Editor's note: At the May 1987 Meeting of the Acoustical Society of America, there was a focus session on vowel perception, organized by Winifred Strange. The following four papers by W. Strange, T. M. Nearey, J. D. Miller, and W. Strange, and the Letter to the Editor in this issue by P. Ladefoged, are based on presentations and discussions during that session.

J. L. MILLER
Associate Editor

Evolving theories of vowel perception[a)]

Winifred Strange[b)]
University of South Florida, Tampa, Florida 33620

(Received 19 November 1987; accepted for publication 19 January 1989)

Research on the perception of vowels in the last several years has given rise to new conceptions of vowels as articulatory, acoustic, and perceptual events. Starting from a "simple" target model in which vowels were characterized articulatorily as static vocal tract shapes and acoustically as points in a first and second formant ($F1/F2$) vowel space, this paper briefly traces the evolution of vowel theory in the 1970s and 1980s in two directions. (1) Elaborated target models represent vowels as target zones in perceptual spaces whose dimensions are specified as *formant ratios*. These models have been developed primarily to account for perceivers' solution of the "speaker normalization" problem. (2) Dynamic specification models emphasize the importance of *formant trajectory* patterns in specifying vowel identity. These models deal primarily with the problem of "target undershoot" associated with the coarticulation of vowels with consonants in natural speech and with the issue of "vowel-inherent spectral change" or diphthongization of English vowels. Perceptual studies are summarized that motivate these theoretical developments.

PACS numbers: 43.71.An, 43.71.Es

INTRODUCTION

In the past 15 years, there has been renewed interest in some basic questions about the perception of vowels. These questions concern the characterization of *what* spoken vowels are as articulatory and acoustic events, and *how* listeners recover speakers' intended vowels from the acoustic signal. In this paper, two directions in which vowel theory has evolved over the last several years are described and some perceptual research that motivates current conceptions of the nature of vowels as dynamic articulatory and acoustic events is summarized.

In order to put the recent developments into perspective, I will begin by sketching the basic assumptions of a generic theory of vowel perception, which I refer to as a "simple target" model. Some version of this view was held implicitly or explicitly by most speech perception researchers in the early 1970s. It is also the "textbook" characterization of vowels and vowel perception that we still teach in introductory phonetics and speech science courses (e.g., Ladefoged, 1982; Pickett, 1980).

Central to this theory is the notion of the *vowel target* as a unifying concept among articulatory, acoustic, and perceptual characterizations of vowels. Vowel targets are conceived of as the canonical forms of vowels, the context-free stored representations of the phonological segments (Daniloff and Hammarberg, 1973). Articulatorily, these canonical targets are best represented by the static vocal tract shapes assumed when a speaker produces sustained, monophthongal vowel sounds. In continuous speech, these static articulatory positions are considered the goal states when coarticulating vowels in syllabic contexts (e.g., MacNeilage, 1970). Acoustically, vowel targets are represented as points in a multidimensional acoustic space whose coordinates are the first two ($F1/F2$) or three ($F1/F2/F3$) oral formants. Formant frequencies are derived from a single spectral cross section through the steady-state portion of the acoustic signal (Joos, 1948; Peterson, 1952, 1961).

According to a simple target model of perception, the target frequencies of the first two formants constitute the primary and often sufficient acoustic information for the perceptual identity of the vowel (e.g., Delattre *et al.*, 1952). Thus articulatory, acoustic, and perceptual descriptions of vowels are unified by this concept of a static, context-free target. Given this characterization of what vowels are in their canonical form, problems in explaining the perception of speakers' intended messages arise from two sources of variation in vowels as actually produced. First, as the classic work of Peterson and Barney (1952) showed, the target formant frequencies are not invariant with respect to intended (or perceived) vowels across men, women, and children, and different vowel categories often overlap in $F1/F2$ space even for speakers of the same age and gender. Thus, in a simple target model, formant frequency information must

[a)] This paper is based on an invited address of the same title presented at the Spring 1987 Meeting of the Acoustical Society of America [J. Acoust. Soc. Am. Suppl. 1 **81**, 516 (1987)].

[b)] Requests for reprints should be sent to the author at the published address.

 0001-4966/89/052081-07$00.80

somehow be adjusted to compensate for speaker variability. "Speaker normalization" has been considered one of the major problems in need of an explanation in any vowel perception theory[1] (see Miller, 1989, for a historical review of formant-ratio theories of normalization).

A second kind of variation in vowel targets is present even for vowels spoken by a single speaker and is often referred to as the "target undershoot" problem. Early studies by Lindblom (1963), Stevens and House (1963), and others showed that when vowels are coarticulated with consonants in consonant–vowel–consonant (CVC) syllables in continuous speech spoken at normal to rapid rates, the canonical acoustic (and articulatory) targets are often not reached. The degree of discrepancy from the target formant frequencies varies with phonetic context, stress, rate of speech, and even with individual differences in coarticulatory strategies (cf. Gay, 1978; Gay *et al.*, 1974; Kuehn and Moll, 1976). Thus formant maxima in vocalic nuclei bear a complex relationship to canonical vowel targets, and it is again argued that the listener somehow compensates for this variability to recover the intended vowel targets (Stevens and House, 1963; Lindblom and Studdert-Kennedy, 1967). Given these two sources of variation in the acoustic representation of vowels—speaker variation and contextual variation—we are left with a classical problem in speech perception, the lack of a simple correspondence between acoustic signal and perceived phonetic segment. An adequate perceptual theory must account for how perceivers compensate for this acoustic ambiguity in recovering the speaker's intended message.

Much of the research on vowel perception in the 1950s and 1960s focused on the speaker normalization problem. Having accepted the characterization of vowels as inherently ambiguous across speakers, researchers such as Ladefoged (1967), Lieberman *et al.* (1972), and others postulated normalization processes whereby listeners used information in ongoing speech to "calibrate" for a particular speaker's acoustic vowel space.[2] A well-known study by Ladefoged and Broadbent (1957) provided empirical support for this hypothesis. Synthetic stimuli with ambiguous target formant frequencies were shown to be identified in relation to the vowel space specified by the formant structure of a preceding carrier sentence (see also, Dechovitz, 1977; Nearey, 1989).

Fewer early research efforts were directed toward the target undershoot problem. In another classic study, Lindblom and Studdert-Kennedy (1967) hypothesized that listeners interpreted formant maxima of coarticulated vowels in relation to the direction and rate of formant transitions. They presented synthetically generated CVC syllables in which formant maxima varied along a continuum and demonstrated that perceptual boundaries shifted as a function of the syllable duration and direction of $F2$ transition; that is, listeners demonstrated "perceptual overshoot" in identifying dynamic formant patterns that compensated for productive undershoot (see also Williams, 1987, and Nearey, 1989, for further discussion and data).

Quite aside from these early perceptual studies of the normalization and target undershoot problems, another body of work published in the early 1960s, by Peterson and Lehiste (1960), reported important, though largely ignored, descriptive information about the acoustic specification of vowels. Analysis of a large corpus of CVC syllables spoken in sentence contexts revealed that there were systematic differences in the time-varying patterns of coarticulated American English vowels. First, vowels were differentiated in terms of intrinsic duration. They also reported systematic differences across vowels in the shape of formant trajectories, specifically in the relative durations of onglides (formant transitions into the syllable nucleus), offglides (formant transitions out of the syllable nucleus) and quasi-steady-state portions of CVC syllables. These differences in formant trajectories distinguished so-called tense and lax vowels. Syllables containing tense vowels have relatively rapid onglides and offglides that are of approximately equal duration. Thus formant trajectories (especially the trajectory of the first formant $F1$) form a symmetric pattern. Syllables containing lax vowels had slightly less rapid onglides but much slower offglides than for tense vowels, and much shorter quasi-steady-state portions. This resulted in asymmetric $F1$ trajectories. Thus independent of differences in overall duration between short, lax vowels and long, tense vowels, there are also differences in formant pattern that differentiate these vowel categories (see Strange, 1989, for an illustration of "tense" and "lax" formant trajectories).

These systematic differences in dynamic acoustic patterns are of potential importance perceptually, because they differentiate vowels with similar articulatory and acoustic targets; that is, vowels with formant maxima that overlap in $F1/F2$ acoustic space when spoken by different speakers, at different rates, and in different phonetic contexts, may nevertheless be distinguished with respect to intrinsic duration and trajectory shape. However, until quite recently, studies of vowel perception have, for the most part, not included these dynamic sources of information. This is clearly evident in studies using synthetically generated stimuli in which formant trajectory shapes and syllable durations are not varied. A number of early experiments (cf. Ainsworth, 1972; Bennett, 1968) demonstrated that, when present, intrinsic duration was an important perceptual cue to vowel identity. However, to my knowledge, no perceptual studies had investigated the relevance of trajectory shapes for the identification of vowels when we began our own research in the 1970s.

Research on vowel perception in the 1970s and 1980s has progressed along two theoretical lines. I will refer to the two approaches as "elaborated target" models and "dynamic specification" models (using the term model in its most general sense of a theoretical framework). Both efforts constitute a return to basic questions about how vowels are best characterized acoustically and perceptually; specifically, both positions challenge the assumption of the simple target model that vowels are acoustically ambiguous. The two approaches differ, however, in that the first focuses, as did earlier models, on the perceptual characterization of vowels as static points in a multidimensional space, while the other emphasizes the role of dynamic sources of information as perceptually critical. They also differ in the emphasis they place on the speaker normalization and target undershoot problems.

I. ELABORATED TARGET APPROACHES TO VOWEL PERCEPTION

Elaborated target models have been formulated primarily to account for speaker normalization. That is, a main concern is to try to characterize vowels produced by speakers of different gender and age in such a way that variations among tokens of the same intended and perceived vowel are minimized and differences between perceptually distinct vowels are maximized.[3] The recent work of Syrdal and Gopal serves as an elegant example of this type of effort (Syrdal, 1985; Gopal and Syrdal, 1984; Syrdal and Gopal, 1986). In Syrdal's model, fundamental and formant frequencies are transformed to the psychophysically motivated critical band or Bark scale; then vowels are arrayed in a multidimensional "auditory space" defined by the coordinates $F1$–$F0$, $F2$–$F1$, and $F3$–$F2$. Finally, vowels are classified into two categories along each dimension according to whether or not they exceed a three-Bark critical difference (Syrdal and Gopal, 1986). Using the Peterson and Barney (1952) multispeaker corpus, this representation was shown to reduce the overlap among tokens of different vowel types produced by different speakers. The dimensions also corresponded quite well to traditional phonetic features that describe articulatory target shapes in terms of tongue, lip, and jaw position.

Other elaborated target models have proposed different algorithms for the transformation of acoustic data (Gerstman, 1968; Skinner, 1977) and recently Hillenbrand and Gayvert (1987) have compared a variety of transformations. While these models differ considerably in detail, they have one characteristic in common. The acoustic raw data for the auditory representation of vowels are taken from a single spectral cross section of the acoustic syllable. That is, all these attempts presuppose that a static spectral configuration captures the essential acoustic input for vowel identity, at least for so-called monophthongal vowels. This poses a "segmentation" problem in the case of coarticulated vowels, because there may be no steady-state portion within the syllable that can easily be identified as most representative of the vowel target.

In the case of Syrdal's 1986 model, formant data were obtained from the Peterson and Barney corpus, in which vowels were produced in /h/–vowel–/d/ citation-form syllables. Stevens and House (1963) showed that vowels produced in this context contain steady-state vocalic nuclei in which formant frequencies correspond closely to those of sustained vowels spoken in isolation; i.e., there is little effect of coarticulation. Thus the problem of determining which spectral cross section is most representative of the vowel target is minimal for vowels spoken in this context, and target undershoot due to coarticulation is nearly nonexistent. An important question about the generality of transformed auditory representations such as the one proposed by Syrdal, then, concerns their adequacy in characterizing vowels coarticulated in different consonantal contexts and at different rates (see Gopal and Syrdal, 1984). At some point, any model that relies on single spectral sections must make explicit the decision criteria for choosing which spectral cross section of the acoustic syllable is to serve as input.

The article by Miller (1989) presents an auditory-perceptual model of phonetic recognition in which speech signals are characterized as paths through a three-dimensional auditory-perceptual space defined in terms of transformations of formant ratios and a "sensory reference" based on the speaker's average fundamental frequency. According to this theory, simple (undiphthongized) vowels are identified when the "perceptual pointer" performs a "segmentation maneuver" within one of the large and irregularly shaped "perceptual target zones" of that space. After a historical review of formant-ratio conceptions of vowels, Miller elaborates the theory and presents data on vowel recognition performance by the model, including recognition of vowels coarticulated in CV syllables. In discussing the performance of the model on coarticulated speech samples, the criteria by which segmentation maneuvers are performed are discussed, as well as the problems associated with diphthongization, rhotacization, and nasalization of vowels. This, then, constitutes an ambitious attempt to elaborate a target model of vowel perception that takes into account both speaker normalization and coarticulatory effects.

II. DYNAMIC SPECIFICATION APPROACHES TO VOWEL PERCEPTION

A second theoretical approach to vowel perception (referred to as the dynamic specification approach), which has evolved over the last several years, stems primarily from research on the perception of coarticulated vowels. Thus the emphasis has been on accounting for how perceivers identify intended vowels in the face of acoustic variability associated with contextual effects (the target undershoot problem) rather than on the problems associated with speaker variability. Renewed interest in this problem came, in part, from some unexpected results of studies of vowels spoken in various stop consonant contexts by multiple speakers (Verbrugge *et al.*, 1976; Strange *et al.*, 1976; see, also, Strange *et al.*, 1979; Gottfried and Strange, 1980).[4] These studies found that coarticulated vowels were identified with surprising accuracy, even by phonetically unsophisticated listeners, despite considerable acoustic ambiguity in formant frequencies measured at the point of closest approximation to the target.[5] In fact, vowels in CVC contexts were identified more accurately than isolated vowels produced by the same panel of speakers even though formant maxima were less distinct in $F1/F2$ vowel space than were the isolated vowel targets. Furthermore, perceptual confusions among coarticulated vowels, when they occurred, were not highly predictable on the basis of similarity of formant frequency maxima.

A number of researchers subsequently reported failures to replicate the finding that coarticulated vowels were identified more accurately than isolated vowels (see Nearey, 1989, for a review of some of the data). Studies by Macchi (1980), Diehl *et al.* (1981), Assmann *et al.* (1982), and Rakerd *et al.* (1984) showed that, under different stimulus and task conditions, both isolated vowels and coarticulated vowels were perceived extremely accurately, even when speakers varied unpredictably from trial to trial (but see Strange and Gottfried, 1980; Gottfried *et al.*, 1985). When dialect variation among speakers and listeners was restricted or when more

sophisticated listeners were tested, performance approached ceiling levels for both coarticulated and uncoarticulated vowels (see, also, Carney *et al.*, 1983). However, it is important to note that in no study using natural (as opposed to synthetic) stimuli were isolated vowels identified *more* accurately than coarticulated vowels, as would be predicted if static targets were the primary source of information for vowel identity.

On the basis of these findings, my colleagues and I hypothesized that there were additional sources of information available in the time-varying acoustic signal, which perceivers utilize in identifying the speakers' intended vowels. Research efforts in our laboratory and elsewhere in the past several years have been directed toward an exploration of the nature of dynamic information available in coarticulated syllables. In order to gain more control over stimulus parameters, while still working with naturally produced coarticulated vowels, a methodology was adopted in which digitized waveforms of spoken CVC syllables were modified by deleting portions of the signal and changing the temporal relationship among remaining portions (see Jenkins *et al.*, 1983; Strange *et al.*, 1983; Strange, 1989, for a detailed description of the procedures). Using this technique, the perceptual efficacy of three types of information has been explored.

First, there is information provided in the syllable nucleus, i.e., the central portion of a CVC syllable in which formants are quasi-steady state or change at a relatively slow rate. These nuclei, which contain formant maxima, are thought to provide the best information about canonical targets; however, it must be noted that, for American and Canadian English vowels, there is systematic movement within the nucleus even of supposedly monophthongal vowels, which may be important perceptually (see Nearey, 1989; Strange, 1989, for further discussion of the role of diphthongization or "vowel-inherent spectral change" in perception). A second source of information is provided by (relative) duration differences intrinsic to the vowel, which, in combination with other factors such as final consonant voicing, stress, and speaking rate, determine the overall syllabic duration. A final hypothesized source of information is provided by the formant transitions into and out of the syllable nucleus. These transitional portions, taken together, may best reflect systematic differences in opening and closing characteristics of vowel gestures, which determine parameters of trajectory shape. Note that none of these sources of information is adequately captured in any single spectral cross section through the acoustic syllable, but rather must be specified as a change over time in spectral structure.

Modified stimulus conditions in which various combinations of these three types of information were available included: (1) silent-center (SC) syllables, in which the entire syllable nucleus was attenuated to silence, leaving the initial and final transitional portions in their original temporal relationship; (2) neutral-duration, silent-center (NDSC) syllables, in which intrinsic duration information was removed by equating the duration of silent intervals between initial and final transitional portions; (3) centers in which (variable duration) vocalic nuclei were presented with initial and final transition removed; (4) neutral-duration centers (NDCs), in which vocalic nuclei were all trimmed to the length of the shortest original nucleus. Control conditions included: (5) unmodified syllables, (6) an "initials only" condition, and (7) a "finals only" condition. The latter two were included to rule out the possibility that vowel identity was preserved in either of the transitional portions of the SC and NDSC syllables alone.

In general, the results of these studies showed a consistent pattern of identification errors. Intended vowels were most accurately identified in SC syllables containing both trajectory shape and duration information. In some studies, errors in this condition were somewhat greater than for unmodified control syllables; in others they were not, but, always, this condition was the best of the modified conditions. NDSC syllables, in which intrinsic duration differences were neutralized, yielded somewhat greater errors for syllables containing lax vowels, but overall identification remained quite accurate, relative to other modified conditions. Center stimuli, which contained target information, vowel-inherent spectral change, and relative duration information, were also identified as well as or nearly as well as unmodified controls. When (most) spectral-change and duration information were removed from syllable nuclei, as in NDC conditions, error rates increased significantly. Finally, performance in the initials only and finals only conditions was extremely poor; thus accurate vowel identification in the SC and NDSC conditions could not be attributed to information present in either transitional component alone (see Strange, 1989, for a more detailed discussion of these results).

The perceptual robustness of SC syllables has been corroborated in a parametric study by Parker and Diehl (1984) in which intrinsic duration information was factored out by constructing separate tests of intrinsically short and intrinsically long vowels. Stimuli were /d/–vowel–/d/ syllables spoken by four different speakers in which 60%, 70%, 80%, or 90% of the total acoustic syllable was removed from the center and replaced with either silence or aperiodic noise. Vowel identification remained well above chance even for the 90% SC condition. Less radical excisions of center portions yielded error rates of about 5% and 15% for long and short vowels, respectively. The noise versus silence manipulation produced no significant differences except for the 90% excised stimuli.

Nearey and Assmann (1986) extended this paradigm to tests of uncoarticulated Canadian English vowels spoken in citation form. They hypothesized that both SC syllables and center stimuli were identified accurately because they contained information about vowel-inherent spectral change. Their results showed that listeners could identify vowels on the basis of two 30-ms portions taken from early and late in the syllable only if the two portions were presented in the appropriate sequence. (These data are reviewed more extensively in Nearey, 1989, and Strange, 1989.)

To examine whether vowels in SC syllables are identified on the basis of dynamic trajectory information or on the basis of interpolated "acoustic targets," Verbrugge and Rakerd (1986) presented SC syllables in which initial and final transitions from the same (citation-form) syllable produced by different speakers (one man and one woman) were re-

combined. They found that identification of vowels in these hybrid SC syllables was no worse than for SC syllables of a single speaker, and significantly better than when either initial transitions or final transitions were presented alone (see, also, Rakerd and Verbrugge, 1987). They concluded that "information...carried relationally (by initial and final transitions of CVC syllables)...is complementary to, and distinct from, formant frequency information present in a syllable's center" (Verbrugge and Rakerd, 1986, p. 39) and that this relational information is speaker independent. This result has recently been replicated in our laboratory with syllables spoken rapidly in a carrier sentence (Jenkins and Strange, 1987).

The results of the preceding studies with modified natural speech strongly suggest that perceptually relevant dynamic information is specified over the initial and final transitional portions of CVC syllables. Furthermore, they suggest that this information is relational in nature; that is, transitional portions define distinctive trajectory shapes in which such parameters as the ratio of transition to quasi-steady-state duration or the symmetry versus asymmetry of onglides and offglides distinguish vowel gestures. Some recent research at MIT using synthetic speech materials sheds further light on the role of trajectory shape in vowel perception.

Huang (1985, 1986) synthesized several series of /d/–vowel–/s/ syllables in which syllable duration and temporal structure of the formant trajectories (symmetrical "tense" versus asymmetrical "lax" temporal trajectory) were independently varied in $F1/F2$ series contrasting both tense–lax and lax–lax vowel pairs. Identification boundaries along the $F1/F2$ frequency continuum shifted significantly for both tense–lax and lax–lax contrasts as a function of both temporal trajectory and syllable duration.

Di Benedetto (1989) extended this paradigm to a study of the role of temporal parameters of $F1$ trajectory in the perception of vowel height and tenseness in synthetically generated patterns of /d/–vowel–/d/ syllables. Tests with relatively sophisticated listeners showed systematic differences in identification of both vowel height and vowel tenseness as a function of the relative durations of $F1$ onglides and offglides. Di Benedetto concluded that vowel identity is dependent upon temporal parameters defined over the entire $F1$ trajectory, that is, on *when* formant maxima are reached, proportional to the overall duration of the acoustic syllable (see Strange, 1989, for a more detailed description of the studies by Huang and Di Benedetto). Further studies of this kind using synthetic stimuli that manipulate dynamic variables will provide important insights into the precise nature of the information provided by these time-varying acoustic patterns.

In summary, then, research from several laboratories, which has been motivated by or interpreted within a dynamic specification framework, points to the critical importance of time-varying acoustic parameters for the accurate identification of coarticulated vowels. This has led to a characterization of vowels as dynamic articulatory and acoustic events. According to this view, American English vowels must be described articulatorily as *gestures* having intrinsic timing characteristics (Fowler, 1980). Canonical vowels are thus defined not only by the target state to be reached by the articulators, but also by their characteristic opening and closing phases, which are coordinated such that both the duration and rate of movement of the articulators are appropriate for the intended vowel. The acoustic consequences of these gestures, when coproduced with consonants, are continuously varying formant patterns. No single spectral cross section adequately captures all the perceptually relevant information; rather, the acoustic information for vowel identity resides in the *changing* spectral structure.

Given this articulatory and acoustic characterization of vowels, the nature of the problems to be explained by a perceptual theory changes. If, as the perceptual results suggest, there is sufficient information within single syllables to allow the listener to identify intended vowels, even when those vowels are coarticulated by different speakers in different consonantal contexts, then the need to postulate psychological processes by which the perceiver enriches otherwise ambiguous sensory input is eliminated. Instead, the goal becomes the development of perceptual models that describe how perceivers detect and utilize the available sources of information for vowel identity during the perception of ongoing speech.

III. CONCLUDING REMARKS

It is perhaps time to merge the efforts of researchers working within the two broad theoretical frameworks outlined in this paper. The Focus Session on Vowel Perception at the Spring 1987 Meeting of the Acoustical Society of America, upon which the papers of this series are based, was an attempt to further that process. It is encouraging to see evidence of such integration already taking place. The second article of this series by Nearey (1989) reviews theory and research dealing both with issues of speaker normalization and with the role of dynamic information in the specification of vowels. He advocates a theoretically eclectic approach, which utilizes the powerful tools of quantitative pattern recognition models in the interpretation of perceptual results. Investigators working on elaborated target models have begun to ask how normalization algorithms can be modified to take into consideration effects of coarticulation and speaking rate (Gopal and Syrdal, 1984). The auditory-perceptual model, described in the third article of this series (Miller, 1989), could, in principle, incorporate dynamic spectral information carried in formant trajectories in his formulation of a mass–spring model of the "perceptual pointer." He deals explicitly with the problem of diphthongization of English vowels. In turn, researchers within the dynamic specification framework have begun to consider issues of speaker variability when they attempt to describe dynamic acoustic parameters of coarticulated vowels (Verbrugge and Rakerd, 1986). The article by Strange (1989) makes a start in this direction by presenting a replication and extension of earlier results with new speakers.

While the research reviewed here and reported in the following papers offers new insights into the nature of vowels and how they are perceived, the work is far from finished.

Clearly, future efforts will benefit from a consideration of issues arising from both theoretical approaches. Quantitative modeling techniques, such as the ones developed by Nearey and others, will be helpful in describing the nature of the information for vowels and will provide a more precise way to relate acoustic and perceptual data. Finally, it is now time to push forward with analytic perceptual studies using carefully constructed synthetic materials to further our understanding of the exact nature of the dynamic acoustic information for vowels. Such methods may also provide more powerful ways to explore how that information is utilized by perceivers in recovering the speaker's intended message.

ACKNOWLEDGMENTS

I wish to thank James J. Jenkins for his many helpful comments on the manuscript and for his continuing collaboration in every aspect of the research reported here. This research has been supported by grants from NIMH (NH-37924) and from NINCDS (NS-22568).

[1]The term "normalization" is used in two different senses in discussions of the acoustic and perceptual representation of vowels. In the engineering sense, it refers to methods by which acoustic parameters can be transformed and represented to "factor out systematic, but phonetically nondistinctive, covariation in signal properties..." (Nearey, 1989). In this sense, no claims about the psychological reality of such transformations are made. In psychological models, the term normalization has been used to refer to physiological and/or mental processes by which incoming sensory data are recoded and "interpreted" in reference to internal representations of phonetic categories. It is possible that several normalization algorithms could be found that served equally well to differentiate tokens of different vowel types (see Hillenbrand and Gayvert, 1987). The problem for speech perception theorists, then, becomes one of determining which (if any) of these normalization routines can be shown to have psychological reality. Experiments showing a strong correlation between outcomes of normalization routines and human perceptual studies are only a first step in this process. Stronger claims for the psychological reality of such procedures must await evidence from "true" experimental designs in which acoustic input is manipulated and perceptual consequences are assessed with respect to proposed models.

[2]These theorists thus postulated psychological normalization processes that used information extrinsic to the test syllable itself. In the terminology of Nearey (1989), this constitutes a theory of extrinsic or transsegmental specification. Thus this type of model can be referred to as an extrinsic normalization model.

[3]Such models can be referred to as intrinsic normalization theories in that they attempt to specify invariant information for vowel categories on the basis of acoustic parameters within the target syllables. Such models can refer to normalization in either the engineering sense or in the psychological sense.

[4]I would like to acknowledge the enormous contribution to this work from the many brilliant colleagues I have had the pleasure of collaborating with over the last several years: Donald Shankweiler, Robert Verbrugge, Brad Rakerd, James Jenkins, Thomas Edman, Thomas Johnson, Terry Gottfried, and Arlene Carney. I am also indebted to Randy Diehl and to Michael Studdert-Kennedy, Carol Fowler, Alvin Liberman, and other colleagues at Haskins Laboratories for many stimulating and challenging discussions about vowel perception.

[5]As Nearey (1989) points out, the absolute error rates in our early studies were considerably higher than for those reported in subsequent studies from our own laboratory and those of other researchers. In these initial studies, productions were obtained from a set of speakers with considerable dialect variation within American English. We did not preselect tokens on the basis of phonetic "goodness," and we urged subjects to speak rapidly and "naturally"; that is, we discouraged "careful" pronunciation of the stimuli. This was done so that our corpus would represent the variability in speech one normally encounters in an environment where speakers come from diverse regional areas of the United States.

Ainsworth, W. A. (**1972**). "Duration as a cue in the recognition of synthetic vowels," J. Acoust. Soc. Am. **51**, 648–651.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**, 975–989.

Bennett, D. C. (**1968**). "Spectral form and duration as cues in the recognition of English and German vowels," Lang. Speech **11**, 65–85.

Carney, A. E., Edman, T. R., Strange, W., and Jenkins, J. J. (**1983**). "Advantage of speaker as listener in a vowel identification study," J. Acoust. Soc. Am. **73**, 2222–2223.

Daniloff, R. G., and Hammarberg, R. E. (**1973**). "On defining coarticulation," J. Phon. **1**, 239–248.

Dechovitz, D. (**1977**). "Information conveyed by vowels: A confirmation," Stat. Rep. Speech Res. **SR 51/52**, 213–219 (Haskins Laboratories, New Haven, CT).

Delattre, P. C., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (**1952**). "An experimental study of the acoustic determinants of vowel colour: Observations on one- and two-formant vowels synthesized from spectrographic patterns," Word **8**, 195–210.

Di Benedetto, M.-G. (**1989**). "Frequency and time variations of the first formant: Properties relevant to the perception of vowel height," J. Acoust. Soc. Am. (in press).

Diehl, R. L., McCuster, S. B., and Chapman, L. S. (**1981**). "Perceiving vowels in isolation and in consonantal context," J. Acoust. Soc. Am. **68**, 239-248.

Fowler, C. A. (**1980**). "Coarticulation and theories of extrinsic timing," J. Phon. **8**, 113–133.

Gay, T. (**1978**). "Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am. **63**, 223–230.

Gay, T., Ushijima, T., Hirose, H., and Cooper, F. S. (**1974**). "Effect of speaking rate on labial consonant-vowel articulation," J. Phon. **2**, 47–63.

Gerstman, L. J. (**1968**). "Classification of self-normalized vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 78–80.

Gopal, H. S., and Syrdal, A. K. (**1984**). "Some effects of speaking rate on spectral and temporal characteristics of American English vowels," J. Acoust. Soc. Am. Suppl. 1 **76**, S17.

Gottfried, T. L., Jenkins, J. J., and Strange, W. (**1985**). "Categorial discrimination of vowels produced in syllable context and in isolation," Bull. Psychon. Soc. **23**, 101–104.

Gottfried, T. L., and Strange, W. (**1980**). "Identification of coarticulated vowels," J. Acoust. Soc. Am. **68**, 1626–1635.

Hillenbrand, J., and Gayvert, F. T. (**1987**). "Speaker-independent vowel classification based on fundamental frequency and formant frequencies," J. Acoust. Soc. Am. Suppl. 1 **81**, S93.

Huang, C. B. (**1985**). "Perceptual correlates of the tense/lax distinction in general American English," unpublished masters thesis, MIT, Cambridge, MA.

Huang, C. B. (**1986**). "The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels," IEEE-ICASSP **86**, 893–896, Tokyo, Japan.

Jenkins, J. J., and Strange, W. (**1987**). "Identification of 'hybrid' vowels in sentence context," J. Acoust. Soc. Am. Suppl. 1 **82**, S82.

Jenkins, J. J., Strange, W., and Edman, T. R. (**1983**). "Identification of vowels in 'vowelless' syllables," Percept. Psychophys. **34**, 441–450.

Joos, M. A. (**1948**). "Acoustic phonetics," Language **24** (Suppl.), 1-136.

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

Ladefoged, P. (**1967**). *Three Areas of Experimental Phonetics* (Oxford U. P., New York).

Ladefoged, P. (**1982**). *A Course in Phonetics* (Harcourt Brace Javonovich, New York).

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Lieberman, P., Crelin, E., and Klatt, D. (**1972**). "Phonetic ability and the related anatomy of the newborn, adult human, Neanderthal man and the chimpanzee," Am. Anthropol. **74**, 287–307.

Lindblom, B. E. F. (**1963**). "Spectrographic study of vowel reduction," J.

Acoust. Soc. Am. **35**, 1773–1781.
Lindblom, B. E. F., and Studdert-Kennedy, M. (**1967**). "On the role of formant transitions in vowel recognition," J. Acoust. Soc. Am. **42**, 830–843.
Macchi, M. J. (**1980**). "Identification of vowels spoken in isolation versus vowels spoken in consonantal context," J. Acoust. Soc. Am. **68**, 1636–1642.
MacNeilage, P. (**1970**). "Motor control of serial ordering of speech," Psychol. Rev. **77**, 182–196.
Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**, 2114–2134.
Nearey, T. M. (**1989**). "Static, dynamic, and relational factors in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.
Nearey, T. M., and Assmann, P. F. (**1986**). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.
Parker, E. M., and Diehl, R. L. (**1984**). "Identifying vowels in CVC syllables: Effects of inserting silence and noise," Percept. Psychophys. **36**, 369–380.
Peterson, G. E. (**1952**). "The information-bearing elements of speech," J. Acoust. Soc. Am. **24**, 629–637.
Peterson, G. E. (**1961**). "Parameters of vowel quality," J. Speech Hear. Res. **4**, 10–29.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.
Peterson, G. E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **30**, 693–703.
Pickett, J. M. (**1980**). *The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception* (University Park, Baltimore, MD).
Rakerd, B., and Verbrugge, R. R. (**1987**). "Evidence that the dynamic information for vowels is talker independent in form," J. Memory Lang. **26**, 558–563.
Rakerd, B., Verbrugge, R. R., and Shankweiler, D. P. (**1984**). "Monitoring for vowels in isolation and in a consonantal context," J. Acoust. Soc. Am. **76**, 27–31.
Skinner, T. E. (**1977**). "Speaker invariant characterizations of vowels, liquids, and glides using relative formant frequencies," J. Acoust. Soc. Am. Suppl. 1 **62**, S5.
Stevens, K. N., and House, A. S. (**1963**). "Perturbation of vowel articulations by consonantal context: An acoustical study," J. Speech Hear. Res. **6**, 111–128.
Strange, W. (**1989**). "Dynamic specification of coarticulated vowels spoken in sentence context," J. Acoust. Soc. Am. **85**, 2135–2153.
Strange, W., Edman, T. R., and Jenkins, J. J. (**1979**). "Acoustic and phonological factors in vowel perception," J. Exp. Psychol.: Hum. Percept. Perform. **5**, 643–656.
Strange, W., and Gottfried, T. L. (**1980**). "Task variables in the study of vowel perception," J. Acoust. Soc. Am. **68**, 1622–1625.
Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.
Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). "Consonantal environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–224.
Syrdal, A. K. (**1985**). "Aspects of a model of the auditory representation of American English vowels," Speech Commun. **4**, 121–135.
Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.
Verbrugge, R. R., and Rakerd, B. (**1986**). "Evidence of talker-independent information for vowels," Lang. Speech **29** (1), 39–57.
Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). "What information enables a listener to map a talker's vowel space?" J. Acoust. Soc. Am. **60**, 198–212.
Williams, D. R. (**1987**). "Judgments of coarticulated vowels are based on dynamic information," J. Acoust. Soc. Am. Suppl. 1 **81**, S17.