
Université Libre de Bruxelles
Brussels, Belgium

Peter Eimas
Department of Psychology
Brown University
Providence, Rhode Island

Janet Fodor
Department of Linguistics
The Graduate School, City
University of New York
New York, New York

Jerry Fodor
Department of Philosophy
The Graduate School, City
University of New York
New York, New York

Carol Fowler
Haskins Laboratories
New Haven, Connecticut

Lyn Frazier
Department of Linguistics
University of Massachusetts
Amherst, Massachusetts

Osamu Fujimura
Division of Speech and Hearing
Research
Ohio State University

Merrill Garrett
Department of Psychology
University of Arizona
Tucson, Arizona

Lila Gleitman
Department of Psychology
University of Pennsylvania
Philadelphia, Pennsylvania

Louis Goldstein
Haskins Laboratories
New Haven, Connecticut

Paul Gorrell
Linguistics Program
University of Maryland
College Park, Maryland

Mark Haggard
Institute of Hearing Research
Medical Research Council
University of Nottingham
Nottingham, England

Daniel Holender
Laboratoire de Psychologie
expérimentale
Université Libre de Bruxelles
Brussels, Belgium

James J. Jenkins
Department of Psychology
University of South Florida
Tampa, Florida

Peter Jusczyk
Department of Psychology
University of Oregon
Eugene, Oregon

Dennis Klatt (deceased)
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Edward S. Klima
Department of Linguistics
University of California
San Diego, California

Masakazu Konishi
Division of Biology
California Institute of Technology
Pasadena, California

Harlan Lane
Department of Psychology
Northeastern University
Boston, Massachusetts

Alvin M. Liberman
Haskins Laboratories
New Haven, Connecticut

Mark Liberman
AT&T Bell Laboratories
Murray Hill, New Jersey

Björn Lindblom
Department of Linguistics
University of Texas
Austin, Texas

Peter MacNeilage
Department of Linguistics
University of Texas
Austin, Texas

Virginia Mann
Department of Cognitive Sciences
University of California
Irvine, California

Daniel Margoliash
Department of Anatomy
University of Chicago
Chicago, Illinois

Ignatius G. Mattingly
Haskins Laboratories
New Haven, Connecticut

Jacques Mehler
Laboratoire de Psychologie
Paris, France

Joanne Miller
Department of Psychology
Northeastern University
Boston, Massachusetts

Helen Neville
The Salk Institute
San Diego, California

Fernando Nottebohm
Rockefeller University
New York, New York

David Pisoni
Department of Psychology
Indiana University
Bloomington, Indiana

Howard Poizner
Center for Molecular and Behavioral
Neuroscience
Rutgers University
Newark, New Jersey

Robert Remez
Department of Psychology
Barnard College
New York, New York

Bruno H. Repp
Haskins Laboratories
New Haven, Connecticut

Lawrence Rosenblum
Department of Psychology
University of California
Riverside, California

Arthur Samuel
Department of Psychology
Yale University
New Haven, Connecticut

Donald Shankweiler
Haskins Laboratories
New Haven, Connecticut

Kenneth N. Stevens
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Michael Studdert-Kennedy
Haskins Laboratories
New Haven, Connecticut

Quentin Summerfield
Institute of Hearing Research
Medical Research Council
University of Nottingham
Nottingham, England

Marilyn Vihman
Haverford College
Haverford, Pennsylvania

Janet Werker
Department of Psychology
University of British Columbia
Vancouver, British Columbia

Preface

This book is the proceedings of a conference held in New Haven, Connecticut, June 5–8, 1988, sponsored by Haskins Laboratories, and entitled "Modularity and the Motor Theory of Speech Perception." The purpose of the conference was to honor Alvin Meyer Liberman for his outstanding contributions to research in speech perception since he joined the Laboratories in 1944.

Liberman's first contribution, in collaboration with Franklin Cooper, Pierre Delattre, and others, was to invent a way to do speech perception research. Natural speech signals are extremely complex: Their perceptually significant components cannot be readily isolated by filtering or by temporal segmentation. But early work at Haskins with the Pattern Playback synthesizer had shown that spectrotemporal patterns modeled on those of natural utterances, but highly simplified, could be used to synthesize intelligible speech. Liberman and his colleagues demonstrated that valid and reliable conclusions about speech perception could be based on naive subjects' judgments of such synthetic speech, generated from carefully controlled patterns.

Using this method, Liberman and his colleagues proceeded to identify and describe the speech cues, the acoustic events that support the perception of particular phonetic categories. During the 1940s and 1950s, they studied the sounds of English, manner class by manner class, opening up the field of acoustic phonetics and laying the foundation for speech synthesis by rule.

Spurred by the observation that speech was far more efficiently perceived than the nonphonetic acoustic substitutes for letters they had hoped to use in a reading machine for the blind, Liberman and his colleagues also investigated differences between the perception of speech and the perception of other acoustic signals. Over the years, they discovered a range of effects, from categorical perception and right-ear advantages in dichotic listening to trading relations and duplex perception, that could not be readily explained on psychoacoustic grounds.

Such findings as these encouraged Liberman to develop the Motor Theory, an account of the psychology of speech perception that had been adumbrated in some of the earliest Haskins papers. As currently formulated, the theory makes two related claims: First, that the entities perceived are not acoustic or auditory events as such, but articulatory gestures; second, that the perception of speech, together with other psycholinguistic processes, is the business of a special neural mechanism—a module, in Jerry Fodor's sense. These ideas have developed over many years. In earlier formulations, it was the listeners' own articulatory productions that guided their perceptions of speech; more recently, it is suggested that an abstract vocal tract model determines both speech production and speech perception. Again, it was proposed earlier that "speech is special," with the implication that speech perception was totally different from any other perceptual process; more recently, speech perception, though still having its own peculiar domain, is seen as one of a class of modular perceptual processes. Finally, and perhaps most importantly, Liberman's perspective has become increasingly biological; on his present view, speech perception has more in common with echolocation in the bat than with perception of Morse code in the human.

It would perhaps have been nice to say, at this point in Liberman's career, that these views of his had found widespread acceptance. Such, however, is far from the case. Liberman's ideas were controversial when first proposed and have remained controversial ever since. What *can* be said is that they have been extraordinarily influential, in the sense that a large fraction of the research in speech perception during the past 30 years has consisted of attempts to corroborate or disprove them.

Under these circumstances, the customary procedures for honoring a distinguished scholar on the occasion of his retirement seemed to us inappropriate. We could, indeed, have planned a conference in which all the participants agreed with Liberman. But surely the best way to honor a controversial figure is to continue the controversies he has provoked. Therefore, we decided to invite both critics and supporters to comment on Liberman's ideas and their implications, not only for speech perception and production but for such arguably related areas at the production and perception of sign language, perception in nonhuman animals, lipreading, language acquisition, sentence processing, reading, and learning to read. An introduction by Franklin Cooper was followed by presentations from fourteen speakers. Each of these presentations was commented on briefly by another speaker. There were also three panel discussions; one member of each panel acted as reporter. A summary by James Jenkins concluded the conference. All this material is included here, except for a few of the comments, written versions of which were not received by our deadline. Finally, it seemed only fair to give Liberman an opportunity to react to the conference after seeing the written versions of the papers and comments; his reflections appear at the end of the book.

The editors would like to thank Yvonne Manning, Joan Martinez, Nancy O'Brien, and Zefang Wang for their generous assistance in preparing the manuscript, and Diana Fish for her skillful indexing. We are particularly grateful to Alice Dadourian not only for editorial advice and assistance, but also for the efficiency and enthusiasm with which she handled the logistics of the conference itself.

I.G.M.
M.S.-K.

Page Intentionally Left Blank

Chapter 1

Introduction: Speech Perception

Franklin S. Cooper
Haskins Laboratories

Welcome to the Conference on Modularity and the Motor Theory of Speech Perception. It is a real pleasure to see so many old friends and to greet those of you whom I have known only by reputation—a pleasure, too, to welcome you graduate students on whom the future of speech research depends. If you are wondering about the viability of a field of research that is already honoring one of its pioneers, the papers you are about to hear will make it clear, I think, that there are more problems ahead of you than there are solutions behind us greybeards. For example: Modularity and the Motor Theory. So welcome to the intellectual challenges as well as to this conference!

To Al Liberman, who is himself an old hand at conferences, this one must be something of a novelty: It was arranged *for* him, not *by* him! It is entirely appropriate that Haskins Laboratories should wish to honor him. Al has been a coworker and a cobeliever in Haskins Laboratories ever since he joined it in 1944 and a continuing inspiration to all of us, both personally and intellectually. He still wanders the halls asking, "What have you discovered today?" It is doubly appropriate that he be honored by a conference on Modularity and the Motor Theory of Speech Perception, since these ideas have been central to his own work and to the many contributions he has made to speech research. I could say more—much more—in the same vein but will limit it to one personal comment: To me, Al has been a friend, and I am the one honored.

Let me consider with you some simple-minded questions. How does it happen that we are here to talk about the Motor Theory of Speech Perception? (I shall leave Modularity aside for a moment.) Part of the answer lies in the history of the field, and as we probe that history—for the benefit of you younger people—we shall find even prior questions. Thus, talking about a theory implies some kind of problem for that theory to explain. Was there such a problem? This may seem a strange thing to ask, since the question of how speech is perceived has been a thorny problem for as long as most of you can remember. Nearly as ancient is the Motor Theory as a proposed solution.

But there was a time when even the problem did not exist—or was not known to be a problem. In the same sense, gravity was not a problem before Newton's time: Everybody knew that apples fell down just as everything else did. So likewise the perception of speech posed no special problem; it and other sounds were heard and recognized all in the same general way.

Let me press the parallel a little farther: Neither Isaac Newton nor Alvin Liberman *discovered* his problem until it fell on him. Newton can now be dismissed, though we should note that it is not every man of science who provides his own problem as well as its solution.

Back to Al and how he discovered *his* problem: Namely, how is speech perceived? He did not begin with speech. The problem that he and I were working on at the end of World War II was the practical one of designing a reading machine for blinded veterans. Our approach was simple and direct: The machine would scan a line of type and convert the distinctive letter shapes into distinctive sound shapes which the blind reader would, with practice, come to recognize—and so to read printed books by ear.

The difficulty that we encountered—as did others before and after us—was that the reading rates were so painfully slow, even after hours and hours of practice, that no one would use the device. We tried many things to make the sounds more distinctive and more easily learned, but reading rates were no better and often worse. Most frustrating was that the performance of our subjects when identifying our machine-made words was much poorer than their performance when identifying nonsense words, spoken by a person.

Thus did Al's problem come down upon him: Finally, he realized that the right question was not why machine-made sounds are so poor but rather why man-made sounds are so good. What is so special about speech that makes its perception so easy?

He then supposed that speech was just a better acoustic alphabet—that it took the phonetic string of a sentence and spelled it out with unit sounds that could be heard easily and rapidly, because they flowed together into words. By studying these unit sounds of speech, he might be able to design a better set of sounds for the reading machine.

But by this time, the Potter, Kopp, and Green (1947) collection of spectrograms had been published, and one could see that finding acoustic invariants for the phonemes would not be so easy. One could pick out some of the acoustic consequences of articulation, but where in all this complex pattern were the acoustic cues for perceiving the individual speech sounds known to be lurking there?

This search for the acoustic cues was the task that Al, Pierre Delattre, and I undertook in the early 1950s using spectrograph and pattern playback. What we found was well known at the time and is still available in the literature. Cues there were—in abundance and extreme diversity. Before the end of the decade, most of them had been found and organized into rules for synthesis that generated quite intelligible speech (Liberman, Ingemann, Lisker, Delattre, & Cooper, 1959).

But it was the diversity and curious character of the cues that needed a better explanation than current auditory theories of perception could provide. The cues for a particular speech sound seemed to make sense only when one considered how that sound had been articulated. Al made these arguments explicit in his 1957 (Liberman, 1957) review paper and offered a motor theory to explain why speech is so exceptionally efficient as a carrier of messages.

Thus history, not logic, is the principal reason we are here to talk about a motor theory of speech *perception* rather than a motor theory of speech *production.*

There were other reasons, too. There was then a bias—which still persists—toward thinking about speech as "that which goes into the ear" rather than "that which comes out of the mouth." Little wonder, since the ear and its roots in the brain are so much more elegant and mysterious than the mouth's crossed-up plumbing and ventilating systems, which can't even breathe and swallow at the same time! Then, too, instrumentation was largely lacking for research on production.

Let me add as an aside that although Al continued to focus on the perception of speech and its many unique characteristics, there were some of us here who did start, in the late 1950s, to look for phonological structure on the production side. The Laboratories still has a major program ongoing in this area, and we are by no means alone.

Now, what would be different if we were talking about a motor theory of speech production instead of a motor theory of speech perception? Surely there must be close linkages between the two processes and their mechanisms unless, indeed, a single mechanism performs both functions. But whatever the internal structure of the speech module (or modules), the input and output signals are very different in kind and structure. This calls for a restructuring operation somewhere in the sequence—one that may put tighter constraints on a model for the speech module than do either perception or production.

So another question: Should we perhaps be talking about a motor theory of speech *per se,* where "speech" stands for "communication by voice?" This would emphasize the communicative function that is served by *both* perception and production. Moreover, it would give central place to the operation that ensures error-free regeneration of spoken messages, even when repeated many times.

You may object to so much emphasis on the relaying of spoken messages from person to person, since it is so rarely done. The point is that it can be done; the mechanism is in place and in use for other purposes. Long ago, this kind of relaying was common; indeed, speech—aided by rhyme—served to repeat epic poems intact across the ages. The trick, just as with long-distance telephony now that it has gone digital, is to regenerate the signal each time it is relayed. The incoming signal, contaminated with noise and distortion, is replaced by a shiny new signal in canonical form. For humans, the regenerated signals serve a further purpose: They are just what is needed for memory, since the bit rate for identifying the message units is so much less than for describing the incoming sounds.

Regeneration is only one of several names for the function I have been talking about. Categorization is an essential part of the function, and with labeling included it provides the recognition stage in models of speech perception. Restructuring, or recoding, are also closely related terms. In models of speech production, the generative part of regeneration corresponds to setting up motor plans or coordinative structures. I have used the term "regeneration," because it relates to both input and output and implies the communicative function of which it is an essential part.

Clearly, regeneration also implies *units*. In their canonical form, these would be the "intended gestures" of the motor theory. But surely these are only a subset of all possible gestures, so what constrains the choice? Speed of execution is one requirement. In fact, people can and do talk at rates of up to fifteen or so units per second—which seems impossibly high for such slow machinery as tongue and jaw. So we should not expect speech gestures to conform to our usual notion of a completed movement such as a nod of the head or a wave of the hand. No amount of coarticulation between such gestures (i.e., overlap along the time line), would crowd them into the time allowed.

But coarticulation across the time line could do it. Given the several articulators that we have and their potential for independent and concurrent action, the total system could achieve a succession of discrete states—nameable as phonemes or intended gestures—and so attain a kind of phase velocity much higher than that of the individual articulators. It may be comforting to note that this way of looking at speech—searching for coincidences and alignments during ongoing gesturing—conforms to the cosmic strategy whereby astrologers seek our destinies in planetary alignments.

Another constraint on the choice of gestures is the fairly obvious one that they must have acoustic consequences. Preferably, the consequences would be as strong and distinctive as they are for [s] and [ʃ], but given the nature of the gestures, most of the sounds are necessarily variable with context and some, to round out the inventory, are even as feeble and confusable as [f] and [θ].

A more demanding requirement is that the units be *permutable*. Thus, assuming speech to be a succession of discrete states that progresses from one intended gesture to the next, then the set of possible "next gestures" from any particular state is small and sharply constrained. It is limited—not by phonological rules—but by circumstances such as that some of the articulators are already in midmovement and must, therefore, continue moving in the next gesture. In a more general way, one of the prices of parallelism is that there is no way to extract a time slice without leaving rough edges, so shuffling its position means finding a place where the edges will match.

It might be useful to turn this argument on its head and use the permutability requirement to reinterpret our knowledge of how real phonemic units combine and recombine. That could help us to arrive at physiological descriptions of the "intended gestures."

Much of what I have been saying has dealt with the constraints that particular processes put on models for speech. Let me now try a different tack and ask about *minimal* constraints on the speech signal at various stages of the communicative process: Thus, what requirements at the very least must the unit signals of speech meet, if they are to be useful in perception, in production, and in such intermediate processing as may be needed to link perception and production? And, having asked these questions, let me propose answers: For perception, the signals must at the very least be audible; for production, they must be utterable; and for the intermediate processing, they must be both regenerable and permutable; it would help, if they were also memorable. The moral I would draw is obvious: The constraints that really bind are the need to regenerate and the need to permute the signal units.

Finally, let me return to my original question, slightly sharpened: We are, in fact, met here to talk about Modularity and the Motor Theory of Speech Perception. Does that emphasis on perception mean that we are "barking up the wrong tree?" Like most simple-minded questions, this one has two answers: YES, if we suppose that perception is all-important, or that it can be dealt with in isolation. NO, if we consider that perception by itself is a very large topic for a single conference, and if we remember that the models we build for perception must be compatible with the rest of the communicative process; that is, they must honor the Throughput Principle: That which goes in at the ear, and out from the mouth, must somehow go through the head.

References

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29,* 117–123.

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P. C., & Cooper, F. S. (1959). Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America, 31,* 1490–1499.

Potter, R. K., Kopp, G. A., & Green, H. (1947). *Visible speech.* New York: Von Nostrand.

Page Intentionally Left Blank

Chapter 2

The Status of Phonetic Gestures

Björn Lindblom

Department of Linguistics, University of Texas, and University of Stockholm

Abstract

In this chapter, I shall argue that speakers adaptively tune phonetic gestures to the various needs of speaking situations (the plasticity of phonetic gestures) and that languages make their selection of phonetic gesture inventories under the strong influence of motor and perceptual constraints that are language independent and in no way special to speech (the functional adaptation of phonetic gestures). These points have implications for a number of issues on which the Motor Theory takes a stance. In particular, the evidence reviewed challenges two assumptions that are central to the Motor Theory—that of modularity and gestural invariance. First, if phonetic gestures possess invariance at the level of motor commands, and listeners are able to perceive such gestural invariance, why is speech production so often nevertheless under output-oriented control? Second, the Motor Theory assumes that speech perception is a biologically specialized process that bypasses the auditory mechanisms responsible for the processing of nonspeech sounds. It also assumes that the motor system for vocal tract control exhibits specialized adaptations. If so, why do inventories of vowels and consonants nevertheless show evidence of being optimized with respect to motoric and perceptual limitations that must be regarded as biologically general and not at all special to speaking and listening?

There are two aspects of phonetic gestures that merit special attention in the context of the Motor Theory (MT), (Liberman & Mattingly, 1985). One striking fact comes from observations of how speech is produced: A large body of experimental evidence suggests that phonetic gestures are highly malleable and adaptive. They exhibit *plasticity*.

The second point emerges from cross-linguistic data on how languages select gestures to build segment inventories: Phonologies are "quantal" in that they use similar gestures drawn from a remarkably small universal set (Stevens, 1989). Moreover, in individual languages, the selection of vowel and consonants from this set is systematic and lawful. It is governed by certain "implicational laws" (Jakobson, 1968; Lindblom & Maddieson, 1988).

As we try to explain why systems of phonetic gestures exhibit these quantal and implicational properties, we are led to argue that they are selected so as to meet collectively a demand for "sufficient perceptual contrast." Developing this point, we shall suggest that phonetic gestures can be seen as *adaptations* to constraints on motoric and perceptual mechanisms that are language independent and not special to speech.

The plasticity of phonetic gestures is a phenomenon that any theory aimed at resolving the issue of *phonetic invariance* (Perkell & Klatt, 1986) must account for. The MT addresses this issue by claiming as Liberman and Mattingly (1985) wrote: "The objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands" (p. 2). Furthermore, viewing phonetic gesture inventories as adaptations to nonspecial input/output mechanisms poses another interesting problem for the MT that argues that both the production and the perception of speech are "modular," biologically specialized processes. Let us see where contrasting these views will lead us.

The Plasticity of Phonetic Gestures

The MT Model of Speech Production

Liberman and Mattingly (1985) took the following stance on the invariance issue:

> Phonetic perception is perception of gesture. . . . [They further state] the invariant source of the phonetic percept is somewhere in the processes by which the sounds of speech are produced. (pp. 21–23)

The authors recognized the complexity and variability that phonetic gestures exhibit in instrumental analyses but claimed (Liberman & Mattingly, 1985):

> It is nonetheless clear that, despite such variation, the gestures have a virtue that the acoustic cues lack: Instances of a particular gesture always have certain topological properties not shared by any other gesture. [They conclude that] the gestures do have characteristic invariant properties, as the motor theory requires, though these must be seen, not as peripheral movements, but as the more remote structures that control the movements. These structures correspond to the speaker's intentions. (p. 23)

Vowel Reduction

We can illustrate this theory of invariance with some examples of the so-called *undershoot* phenomenon (Lindblom, 1963). Figure 2.1 shows spectrograms of three English words containing one, two, and three syllables: *muse, music, musically*. The increase in word length is correlated with a shortening of the initial, stressed vowel. This durational variation is associated with shifts in the extent to which mid-vowel formant patterns approach a hypothetical target. Note the extent of the F2 contour, which shows a clear dependence on vowel duration. The tongue, initially in a palatal position, undershoots its velar /u/ target more and more as the vowel becomes shorter. Note that these samples are all from syllables carrying lexical main stress. Therefore, we are justified in calling the phenomenon illustrated in Fig. 2.1 *duration-dependent* undershoot.

In conformity with the Liberman–Mattingly model of speech production, it seems possible to suggest that the undershoot effect is due to the spatial and temporal overlap of adjacent motor commands. The durational variations induced by changing word length cause differences in timing of the motor commands, and, provided that the time constants of the articulators are assumed not to change, the MT makes the correct prediction that in a particular context reaching the target configuration of the stressed vowel is a function of the duration of the vowel. Because undershoot is lawfully related to the duration and the context of the vowel gestures, it is possible to claim that something nonetheless remains invariant: the underlying intention, or "Lautabsicht" (Lindblom, 1963). On this view of speech, therefore, the task of the listener becomes that of inferring the intended gestures from highly encoded and indirect acoustic information.

For biomechanical reasons, the simple undershoot model may still be said to have a certain validity. However, there are complications. Mainly they arise from the fact that in natural speech a speaker's intentions go far beyond that of merely producing a sequence of invariant phonetic gestures. We begin to see these complications as soon as we broaden the scope of our inquiry and approach slightly more ecological speaking conditions than those normally studied in our laboratories.

Apparently speakers are free to vary degree of undershoot somewhat independently of vowel duration. This is evident from studies indicating on the one hand that in fast speech, articulatory and acoustic goals can be attained *despite* short segment durations (Engstrand, 1988; Gay, 1978; Kuehn & Moll, 1976) and on the other hand that reductions can occur *despite* adequate duration (Nord, 1986). How talkers go about varying degree of undershoot is not known. One possibility is that deviations from duration-dependent undershoot might be due to processes such as "overarticulating" and "underarticulating" (cf. discussion of "clear speech" following). The observed deviations of duration-dependence obviously constitute an embarrassment for the simplest version of the undershoot model

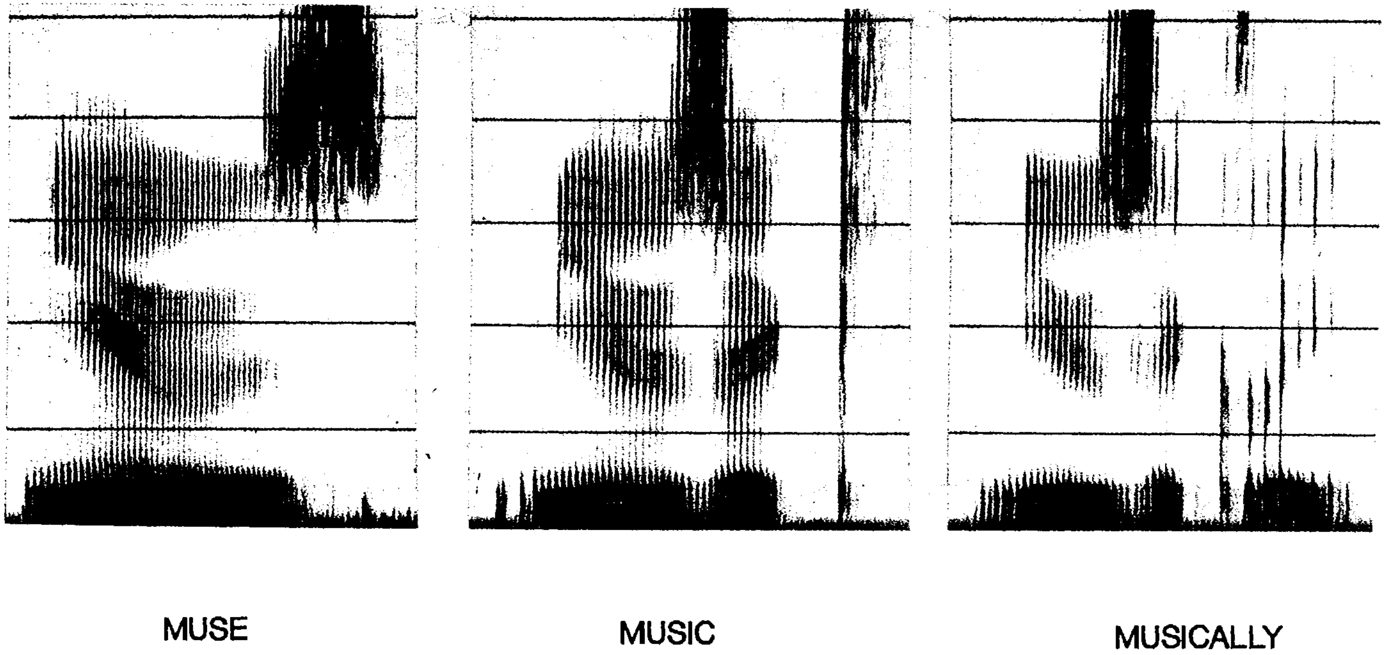


FIG. 2.1. Vowel reduction and 'duration-dependent formant undershoot" (Lindblom, 1963). Spectrograms of three English words (from left to right): *muse, music, musically.* Note the variation in the duration of the initial vowel and the associated changes in the frequency contours of F2.

(Lindblom, 1963). An improved model is clearly needed capable of capturing the malleability of phonetic gestures.

Compensatory Articulation

Speakers are in fact capable of reorganizing phonetic gestures so as to reach constant acoustic and perceptual goals. This has been shown most clearly by experiments on *compensatory articulation* in which atypical jaw positions are induced by means of so-called "bite blocks" (Lindblom, Lubker, & Gay, 1979; Lindblom, Lubker, Lyberg, Branderud, & Holmgren, 1987). The relative ease with which speakers adapt to an unnatural bite block can be accounted for by assuming also that normal speech motor control is intrinsically compensatory. Although the bite block must be overcome by invoking rather extreme articulations, the compensation occurs effortlessly, because not only speech but motor behavior in general is organized to be compensatory. For the sake of those who take a dim view of bite block experiments and remain unconvinced by claims that bite block speech tells us anything at all about normal speech, let us examine another case of compensation, but one found in a more ecological speaking situation.

Loud Speech

Consider the control of vowel duration in *loud speech*. Speakers have been shown to use larger jaw openings when speaking louder. The effect is independent of vowel identity and has been demonstrated for several languages (Schulman, 1989). Now this raises a problem for the production of loud vowel duration in the following way. Recall the Extent of Movement Hypothesis proposed by Fischer-Jørgensen (1964). It explains why, everything else being equal, open vowels universally tend to be longer than close vowels. The main effect is that in an open vowel occurring in a CVC environment, the jaw moves further than in a close vowel in the same context. Using a quantitative articulatory model formalizing Fischer-Jørgensen's idea, I showed for /ibVbi/-utterances (Lindblom, 1967) that owing to these differences in jaw movement, the release of the first /b/ will occur more and more prematurely, and the implosion of the second /b/ will be increasingly delayed as the degree of opening of the vowel is increased. In addition to supporting the Extent of Movement Hypothesis, these model experiments indicate that the effect can in fact be so drastic that unless the lip gestures for the /b/:s are reorganized to compensate for the jaw movement, unacceptably large durational differences between open and close vowels will result. The need for such compensation was indeed substantiated by the lip and jaw measurements of the same study (Lindblom, 1967, e.g., Fig. I-A-14).

Because loud speech uses more open jaw positions, the Extent of Movement Hypothesis applies also to that style of speech. Experimental data (Schulman, 1989; Lindblom, 1987) show that the increased jaw openings of loud vowels are compensated for by other articulators in order to make vowel durations of loud and normal conditions more similar than they would have been without compensatory maneuvers.

Clear Speech

We recently began a series of studies aiming at describing the acoustic properties of clear speech. Presumably when people speak more clearly, they do so in an effort to become more intelligible. One issue is whether this speaking style differs from more neutral speech mainly in that its signal-to-noise ratio is better or whether it also involves a reorganization of phonetic gestures and acoustic patterns. There is evidence indicating that such reorganization does indeed take place and can be rather extensive (Picheny, Durlach, & Braida, 1986; Uchanski, Durlach, & Braida, 1987).

We have preliminary data on American English vowels (Moon & Lindblom, 1989) produced in contexts that meet the following conditions:

1. The vowels and their consonantal environments should be chosen so as to maximize large "locus-to-target" distances (e.g., front vowels occurring in a labio-velar environment: *wheel, will, well, wail*)
2. The vowels should carry lexical main stress
3. They should vary in duration.

The latter two requirements were met by making use of the so-called "word length effect." The length of the test words was varied by adding *-ing* and *-ingham* to the CVC sequence under analysis, which produced series such as *will, willing, Willingham,* and so forth. Subjects were asked to read randomized lists of such tokens. Initially they were instructed to adopt a comfortable tempo and vocal effort but received no specific instructions otherwise. We refer to these speech samples as citation form speech (CF). In the second half of the recordings, they read similar lists but were now explicitly told to overarticulate and to speak as clearly as possible (CS lists). Measurements were made of vowel duration and of formant frequencies at points of minimum rate of change in the vowels and of the locus pattern of the consonants.

Plots of formant frequencies versus vowel duration were prepared for all the test items. The vowel formant patterns of both CF and CS samples were found to exhibit duration-dependent undershoot. For both styles, the data points tended to cluster in` ways that could be described in terms of exponential curves similar to those used in Lindblom (1963). However, there were significant differences: Overarticulated vowels were consistently of longer duration, and for every vowel examined, the CS undershoot curve was different from the corresponding CF curve. These differences can be summarized by saying that for each individual vowel the asymptotes of the exponentials tend to be located much closer to the formant values observed for null-context environments such as /h–d/. Plotting the data on an F1/F2 vowel chart, we observe that the CS vowel space invokes values that are more peripheral and closer to the /h–d/ targets than the CF tokens, which are more context-sensitive and, hence, more centralized in the formant space.

The analysis of the investigation from which these observations are taken is still in progress. In the near future, we expect to be able to give a more comprehensive report on the robustness and generality of the observed effects across a wide range of speakers and contexts. Nevertheless, a trend fully compatible with previous work on CS acoustics (Picheny et al., 1986, Uchanski et al., 1987) is evident in the patterning of the data, which so far suggest that it does not merely improve the S/N ratio. Clear speech is a transform that tends to enhance the acoustic contrast among vowel phonemes, making their formant patterns less dependent on context and more widely dispersed.

If our preliminary results are further corroborated, we must ask: Why should there be such a thing as clear speech? Why do talkers bother to make extensive adjustments of their phonetic gestures and the associated acoustic patterns? Is it because in so doing they facilitate the listener's access to the distal objects of perception: the underlying phonetic gestures (cf. the Motor Theory)? Instead, is it because they thereby make acoustically stable and salient properties of the signal easier to identify (cf. the Quantal Theory of Speech)? Finally, is it—as we prefer to argue—because lexical access is based on "sufficient contrast" (cf. the Theory of Adaptive Dispersion as presented in the following)?

Is Invariance Necessarily Phonetic?

How do we account for the variance of phonetic gestures that we observe in compensatory articulations, in loud speech, and in clear speech? No doubt proponents of the MT would pin their hopes on future research demonstrating how the speech system succeeds in computing a family of gestures that, in spite of substantial surface variability, topologically share certain unique properties and nevertheless manage to remain *motorically* invariant.

However, faced with a rather impressive body of evidence on the plasticity of motor gestures in general and phonetic gestures in particular, we are easily persuaded by an alternative vision according to which invariants will ultimately have to be defined in terms of the *purpose* and *primary ecological function* of the gestures, namely lexical access, comprehension, and social interaction. On this view, phonetic gestures should not be expected to be motorically invariant, because they are merely adaptive and malleable means to more global communicative ends.

Why then are we looking for phonetic invariance? Is it not needed for satisfactory lexical access? Here is a summary of an argument that leads us to conclude that in principle it is indeed dispensable.

We begin by noting that the structure of all languages exhibits redundancy and that the perception of speech is the product of two types of information: signal-driven and signal-independent information. As a consequence of redundancy, the words and phonemes of individual utterances show short-term variations in predictability. Consider the following two utterances[1]:

Utterance A: *A stitch in time saves* ______.
Utterance B: *The next number is* ______.

A reduced, articulatorily simplified pronunciation of *nine* would stand a better chance of being correctly identified in utterance A than in utterance B. Whether reduced or not, any phonetic form that is correctly identified would by definition be perceptually adequate (sufficiently rich). From the viewpoint of lexical access, such a form can be said to exhibit sufficient perceptual contrast.

These considerations lead us to conclude that phonetic invariance is not necessarily essential for lexical access. Speech signals will be adequate for lexical access as long as they are rich enough to match in a complementary fashion the listener's running access to signal-independent information. According to this theory, therefore, the critical condition that phonetic gestures must meet is that they be *perceptually sufficiently contrastive.*

Coarticulation

With the idea of "sufficient perceptual contrast" in mind, let us take a new look at some well-known measurements often referred to in discussions of consonant–vowel coarticulation. Early work on the acoustic patterns of synthetic speech led Haskins researchers to conclude that the objects of speech perception were not to be found at the acoustic surface but might be sought in upstream invariant motor processes. In 1966, Öhman published his spectrographic measurements on V_1CV_2 sequences. His results give a vivid demonstration of massive coarticulation effects and seem at least at first glance to lend strong support to the Haskins idea that there is simply no way to define a phonetic category in purely acoustic terms.

To make this point, we reproduce one of Öhman's diagrams in Fig. 2.2, an illustration, as good as any, of the observation that place information for a given consonant is carried by a rising transition in one vowel context and a falling

[1] In my choice of these examples I am indebted to Lieberman (1963).

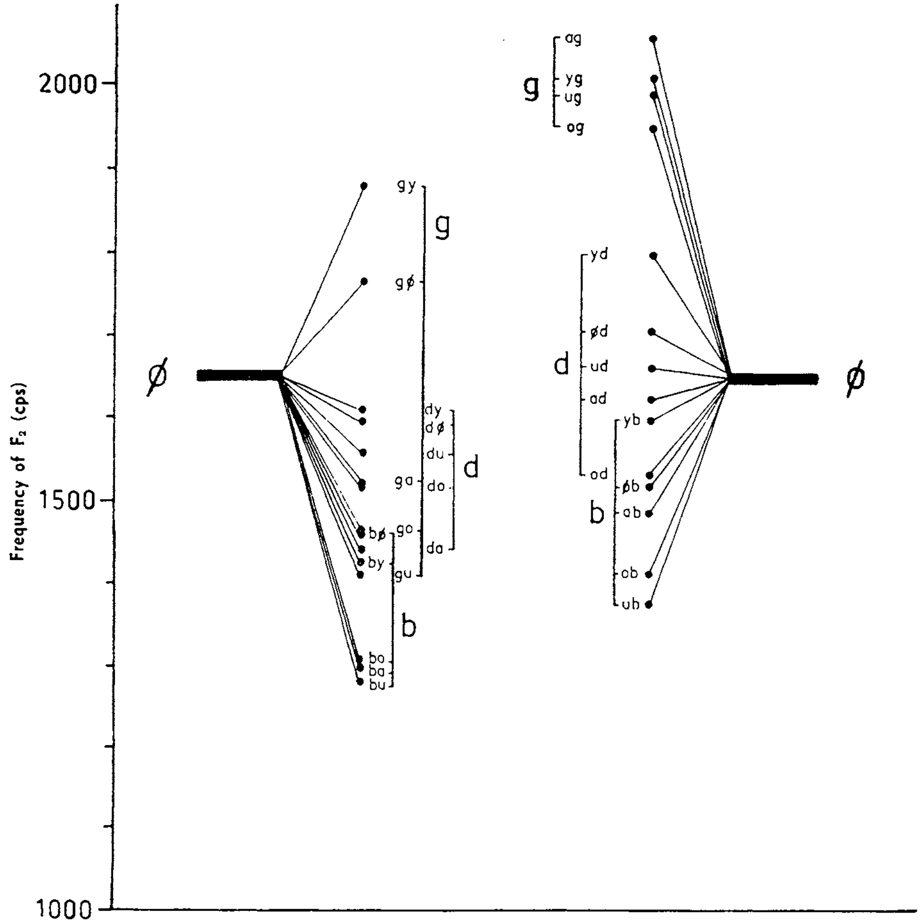


FIG. 2.2. Formant transitions and consonant-vowel coarticulation. Stylized second-formant transitions observed in VCV utterances. The symbols at transition endpoints identify the following and preceding contexts respectively (adapted from Öhman, 1966, with permission).

transition in another (Liberman, Delattre, Cooper, & Gertsman 1954).

However, although admittedly complex, do acoustic patterns of this kind really justify the conclusion that there is simply no way to define a phonetic category in purely acoustic terms? Let us replot the Öhman data as shown in Fig. 2.3.

The data points pertain to F2 and F3 of the CV_2-boundary (x- and y-axes) and

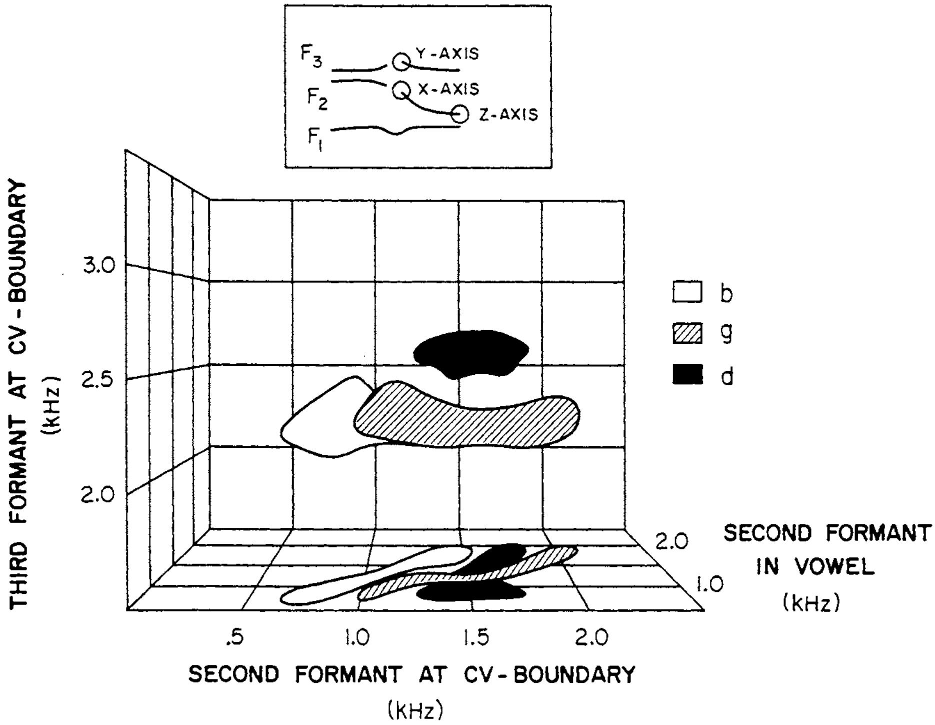


FIG. 2.3. A three-dimensional representation of formant measurements at CV-boundary of VCV sequences (Öhman 1966). The "clouds" of the diagram includes all the data in Tables II and IV of the Öhman (1966) article. X-axis: Second formant at CV-boundary. Y-axis: Third formant at CV-boundary. Y-axis: Second formant in final vowel.

to F2 of the V_2 vowel (*z*-axis) and are from his Tables II and IV (Öhman, 1966). We see a three-dimensional view of three "clouds" that correspond to samples of V_1bV_2, V_1dV_2 and $V_1g_2V_2$ utterances, respectively, and that, in spite of all the vowel–consonant coarticulation, do not overlap and hence, are sufficiently distinct from each other.

The implication of this result is this: If we make the reasonable assumption that perception has access to (at least) these three parameters of the VCV utterances, the information available in the acoustic signal should be sufficient to disambiguate the place of the consonants. Needless to say, the three dimensions selected here do not by any means exhaust the signal attributes that might carry place information. One obvious omission is the spectral dynamics of the stop releases. Spectra for /b/ would be relatively weak and flat whereas those for /d/ and /g/ would show distinct stronger energy concentrations of mainly front cavity dependence (Stevens, 1968). Adding such dimensions to the consonant space would be an effective means of further increasing the separation of the three clouds and thus enhancing their distinctiveness.

Please note the following: Given the preceding analysis, we do not, unlike proponents of the MT, need to postulate that a specialized mechanism evolved to handle coarticulation in CV syllables. Phonetic categories are "polymorphous" phenomena (Kluender, Diehl, & Killeen, 1987) that, if sufficiently contrastive perceptually, do the job of differentating lexical items from each other. Their polymorphous nature and the notion of sufficient contrast imply that there is no single necessary or sufficient cue that must always be present for category membership.

This analysis is supported by work on speech perception by animals. Most recently Kluender et al. (1987) demonstrated the ability of Japanese quail to learn to discriminate place in stop consonants and to generalize their judgments to new vowel contexts. These birds are also capable of using cues for voicing, vowel height, and sex of talker. These findings strongly suggest that quail perform well on the discrimination tasks not because they are equipped with a specialized processor for speech, but because they are able to exploit the stimulus properties, and because these properties are acoustically sufficiently rich.

The Linguistic Selection of Phonetic Gesture Inventories: Adaptation to Non-specialized Input/Output Constraints

It appears reasonable to assume that the factors that shape the vowel and consonant inventories of the languages of the world originate in the interactive behavior of speakers and listeners. What is the nature of the selection criteria that might govern the evolution of phonetic systems?

The Quantal Theory of Speech (Stevens, 1989) hypothesizes that languages tend to seek out regions of high *acoustic* and *auditory stability* in the universal phonetic space and that these regions represent the physical correlates of the distinctive features of phonological systems. Both talker-oriented and listener-oriented factors motivate the choice of acoustic stability as a basis for selections.

An alternative theory, the Theory of Adaptive Dispersion (Lindblom, MacNeilage, & Studdert-Kennedy, forthcoming), shares with the Quantal Theory the assumption that the factors shaping phonetic inventories originate in on-line speaker–listener interactions but differs in that it explores the consequences of adopting another selection criterion, namely *sufficient perceptual contrast*. Some of the results obtained within that paradigm bear on the present discussion.

Perceptual Contrast

Let us first look at dispersion and the notion of perceptual contrast. Typological studies of vowel systems (Crothers, 1978; Maddieson, 1984) show that the most favored inventories are drawn from a small subset of the total set of observed qualities. The data of Table 2.1 are from Crothers (1978).

It is evident that languages favor peripheral vowels and that there is a tendency to use many more sonority (open/close) contrasts than chromaticity (front/back and rounded/unrounded) contrasts.

Suppose we approach these observations from the following point of view: *If vowels systems were seen as adaptations to the universal auditory constraints of human hearing, what would they be like?* This is essentially the question that we have addressed in a number of studies. Here is a brief summary of some of the results.

Three studies explore the notion of "maximal perceptual contrast." In Liljencrants and Lindblom (1972), a formant-based distance metric was used to quantify the notion of perceptual contrast and to predict the phonetic values of vowel systems as a function of inventory size. The predictions were successful in reflecting the patterns of *dispersion* clearly evident in the typological data. Their major failure was that, in large systems, too many high vowels were generated.

In Lindblom (1986), the simulations were repeated with a psychoacoustically better-motivated distance metric (Bladon & Lindblom, 1981). This revision led to clear improvements, implying that as our description of the auditory constraints becomes better, so will our predictions. A third study (Lindblom, in press) combines the 1986 model with the results of experiments using Direct Magnitude Estimation (DME). The DME technique was used to compare subjects' judgments of movement along the dimensions of jaw opening and anterior–posterior positioning of the tongue. The results indicated that jaw movements appeared subjectively more extensive than tongue movements, although displacements were equal in terms of physical measures (Lindblom & Lubker, 1985). Incorporating those results into the simulations, we revised the optimization criterion to encompass also articulatory discriminability, departing from the

TABLE 2.1
Most Favored Vowel Systems Observed in a Corpus of Over 200 Languages (Crothers, 1978)

Inventory Size	*Vowel Qualities*	*No of LG's*
3	i a u	23
4	i a u ɛ	13
4	i a u ɨ	9
5	i a u ɛ ɔ	55
5	i a u ɛ ɨ	5
6	i a u ɛ ɔ ɨ	29
6	i a u ɛ ɔ e	7
7	i a u e o ɨ ə	14
7	i a u ɛ ɔ e o	11
9	i a u ɛ ɔ e o ɨ ə	7

TABLE 2.2
Predicted Vowel Systems Derived From Quantitative Simulations (Adapted from Lindblom, MacNeilage & Studdert-Kennedy, forthcoming).

Inventory Size	*Vowel Qualities*
3	i a u
4	i a u ɛ
5	i a u ɛ ɔ
6	i a u ɛ ɔ ʉ
7	i a u ɛ ɑ ʉ ɤ
9	i a u ɛ ɑ e o ʉ ə

assumption that vowels tend to evolve so as to both sound and feel sufficiently different.

Evaluating the results, two things should be noted. The probability of selecting a correct system by pure chance is less than 10^{-3}, irrespective of system size. The predictions are perfect, if we measure agreement between model and data in terms of the number of sonority and chromaticity contrasts. Bearing these points in mind, we see from Table 2.2 that the simulations achieve an extremely close agreement with the typological data.

Adaptive Dispersion

In the three studies reviewed above, articulatory factors play a role in delimiting the phonetic space of "possible vowels" (Lindblom & Sundberg, 1971), but beyond that they are essentially neglected. There is a great deal of evidence (Lindblom et al., forthcoming) indicating that articulation plays an important role and that production constraints tend to counterbalance demands for perceptual contrast. Briefly let us mention a single example due to Maddieson (1984). The optimal five-vowel system is / i e a o u/ not /iː ẽ a̤ o̰ uˤ/. He suggested that a principle of "sufficient contrast," rather than of maximal contrast, may underlie such patterns.

Recent work (Lindblom et al., forthcoming) indicates that both vowel and consonant systems appear to be organized so as to meet a demand for "sufficient contrast." This becomes clear, once we begin to examine the contents of phonetic systems in relation to inventory size.

Our source of information is the UPSID database (Maddieson, 1984), which contains typological data on the segment inventories of 317 languages. Figure 2.4 exemplifies the results of sorting the consonant segments of UPSID into three categories—Basic, Elaborated, and Complex articulations—and then plotting the number of segments that a language uses in each category as a function of the total

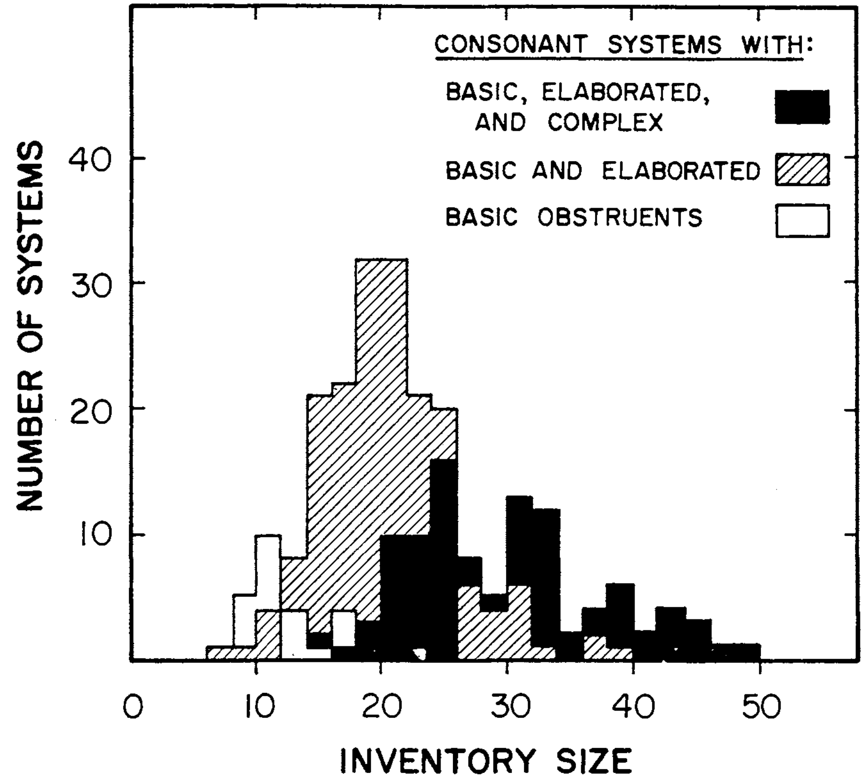


FIG. 2.4. Inventory size as a determinant of the contents of obstruent systems. Small inventories invoke Basic articulations, medium systems Basic and Elaborated segments and large inventories recruit Basic, Elaborated as well as Complex articulations. Data from the UPSID database (Maddieson, 1984).

number of consonants in that language[2]. Figure 2.4 shows a histogram plot describing the distribution of obstruents in the UPSID corpus. The diagram tells us that the contents of UPSID inventories is determined by inventory size. First, they invoke Basic articulations, then Basic and Elaborated, and ultimately all three types, including the Complex.

This Size Principle makes sense, if we assume that, in small systems, elemen-

[2] Elaborated articulations are place, source, and manner mechanisms that can be seen as elaborated versions of more elementary, or Basic, articulations. Segments containing combinations of Elaborated articulations are classified as Complex.
Basic: b, m, t, i, a . . .
Elaborated: p', ɗ, ɹ, mb, ʈ, q, p^j, . . .
Complex: qh, ãũ, q', ^{h}ʈ, . . .

tary articulations achieve sufficient contrast, whereas in larger systems, demands for greater intrasystemic distinctiveness cause additional dimensions (elaborations) to be recruited and combined to form complex segments. Data of this sort lend support to the Theory of Adaptive Dispersion (Lindblom & Maddieson, 1988; Lindblom et al., forthcoming) and suggest that the Size Principle, combined with quantitative measures of perceptual distinctiveness and articulatory complexity, should play a significant role in accounting for the contents of phonetic inventories.

The conclusions relevant to the present context are as follows: The results are compatible with claiming that inventories of phonetic gestures are selected so as to optimize both the distinctiveness and the pronounceability of individual segments. Phonetic gestures can, thus, be seen as *adaptations* to motoric and perceptual constraints that are language-independent and in no way special to speech. Facts about the way humans respond to psychophysical, nonspeech stimuli are sufficient to enable us to predict with good accuracy the essential contents of vowel inventories in a large number of languages. If human speech perception is a biologically specialized process that bypasses nonspeech hearing, why do vowel system patterns show such clear adaptations to auditory constraints not special to speech?

Conclusions

Plasticity and Invariance

Our interpretations are in agreement with the MT in that the distal object of speech perception is the speaker's intention. However, we differ by claiming that a speaker's intentions go beyond the production of phonetic gestures. We see the gestures as no more than a variable and adaptive means to the more global, ecologically more primary ends of speech acts: lexical access, comprehension, and social interaction. On this view, phonetic gestures are not strong candidates for the invariant units of speech. In fact, we argue that phonetic invariance is not necessary at all for adequate lexical access, because successful speech understanding presupposes gestures that are sufficiently contrastive but not necessarily physically constant.

Modularity and Phonological Adaptations

Assuming that speech perception is modular and operates by by-passing the general-purpose mechanisms of auditory perception, we face the question: Why are the fossilized gestures of phonological inventories so well adapted to biological properties of production and perception not special to speech? There appears to be a clear problem here for the MT.

Consider also the quantal and implicational nature of sound structure, which is the fact that languages tend to use similar gestures drawn from a very limited universal set and that the subsets they select show a strongly hierarchical organization internally. How does the MT account for such facts?

One possibility would be to suggest that all of these properties reflect the way that the "speech-processing module" works. We might assume that the module accepts only a limited number of gestures and that it somehow imposes an implicational structure on phonological systems. If so, we are led to ask: How did the speech-processing module get that way in the first place? It seems clear that if at an early stage of the game we claim that "speech is special," we shall *a priori* deprive ourselves of all opportunities to provide performance-based explanations of phonological facts. Consequently, we are forced to conclude that suggesting that the quantal and implicational organization of sound systems reflects the way that the speech-processing module works is a solution that completely begs the question on an issue that must be regarded as central to linguistic theory.

Speech Evolution

Admittedly, postulating biologically specialized systems for the production and perception of speech—as the MT does—appears not only reasonable but necessary in the light of a great deal of evidence. Claiming that linguistic perception does not in some sense presuppose specialized neural architecture would clearly be counter-factual. Why then have we pursued a line of reasoning that consistently sets out to deny the existence of such specializations? The answer is that denying the existence of specializations is not the expression of a belief or a conviction. It simply reflects a methodological strategy.

As we compare spoken language with the input and output structures underlying its use, we note that the motoric and perceptual mechanisms were in place long before language entered the stage. An initial task on the agenda of an evolutionary research program on spoken language would, therefore, seem to be to investigate how the newcomers, speech and language, could acquire some of their properties by adapting to the phylogenetically older structures rather than the other way around. The question would be: If language were seen as a set of adaptations to the constraints of early man's vocal, auditory, and cognitive systems, what would it be like?

The MT reverses this query completely, responding instead to: If speech production and speech perception were seen as adaptations to language what would they be like? Consider the following statements (Liberman & Mattingly 1985):

> Adaptations of the motor system for controlling the organs of the vocal tract took precedence in the evolution of speech. These adaptations made it possible, not only to produce phonetic gestures, but also to coarticulate them so that they could be

> produced rapidly. A perceiving system, specialized to take account of the complex acoustic consequences, developed concomitantly. (p. 7)

Perhaps Liberman and Mattingly were correct in saying that their theory "is neither logically meaningless nor biologically unthinkable" (Liberman & Mattingly 1985, p. 3). Once evolved, language could conceivably continue to develop in coevolution with the input/output mechanisms.

However, this approach has a methodological problem. How do we go about reconstructing the path of development towards specialization and uniqueness without running the risk of prejudging the issue? One possible answer—the one favored here—is that we can minimize this risk, if, in attempting to derive language from nonlanguage, we first make the most of the non-special mechanisms.

Acknowledgments

The author is indebted to Randy Diehl and Peter MacNeilage for helpful comments on this manuscript.

References

Bladon, R. A. W., & Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *Journal of the Acoustical Society of America, 69,* 1414–1422.

Crothers, J. (1978). Typology and universals of vowel systems. In J. H. Greenberg, C. A. Ferguson, & E. A. Moravcsik (Eds.), *Universals of human language* (vol. 2, pp. 99–152). Stanford, CA: Stanford University Press.

Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America, 83,* 1863–1875.

Fischer-Jørgensen, E. (1964). Sound duration and place of articulation. *Zeitschrift für Sprachwissenschaft und Kommunikationsforschung, 17,* 175–207.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America, 63,* 223–230.

Jakobson, R. (1968). *Child language, aphasia, and phonological universals.* The Hague: Mouton.

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237,* 1195–1197.

Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4,* 303–320.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant–vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs, 68,* 1–13.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6,* 172–187.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48,* 839–862.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35,* 1773–1781.

Lindblom, B. (1967). Vowel duration and a model of lip mandible coordination. *Quarterly Progress and Status Report, 4,* 1–29.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology* (pp. 13–44). Orlando, FL: Academic Press.

Lindblom, B. (1987). Absolute constancy and adaptive variability: Two themes in the quest for phonetic invariance. *Proceedings of the XIth International Congress of Phonetic Sciences* (vol. 3, pp. 1–18), Tallinn: Academy of Sciences of the Estonian S.S.R.

Lindblom, B. (in press). A model of phonetic variation and selection and the evolution of vowel systems. In W. S.-Y. Wang (Ed.), *Language transmission and change.* New York: Blackwell.

Lindblom, B., & Lubker, J. (1985). The speech homunculus and a problem of phonetic linguistics. In V. A. Fromkin (Ed.), *Phonetic linguistics* (pp. 169–192). Orlando, FL: Academic Press.

Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics, 7,* 147–161.

Lindblom, B., Lubker, J., Lyberg, B., Branderud, P., & Holmgren, K. (1987). The concept of target and speech timing. In R. Channon & L. Shockey (Eds.), *In honor of Ilse Lehiste* (pp. 161–182). Dordrecht, Holland: Foris.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (forthcoming). *Evolution of spoken language.* Orlando, FL: Academic Press.

Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. In L. M. Hyman & C. N. Li (Eds.), *Language, speech and mind* (pp. 62–78). New York: Routledge.

Lindblom, B., & Sundberg, J. (1971). Acoustical consequences of lip, tongue, jaw and larynx movement. *Journal of the Acoustical Society of America, 50,* 1166–1179.

Maddieson, I. (1984). *Patterns of sound.* Cambridge: Cambridge University Press.

Moon, S.-J., & Lindblom, B. (1989). Formant undershoot in clear and citation-form speech: A second progress report. *Quarterly Progress and Status Report, 1,* 121–123. Stockholm: Department of Speech Communication, RIT.

Nord, L. (1986). Acoustic studies of vowel reduction in Swedish. *Quarterly Progress and Status Report, 4,* 19–36.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39,* 151–168.

Perkell, J., & Klatt, D. (1986). *Invariance and variability in speech processes.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research, 29,* 434–446.

Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America, 85,* 295–312.

Stevens, K. N. (1968). Acoustic correlates of place of articulation for stop and fricative consonants. *Quarterly Progress Report, 89,* 199–205. Cambridge, MA: Research Laboratory of Electronics, MIT.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3–45.

Uchanski, R. M., Durlach, N. I., & Braida, L. D. (1987, November). *Clear speech.* Paper presented as part of a seminar on "Hearing-aid processed speech" at a meeting of the American Speech-Language-Hearing Association, New Orleans.

Comment: Beyond the Segment

Osamu Fujimura

The Division of Speech and Hearing Science, Ohio State University

Segmentalism

Lindblom gives an intriguing argument that the linguistic system for speech communication is largely determined by biological constraints that are not speech specific and by the requirement that the verbal communication be efficient. Such conditions can be shown to be satisfied by the sound patterns of existing natural languages, he argues, when we examine the observable characteristics of acoustic signals or their auditory perceptual values that represent the phonetic units of those languages. Phonetic units such as vowels and consonants form a system of segmental units in such a way that their auditory effects are maximally distinct under the peripheral perceptual constraints. Therefore, if I understand his point correctly, the reference to motor gestures, which Al Liberman and his colleagues' motor theory assumes as the basic principle of speech perception, should not necessarily be of primary concern for us; what we need is to assume that perception sorts distinct signals into categories of patterns segment by segment.

From my point of view, the segment inventory is only a part of language as a system of verbal codes. The human competence in verbal communication processes goes well beyond what a theory of phonemes can describe. What one has to depend on in perceiving speech messages is not limited to the capability of identifying isolated segmental units, even if that constitutes part of the actual process. As for the "segmental aspects" of speech, furthermore, a more interesting question would be to ask to what extent the concatenation-coarticulation approach would work, if we took syllables or demisyllables instead of the phonemic segments as the categories, because we know segment-by-segment identification simply does not work in the way the phonemic theory would prescribe. At best, such a system would involve very complex and abstract processes. As a concatenative unit, if we take the concept literally in analysing acoustic or auditory signals as they are, we will probably need to consider some rather large phonetic phrases, something like the stress group or foot in English (for some relevant observations about cognitive programming of motor execution, see Sternberg, Knoll, Monsell, & Wright, 1978, 1988) or an intonation phrase at some level (Pierrehumbert, 1980). This situation may be in some sense realistic, when we discuss initial exposures of infants to the surrounding language, where, for example, crucial parts of utterances, such as key words in focus, are marked with readily accessible prosodic cues, and those materials are drawn from a relatively limited vocabulary of simple words or phrases. Within such a prosodically coherent phrasal unit (including single words under certain environments), I think the organization is multidimensionally interwoven in the sense that there are no internal breakpoints in time that strictly synchronize phonetic events in different dimensions, such as voice pitch, movements of the mandible, tongue body, lips, velum, and so forth, and, depending on the descriptive scheme, temporal modulation (Fujimura, 1987b; Edwards & Beckman, 1988). To decompose such a phonetically coherent unit into constituent phonetic elements normally requires a complex procedure to map signals into abstract structures. A straightforward and transparent principle like concatenation and smoothing does not govern the phenomena under such circumstances.

I do not think any of the organizational issues such as temporal organization principles of speech, either within or among such phonetic phrases, can be accounted for by the biological/physical principles as we know them, except their rather peripheral constraints, such as declination and smoothing, pausing or decelerating for preparing the next phrase in the cognitive process, and, of course, some aspects of local characteristics of articulatory dynamics (see Browman & Goldstein, chapter 13 this volume). There are many linguistic structural choices that have to be made for producing or identifying specific structures. The choices are made lawfully within a very specifically selected framework, as we all know, but all we can say at present from a biological or physical point of view about the principles governing the rule systems of language seems to be that they are often (but not necessarily always) crucially specific to humans.

There are cases where sophisticated and careful consideration of necessary conditions can narrow down possibilities to a striking extent within a very limited domain. Lindblom's explanation of existing vowel systems may be cited as such a case, assuming that his demonstration holds for phonetic substances rather than the symbols used by linguists for transcribing different languages, as Mark Liberman aptly questioned in the conference. Local phonetic characteristics such as the "target values" of vowels are presumably the most likely aspects of speech phenomena that are significantly dictated by biological/physical constraints.

However, that there are such aspects of speech phenomena does not mean that that *is* the primary principle.

We should not be confused about the distinction between necessary conditions and sufficient conditions. It is obvious that both production and perception have to be under given physical and biological constraints as necessary conditions for any human activity. Given that Lindblom clearly admits that postulating specialized systems appears not only reasonable but also necessary, there is no disagreement about this. It may be, as he asserts, a matter of methodological strategy of research that he advocates a different characterization of the process of speech perception. However, I think it is also a matter of focus of interest (i.e. whether we are interested in the mental representation and a synchronic description of language or in providing ecological accounts of language evolution). I accept his assertion that general biological apparatus had developed before speech functions were needed. However, the highly evolved speech processing functions and mechanisms still can be special, because clearly there are needed functions other than swallowing and breathing, in order to utter and understand speech. That some phonetic capabilities are observed in animals does not lead us to the conclusion, as Lindblom appears to suggest, that biological principles commonly applicable to animals can account for phonetic capabilities. The respiratory mechanism, for example, is clearly a necessary component of speech production, and it does give some relevant constraints about what language must be like, but nobody argues that the characteristics of this biological mechanism are sufficient to account for the characteristics of speech.

I believe that it is important to distinguish the principles governing the real-time process of speech production and perception from those prescribing the evolution of linguistic systems. The process of evolution must respect a number of factors, and it is a slow (quasi-static) process that can balance out all different types of influences into an equilibrium of a synchronic system. The process is slow in reference to the time constant of developing and adapting the biological neural network. Speech production and perception are not such slow processes. It is a process of selecting elements of information to be conveyed according to a fixed and recognized framework of coding messages. Fixed elemental patterns can be organized into a seemingly variable component of the whole of an utterance. In my understanding, it is the issue of how listeners identify such organizational structures of utterances, rather than how elemental units evolve as the ingredients of phonetic forms, that the motor theory is or should be primarily concerned with.

Likewise, it is important to distinguish what a speaker or hearer can do under special circumstances from what he or she usually or typically does. Whatever a speaker makes use of as the program for generating specific utterances, it must be readjustable according to various situations and disturbances, and the perceptual system also must be able to conform to and recover from the variable effects of such disturbances and readjustments. Such readjustments may occur in part in anticipation of specific effects that are assessable by nonspeech measures, such as sensing the bite block via tactile, proprioceptive, and other means. However, in any case, such readjustability should not be taken as a proof that speaking strategies are not composed of fixed components of control programs. The entire configuration of control may well be formed out of fixed patterns involving adjustable parameters based on a certain framework of representation.

In order to understand the general characteristics of how speech is organized, uttered, and perceived, we need to identify speech organization principles that handle inherently multidimensional temporal structures. I think Sven Öhman was correct, when he proposed the consonantal perturbation theory (1967) as the result of encountering some difficulty in acoustic data interpretation using the concept of coarticulation. It is remarkable that he did it twenty years ago; the current nonlinear theory of phonology, its reference to articulatory organs (McCarthy, 1988), and many new observations in speech articulation, all are consistent with his insight. Obviously what he did was only a beginning. We now know much more about the abstract representation of sound patterns. The currently emerging multidimensional (multi-tier) theory of phonology may or may not succeed, and any model reflecting such representations will be inevitably complex. The mapping between abstract phonological representations onto observable speech phenomena must be rather opaque, in spite of the theory's direct reference to physiological apparatus.

I think pursuing a theory of perception as well as production referring to articulatory gestures is promising not only to understand what a speaker actually does and how signals are characterized accordingly but also given the independent linguistic justifications of phonological rules referring to such gestures. The classical concept of coarticulation as *the* principle of speech organization by concatenation is necessary but far from sufficient. There are more than assimilatory effects in phonological representations and the temporal organization of speech. As the first step, however, the question we need to ask may be what other principles we will have to introduce after generalizing the concept of coarticulation to inherently multidimensional representations, where smoothing works in different dimensions separately. In such a model, timing relations among elemental gestures in different articulatory dimensions seem to provide critical information (Fujimura, 1986).

Nonlinearity in the Sense of Superposition Principle

Suppose the mapping relation between the two representations of speech messages at the output of the production description (say, patterns of motor control) and the input to the perceptual system (say, auditory patterns) were described by a linear transformation in the sense of superpositionality. Then we would specify a message at the production output level by a set X of entities x_i in the form of a linear combination, and the same message can be represented as a linear combination based on a set Y of input patterns y_j. Each entity, for example y_j, could be an elemental time function, and overlapping signals could be decomposed into the constituent elements and their contributions to the given composite signal. Such a system would allow us to analyze many data of utterances by automatic statistical processing to derive a set of effective constituent elements empirically (Atal, 1983; van Dijk-Kappers & Marcus, 1988). Between such descriptive systems, based on X and Y for input and output, respectively (if they were available for speech descriptions), there would be no point in arguing which system is primary and which is secondary as Liberman and Mattingly (1985) aptly cautioned. The causal primacy of the production description is clearly valid, but that simply says that speech is physically produced and only then is heard.

It is the lack of linearity, in the sense of additivity or the applicability of the superposition principle, that makes motor theory a nontrivial theory. Either articulatory or auditory description may be formulated as a mostly linear process by choosing appropriate input–output levels and a framework of description. However, the mapping relation between the two levels representing the motor commands on the one hand and the auditory patterns on the other cannot be superpositional in any significant sense. This difficulty is there, whichever levels of representation we may choose for production and perception, as long as we take the production representation at a level reasonably transparently related to what we know as phonetic units, such as distinctive features, and the perceptual representation similarly transparent to phonetic units. Because the mapping is not linear, the usual and the most powerful reasoning method, first treating two factors of the problem separately and then combining them to predict the result for more general and complex situation, simply does not work. In my interpretation, the motor theory is an attempt to represent the auditory or some cognitive perceptual patterns of speech in terms of units in the motor-level representation of the message. The claim is in essence, if I am correct, that the information at this level is representable in such a way that a phonological representation has an approximately linear mapping into this motor-level representation. If this claim can be maintained, it seems at least to me that our study of speech organization can be reduced to components of tractable forms.

We wish to describe the principles of phonetic organization through a composition of effects of features. According to the classical theory (Jakobson, Fant, & Halle, 1963), each distinctive feature has its inherent phonetic target pattern, and each segment is represented by a simultaneous bundle of distinctive features. However, when we bundle up a set of distinctive feature values, each of which is evaluated under a certain condition of other feature values, all of a sudden the component values may change. A combination of elementary articulatory gestures does not necessarily result in a combination of acoustic or auditory characteristics, if the individual effects are evaluated separately under certain conditions that are not satisfied for the particular combination in question. Here we are talking about the mapping between phonological representations and articulatory or motor command representations. We do not know yet how we can describe the entire system based on invariant manifestations of elemental phonological units and simple and effective organizational principles to organize a phonetic material to be uttered.

My argument is that the system linearity is the critical issue for a successful endeavor in this area, and whether the system is linear or not depends crucially on the choice of the descriptive framework. Furthermore my conjecture, based on articulatory observations, is that certain organizational processes expressed in terms of articulatory or motor events are more nearly linearly related to the phonological description than most other descriptions (see Fujimura, 1987a). This would mean that a certain framework of articulatory representation with respect to temporal organization processes more or less allows a composition of complex cases out of a superposition of elemental mapping relations between a phonological representation and the articulatory characterization. More specifically, a correspondence between phonological features and elemental gestures must be representable by an approximately linear relation by appropriately choosing both the phonological and articulatory frameworks of representation. In the case of prosodic modulation, features representing phrase boundaries and stress/emphasis, and so on, or, in other words, configurational as well as prosodic specificatory features in the sense of Jakobson, Fant, and Halle, must be shown to have linear relations with, say, timing values of gesture organization control (cf. the gesture score in Browman & Goldstein, chapter 13 this volume, and Fujimura, 1987b).

The general framework of the phonological specifications for a message is inherently multidimensional, as theorists of nonlinear phonology advocate, and it is now becoming clear that specifications cannot be fully provided to be complete for each segment in terms of feature values. The phonetic implementation process has to consider both partial specifications of "segmental" features and "suprasegmental" features simultaneously. In such an opaque and complex system as the multidimensional representations with partial specifications and dimension-by-dimension implementation rules with linking constraints, the need for an explicit model of temporal organization is immense. Browman and Goldstein's work aims at such a model, working at the moment within a very limited local domain. A nonlinear model (in the sense of superposition) of this type eventually has to incorporate all factors that affect speech utterances under various circumstances, at least to first-order approximation, so that the setting is roughly correct for all the feature values, for the entire stretch of the phrasal unit that is approximately independent. It also must contain parameters that are sensitive to the context external to such a phrasal unit. Only then we can assume local superpositionality, so that we can reason the effects of different features involved, one by one, implicating their general phonetic significance.

I hope we will be able to achieve this stage of study soon. Our capability is probably not limited by the computational complexity. We need good insights into the problem, and the best insight, apart from intuition, can be obtained only through observations and interpretations of direct articulatory facts. Such observations will not automatically lead us to a good model; but they will delimit the domain of search, and we need such a guidance as much as possible, along with insights about other cognitive behaviors than speech production.

I do not believe we can understand speech and language only by statistical processing of observable data, nor by any elementary inference of automatic learning, based on a less than minimal descriptive framework that is inherent to language. I highly appreciate Lindblom's pointing out very interesting observations about some elements of language, but I also need more information about its inherent structure and organization.

References

Atal, B. S. (1983). Efficient coding of LPC parameters by temporal decomposition. *ICASSP Proceedings* (Vol. 2, 81–84). New York: IEEE Acoustics, Speech and Signal Processing Society.

Edwards, J., & Beckman, M. E. (1988). Articulatory timing and the prosodic interpretation of syllable duration. In O. Fujimura (Ed.), Articulatory organization—from phonology to speech signal [Special issue]. *Phonetica, 45,* 140–155.

Fujimura, O. (1986). Relative invariance of articulatory movements: An iceberg model. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 226–242). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fujimura, O. (1987a). Fundamentals and applications in speech production research. *Proceedings of the XIth International Congress of Phonetic Sciences* (Vol. 6, 10–27). Tallinn: Academy of Sciences of the Estonian S.S.R.

Fujimura, O. (1987b). A linear model of speech timing. In R. Channon & L. Shockey (Eds.), *In honor of Ilse Lehiste* (pp. 109–123). Dordrecht: Foris.

Jakobson, R., Fant, G., & Halle, M. (1963). *Preliminaries to speech analysis.* Cambridge, MA: MIT Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. In O. Fujimura (Ed.), Articulatory organization—from phonology to speech signal [Special issue]. *Phonetica, 45,* 84–108.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41,* 310–320.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Sternberg, S., Knoll, R. L., Monsell, S., & Wright, C. E. (1988). Motor programs and hierarchical organization in the control of rapid speech. In O. Fujimura (Ed.), Articulatory organization—from phonology to speech signal [Special issue]. *Phonetica, 45,* 175–197.

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 117–152). New York: Academic Press.

van Dijk-Kappers, A., & Marcus, S. (1988). Temporal decomposition in speech. Manuscript submitted for publication.