

# Virtues and perils of an empiricist approach to speech perception

Keith R. Kluender<sup>a)</sup>

*Department of Psychology, University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706*

Andrew J. Lotto

*Department of Psychology, Parnly Hearing Institute, Loyola University Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626*

(Received 2 January 1997; revised 20 August 1998; accepted 6 October 1998)

Nearey's "double-weak" approach [J. Acoust. Soc. Am. **102**, 3241–3254 (1997)] advocates a pattern-recognition model in answer to fundamental problems of speech perception. Both theoretically and empirically, there is much to recommend double-weak. However, there is some question whether this approach avoids long-standing disagreement with respect to the objects of speech perception. In addition, the descriptive power of even relatively simple empiricist models such as Nearey's can mislead with respect to fundamental articulatory and auditory processes underlying speech perception. Here, the positive contributions are celebrated, and several cautionary observations—germane to broader questions of experience and learning—are offered. © 1999 Acoustical Society of America. [S0001-4966(99)04601-9]

PACS numbers: 43.71.An, 43.71.Es [WS]

## INTRODUCTION

Nearey (1997) presents an approach to understanding phonetics and speech perception, which he labels "double-weak." This approach, in some ways new and in some ways familiar, provides the motivation for this commentary. Nearey suggests the name "double-weak" because, in contrast to some previous theorization concerning speech perception, his approach assumes relatively weak constraints on both articulatory and auditory processes. In place of substantial demands upon articulation or audition, Nearey argues that relatively simple pattern-recognition processes may be adequate to yield phonological units from patterns in the acoustic signal. He suggests, however, that talkers do control their productions in a fashion that permits these pattern-recognition processes to be successful. Nearey simulates simple pattern recognition using logistic regression analysis, which shares characteristics of several other modeling approaches, and he demonstrates the efficacy of this approach with old and new data from perception studies. Each of these major points will be considered in what follows. A number of positive contributions will be celebrated, and several cautionary observations will be offered.

## I. OBJECTS OF PERCEPTION

Nearey begins with some reflection upon the tension among theories with respect to whether objects of speech perception are articulatory gestures or acoustic patterns typically produced by vocal tracts. Describing this division as one of mere "tension" may be an understatement. Together with issues concerning whether perception of speech requires special mechanisms (specific to humans and separate from

processes responsible for perception of other events or complex sounds), polemics concerning the primacy of gestures versus acoustics have been longstanding.

Over the past 40 years and across evolving versions of motor theory, perception of gestures lies at the heart of the matter. Liberman and Mattingly asserted that "the objects of speech perception are the intended phonetic gestures of the speaker" (1985, p. 1). In recent versions of motor theory (e.g., Liberman and Mattingly, 1985, 1989), this specialization takes the form of a "phonetic module." Although this hypothesized module shares duties of production and perception, the emphasis has been upon how such a module is exquisitely adapted for recovering the intended gesture when its existence is considerably obscured as portrayed in the acoustic signal. The force of this gestural approach<sup>1</sup> has been to accept articulation and its seemingly inscrutable acoustic products as part of the problem and to rely upon specialized perceptual processes as the solution.

In contrast, other researchers and theorists have emphasized the acoustic signal and its auditory impression as most clearly defining the phonetic message. Liberman refers to such approaches as "conventional" because, for better or worse, these views rest most comfortably within traditions of perception and cognition. Pastore (1981), for example, draws attention to psychoacoustic factors and similarities between perception of speech and perception of other complex sounds. Arguments supporting acoustic objects of speech perception are not based solely upon general similarities. Diehl and Kluender (1989a, b) argue that acoustic patterns and their auditory consequences are the appropriate objects of speech perception on the basis of an array of phonetic regularities across languages and a number of empirical findings (e.g., Diehl, 1987; Diehl *et al.*, 1990; Kluender *et al.*, 1987).

Diehl and others (e.g., Diehl and Kluender, 1989a, b; Kingston and Diehl, 1994, 1995; Lotto *et al.*, 1997a) have

<sup>a)</sup>Electronic mail: kluender@macc.wisc.edu

extended this approach to suggest that talkers actively manage articulation to produce acoustic patterns most congenial to general operating characteristics of auditory systems. In this case, one may be able to conceptualize the perceptual side of the equation as more general or tractable, but at the expense of increased flexibility and control of articulation (e.g., Kingston and Diehl, 1994).

Nearey (1997) attempts to distinguish his approach from more polarized perspectives. In a field he presents as being cleaved by theories that are “strong gestural” versus those that are “strong auditory,”<sup>2</sup> Nearey (1997) proposes an alternative approach which he labels “double-weak.” He describes his position as double-weak because he envisions relatively loose constraints both on production and perception. By this account, much of speech perception is achieved through simple pattern-recognition processes which map acoustic attributes onto symbolic phonemic labels. Although Nearey adopts acoustic patterns to be the objects of perception, he assumes relatively weak constraints to be imposed by audition. Instead, the theory suggests that acoustic covariation is the product of real-time demands on articulation, but only up to a point. Nearey does not suggest that pattern recognition operates upon whatever acoustic patterns happen to arise from gestures. Instead, he suggests that talkers find ways to accommodate processing capabilities of pattern recognizers such as those he proposes through “stylized output patterns.”

Avoidance of strong claims, as Nearey attempts to do, does not necessarily result in a model that is less parsimonious. This is because, in some cases, parsimony on one side of the communication channel can incur costs on the other side. Based upon the conclusion that more general processes are inadequate, motor theorists (e.g., Liberman and Mattingly, 1985) posit specialized perceptual processes in the form of a speech module.<sup>3</sup> On the other hand, a strong auditorist may have confidence in the power of default perceptual processes so long as the talker takes on the burden of added articulatory planning and control. Parsimony may suffer no more by a double-weak approach than by either a strong gesturalist or strong auditorist approach.

To the extent that Nearey (1997) assumes some middle ground, however, it bears note that the middle ground is not virgin territory. As Nearey notes, for even the most ardent gestural and auditory theorists, positions have not been exclusively one-sided. Gesturalists (e.g., Fowler, 1989) note the importance of perceptual constraints in development of phonetic inventories, and auditorists admit the importance of minimizing articulatory effort (e.g., Kingston and Diehl, 1994). Further, a good many researchers make progress without expressing any strong commitment to gestures or acoustics; although, assumptions typically are implicit in designs of experiments and the questions they address.

It is unlikely that the double-weak approach will provide a harmonious midpoint between gestures and acoustics. In large part, this is because from the start, Nearey adopts auditory patterns as the objects of speech perception. In fact, it may be inescapable that researchers explicitly or implicitly embrace some sense of what it is that is perceived, be it gestural or auditory.

Further, although Nearey does not emphasize control of production for auditory effect, he looks favorably upon articulation being actively managed to be more amenable to pattern-recognition models such as his own. In particular, he suggests that speech is decodable via simple pattern recognition because talkers adhere to “orderly output constraints” such as those proposed by Sussman and his colleagues (Sussman *et al.*, 1995, 1998).<sup>4</sup> Given the embracing of acoustic/auditory patterns as objects of speech perception and acceptance of articulation organized to facilitate perception, Nearey’s approach tilts clearly away from gestural approaches. Relative to other approaches, double-weak bridges the divide between auditory and gestural approaches modestly if at all.

## II. THE MODEL

One way that the double-weak approach *does* distinguish itself from other contemporary psychological approaches is that it can be viewed as “a crassly empirical approach to phonological contrast” (Nearey, 1997, p. 3252). Nearey’s approach simply is to model the mapping between individual acoustic attributes of the signal and phonological elements. Using logistic regression analysis, Nearey generates *territorial maps* depicting regions in space corresponding to particular percepts with boundaries separating these regions. The essential feature is not logistic regression analysis *per se*. The framework is one of general linear analysis; there are strong commonalities between Nearey’s models and analysis of covariance. Similarly, the models are closely related to fuzzy logical models (Massaro, 1987; Massaro and Oden, 1980; Oden and Massaro, 1978). The force of logistic regression models is generality and simplicity.

In the simplest form of the logistic regression model presented by Nearey, the *primary cue model*, a particular property of the signal (cue) contributes solely and independently to define a phonetic distinction. In Nearey’s (1997) example, the attribute “vocoid duration” (vowel duration) in VCs varying in tenseness–laxness<sup>5</sup> of the vowel (/æ/–/ɛ/) and voicing of the consonant (/d/–/t/) could contribute to identification of the vowel or the consonant, but not both. Nearey uses this model to set a lower bound on simplicity, for such a model is at odds with perceptual data because a single acoustic property can contribute simultaneously to perception of more than one phonetic unit (e.g., Mermelstein, 1978; Repp *et al.*, 1978).

Slightly more complex is what Nearey refers to as a *secondary cue model* in which a given auditory property can contribute to defining more than one phonetic distinction. For Nearey’s VC example, vowel length can contribute to distinguishing both vowel (longer→/æ/) and consonant (longer→/d/). The term *secondary* refers to the notion that one acoustic attribute may be weighted less heavily than another. Hence, the same attribute may be weighted heavily for one distinction (*primary*) and be weighted less heavily (*secondary*) for another distinction. Although these distinctions may be helpful in understanding particular applications, the terms *primary*, *secondary*, *tertiary*, etc., relate to *territorial maps* only to the extent that they reflect relative magnitude of coefficients of the logistic regression model. Formally, if

not in practice, two attributes could have exactly the same coefficients. As a general linear model, there also is nothing that constrains logistic regression analysis to include no more than two attributes. However, it is easier to visualize results on a plane, and it is more practical to study perception of stimuli when only two attributes simultaneously covary.

To the *secondary cue model*, Nearey (1990, 1992) adds a final term which he refers to as a *diphone bias*. This term permits interaction between the identity of adjacent phonemes. What this term does is permit some diphones (/æɪ/ and /ɛɪ/ in Nearey's example) to be more probable outputs (relative to /ɛɪ/ and /æɪ/). In this example, more lax+voiceless also results in more tense+voiced VCs. In a *territorial map*, *diphone bias* permits regions to be translated along a dimension such that the areas of two diphones are larger at the expense of the areas for the other two phoneme combinations.

Nearey (1997) speculates about why listeners may behave as if such bias effects exist. He considers the possibilities that they represent phonotactic constraints (some VC sequences are more common than others), or they could reflect underlying lexical biases (e.g., some VCs occur in more words or in words that are heard more often.) Recent studies with infant listeners (Aslin *et al.*, 1998; Saffran *et al.*, 1996) are consistent with this approach.

Overall, the *diphone-biased secondary-cue* model does an admirable job accommodating the variance in listeners' responses to VCs varying in two dimensions. In either case, experience with probabilities of co-occurrences could serve to induce *diphone bias*, thus maintaining Nearey's empirical commitment. It may prove important that this bias must be symmetric in Nearey's models to the extent that symmetrical bias effects do or do not generally coincide with actual frequencies of co-occurrence.

Nearey's models are admirably frugal when it comes to computational power. He considers how the fit of the model could be improved by adding additional terms. While more free parameters always will improve performance, Nearey demonstrates not only that the limited *diphone-biased secondary-cue* logistic regression provides quite accurate description of the data, but also that additional variables do not account for sufficient variance to justify their inclusion. The models embody welcome elegance and tractability. Unlike, for example, multi-layer connectionist simulations in which "hidden units" permit nonlinear mapping and render difficult or impossible the task of understanding how a network succeeds if and when it does, logistic regression models and their general linear relatives are simple and understandable. In the cases Nearey presents, these models work and one knows why they work. As Nearey himself shows (Nearey and Shamma, 1987; Nearey, 1997), there likely will be cases when they do not work and linearity must be violated. These failures will be the kind of useful failures upon which good science prospers. For now, these models are simple and general, and it will be possible to add power prudently to these models as required by the data.

### III. LEARNING AND EXPERIENCE

Nearey's empiricist approach also implies a welcome emphasis upon learning, or more accurately, learnability. Although Nearey mostly describes his modeling efforts as investigations of processes of pattern recognition, he makes the point that his modeling should be taken, at least for now, as a simple case of auditory-perceptual learning.<sup>6</sup> Most speech perception theorists acknowledge the importance of learning, and most appeal to experience in a language environment as explanation. One lapse in Nearey's depiction of strong gesturalists and strong auditorists is the fact that he does not acknowledge that they, too, appreciate the importance of learning. Best (1995), as a strong gesturalist and proponent of direct realism, focuses upon the role of experience. Diehl and Kluender (1989a, b; Kluender and Diehl, 1987; Kluender *et al.*, 1987), among the strong auditorists Nearey notes, clearly embrace the role of experience and learning in shaping perception of speech.

In the last decade, investigators have been keenly interested in studies demonstrating changes in how infants respond to speech sounds as they have increasing amounts of experience in a language environment (e.g., Best *et al.*, 1988; Greiser and Kuhl, 1989; Kluender *et al.*, 1998; Kuhl, 1991; Kuhl *et al.*, 1992; Werker and Lalonde, 1988; Werker and Tees, 1984a, b). What is lacking, however, is sufficient modeling concerning how experience brings the perceptual system to behave in a manner that respects distributional properties of speech sound attributes.

Nearey's (1997) major contribution is making explicit some of the characteristics that he sees to be embodied by a model for how language-users are capable of exploiting correlations between multiple stimulus attributes and linguistically significant units. To the extent to which Nearey constrains his models to use as little computational power as possible, he has something in common with gesturalists<sup>7</sup> and auditorists. While they work to require relatively less from speech production and speech perception, respectively, Nearey's models require relatively little computational (brain) power. Nearey's models of pattern recognition are constrained to be linear and he argues for the preeminence of the segment as an economical unit of phonetic perception. Although the present authors are not sanguine regarding the role of phonetic segments *per se*, if one were to develop a language system that is easily learnable by most members of a group, a perceptron-like process like Nearey's may be about as simple as one could desire. If the architecture of phonetic perception can be so elegant, effects of experience evidenced in six-month-old infants probably ought not be surprising.

Nearey's approach also can be viewed as being constrained by ecology in as much as processes are tuned by distributional properties encountered through experience. In many ways, his empirical approach has a precedent in the theory of Egon Brunswik over 50 years ago (Brunswik, 1937, 1940, 1944, 1955; Postman and Tolman, 1959). For Brunswik, too, no single stimulus attribute is necessary and sufficient for perception of an object or event. Multiple stimulus attributes are used, each weighted to reflect probability and manner of occurrence in the world.<sup>8</sup> In this re-

gard, perception of multiply-specified phonetic units would be much the same as for other objects and events (Kluender, 1994).

It has been demonstrated that quite simple processes may be adequate to account for perceptual learning of multiply-specified phonetic equivalence classes such as that for alveolar stops (e.g., Kluender, 1998). Kluender *et al.* (1987) found that Japanese quail (*Coturnix coturnix japonica*) could learn the ostensibly complex mapping between multiple acoustic attributes and a response that maps onto alveolar stop consonants. Despite arguably feeble mental capacity, quail learning generalized accurately to novel syllables beginning with /d/.

Kluender and Diehl (1987) proceeded to explore ways in which multiple acoustic properties could be used by explicitly training quail to learn functional mappings without benefit of any single necessary or sufficient property. They used syllables produced by ten different talkers that varied on three dimensions: initial-consonant type (voiced versus voiceless, velar versus nonvelar), vowel type (front versus back, high versus low), and sex of talker. Two quail learned different equivalence classes based upon these dimensions. For one bird, the positive-class properties were voiced stop, front vowel, and male talker, while for a second bird, positive properties were velar stop, high vowel, and female talker. During training, birds were reinforced for pecking when hearing stimuli that were positive on exactly two of the three dimensions, and were not reinforced when hearing stimuli that were negative on two of the dimensions.

As formidable a task as this may appear, both quail learned to use all three constituent dimensions in mastering the task. Each quail generalized to novel instances of its equivalence class, pecking significantly more to novel positive stimuli than to novel negative ones. This happened when novel stimuli were positive on only two of three dimensions, and the highest level of responding occurred when the novel stimulus was positive on all three dimensions (a situation never encountered during training). Such behavior is consistent with theories of human categorization that account for the fact that certain category instances seem more representative than others and are generally easier to learn and remember, even when they have never been directly experienced (Posner and Keele, 1968; Rosch, 1978). With the exception of triple-positive stimuli, quail responded more strongly to stimuli used in training than to novel tokens, consistent with accounts of human category formation that emphasize experience with individual exemplars (Medin and Schaffer, 1978).

More recent data imply that relatively simple learning processes may account for the fact that some acoustic instances of vowel sounds are phonemically more compelling than others. Kluender *et al.* (1998) trained eight European starlings (*Sturnus vulgaris*) to respond differentially to vowel tokens drawn from stylized distributions for the English vowels /i/ and /ɪ/, or from two distributions of vowel sounds (approximate to /y/ and /ʉ/) that were orthogonal in the  $F1-F2$  plane. Following training, starlings' responses to novel stimuli drawn from these distributions could be predicted well on the bases of frequencies of the first two for-

mants as well as by distance from the centroid of these distributions of vowel sounds. Graded responses relative to the centroid were much like those often taken to imply the existence of category prototypes.<sup>9</sup> Starling responses corresponded closely to adult human judgments of "goodness" for English vowel sounds. Most germane to Nearey's efforts, a simple linear association network model trained with vowels drawn from the birds' training set captured 95% of the variance in birds' response rates for novel vowel tokens. Nearey's logistic models can successfully accommodate all of these avian response patterns because these are problems that require linear combinations of stimulus attributes.

Another virtue of Nearey's logistic models lies in what they cannot do. Following the successful Kluender and Diehl (1987) experiments in which quail conquered the simultaneously introduced complexities of vowel frontness/backness, consonantal voicing and place of articulation, and sex of talker, it was questioned whether a nonperceptron or XOR-like problem was beyond the pale of quail performance. In this study, in order for a syllable to be positive, it must **either** have a labial [b,p] stop and a high [i,u] vowel, **or** have a nonlabial stop [d,t,g,k] and a low [æ,ɔ] vowel. Despite training for a full year (the average lifespan for the species), performance on training stimuli never approximated avian performance for any of the forementioned tasks, and there was no positive transfer of performance to novel stimuli that obeyed the rule.

Nearey's linear logistic models would do no better than quail at this task. This is good. Despite the potential power of logistic or quail models of learning, there exist principled constraints upon the sorts of processes these models can accommodate. In the case of Nearey's logistic model, one reason this XOR problem cannot be accommodated is that the solution would require the functions for dimensions of voicing and of vowel height to be nonmonotonic. Such "failures" can be taken as encouraging in as much as Nearey's models have an elegance that constrains the domain of solutions in a principled and informative manner.

One may view as a potential weakness of Nearey's and similar empiricist approaches the fact that few predictions can be made beyond maintaining that these models ought to reflect distributional properties of the input. It then becomes critical that, as simple learning models, one can predict that models will fail when the task exceeds certain computational bounds. And, very significantly, Nearey has done admirable job of demonstrating how one may be able to account for the vast majority of perceptual data with a nicely constrained segment-based model. Moreover, perceptron-like linear logistic models may have some biological plausibility, as analogous processes apparently can be embodied in the neural substrate of an animal with presumably quite limited intellect. These are important contributions, but they are only a part of the story.

## IV. PERILS

### A. What cannot be explained

Is there any reason one may be reluctant to be fully satisfied with what Nearey calls a "crassly empiricist ap-

proach''? For example, why might one continue to search for constraints on auditory perception with concomitant demands upon articulation?

Purely empiricist models are quite silent regarding why linguistic sound inventories are structured as they are. Why, for example, are vowels longer before voiced consonants in all or almost all languages (see, e.g., Kluender *et al.*, 1988)? What logistic models will do is describe response patterns that are sensitive to the fact that such covariation exists. These models do not explain **why** covariation exists. To better appreciate why searching for articulatory and auditory explanations continues to be pivotal, imagine what would happen if, contrary to covariation typical of sound systems, [ɛ] was typically produced with longer duration than [æ]. In this revised bæd–bet perceptual study, listeners should perceive more long vowels as [ɛ] and more short vowels as [æ]. And, the logistic model will fit the data exactly as well, accounting for the same amount of variance. This is not to say that there is no virtue in a model of perceptual learning by which lawful variance in the world is reflected in lawful perception and behavior. But, a model that, for the most part, reflects probabilities in an inverted imaginary world as well as it does in a real world cannot be taken as fully explanatory. Nearey appears to be appropriately cautious about this general problem when he suggests that real-time demands upon articulation constrain the range of realized covariation. Nearey also portrays some regularities as products of articulation “stylized” to accommodate pattern recognition; however, pattern recognition would work exactly as well if vowel duration and voicing were stylized in reverse.

## B. What may be missed

A related concern is that, sometimes, the success of such an empiricist model can clearly mislead. Kluender (1991) conducted a series of perceptual studies with quail that were designed to investigate the role of *F1* transitions on perception of voicing for syllable-initial stops. When *F1*-onset frequency is lower, a longer *F1* cutback (one of the acoustic products of voice onset time) is required for human listeners to perceive synthesized stop consonants as voiceless. He synthesized CVs with four types of natural-like rising *F1* transitions that varied in duration of *F1* cutback. Thus, endpoint training stimuli (with 5- or 65-ms *F1* cutback) always differed in onset frequency of *F1* because a longer cutback resulted in a higher *F1* onset. (See Fig. 1.) Comparable effects of *F1*-onset frequency across the four *F1* types were found for human and avian listeners. Lower *F1*-onset frequency shifted the boundary to higher values of *F1* cutback (more “voiced” responses).

These data can be interpreted as evidence of quail and humans responding to simple covariations in the input and, thus, are amenable to Nearey’s logistic model approach. The *F1*-onset frequency covaried with *F1*-cutback for all the series in the Kluender (1991) study, and earlier work had demonstrated the ability of quail to learn covariations between acoustic attributes (Kluender and Diehl, 1987; Kluender *et al.*, 1987). As is the case for data from these previous studies, Nearey’s models would seamlessly incorporate this

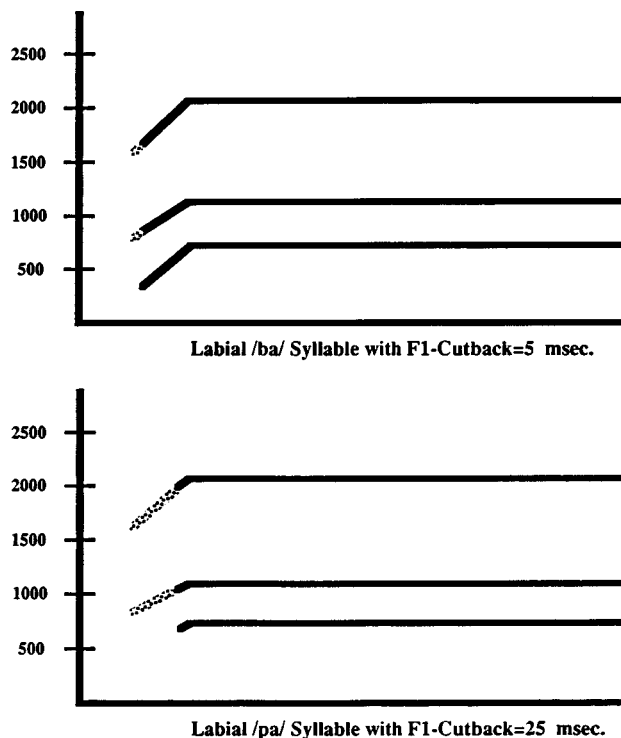


FIG. 1. During release of an initial stop consonant (in this case labial), the frequency of *F1* increases. Prior to voice onset, there is little or no energy in the region of *F1*. Consequently, when delay in voice onset is longer, the onset frequency of *F1* is higher.

covariance with no reference to auditory processes and the model would provide a reasonable fit.

However, a simple pattern-recognition approach would fail to discover that auditory processes were at least in part responsible for the shift in birds’ responses (and, by extension, humans’ responses). Another quail experiment was conducted with synthetic continua having either a relatively low (375 Hz) or high (750 Hz) constant-frequency *F1* (Kluender and Lotto, 1994). Because *F1* did not vary in frequency, there was no opportunity for birds to use covariation of *F1*-onset frequency and *F1* cutback. Despite this, quail exhibited reliably shorter “labeling” boundaries (more voiceless stops) for intermediate stimuli of continua when *F1* frequency was higher. Human performance with the same stimuli was like that for birds.

Essentially the same effect was observed for human and quail whether or not there was any possibility to use covariation between *F1*-onset frequency and *F1*-cutback. Absent the opportunity for learning the natural relationship between *F1*-cutback and *F1*-onset frequency, these results suggest that the effect of *F1*-onset frequency is due to quite general auditory processes common to humans and quail. The problem is that an empiricist model would have fit the original data very well, and there would have been little reason to suspect what was true all along; the effect is of general auditory—not learning—origins (see also Parker, 1988). Without denying the likely event that experienced covariance can play a role as well, it appears that modeling these data solely as an empiricist process of simple pattern recognition would have been misleading.

In other cases as well, one may find that, despite the fact

that a logistic model will accommodate labeling data with near-perfect accuracy, such empiricist models may be misleading. Worse is that these mistakes can deprive one's model of economy if multiple acoustic attributes collapse onto a smaller number of auditory properties. This argument does not depend upon constructs such as Kingston and Diehl's (1994) putative "intermediate perceptual properties." Simpler more obvious instances exist. For example, at relatively short durations (<200 ms), loudness is the product of both physical time and intensity (e.g., Blodgett *et al.*, 1958; Zwicker and Wright, 1963; Zwislocki, 1960). The fact that auditory systems "summate energy" in this fashion may well account for the existence of trading relations between duration and amplitude of aspiration energy for English stops (Repp, 1979). Similarly, loudness integration likely explains the finding that identification of heavily aspirated stops by Korean listeners is predicted by both duration and amplitude of aspiration energy (Kim *et al.*, 1994).

Nearey does anticipate the virtue of limiting the inventory of attributes and, thus, decreasing the degrees of freedom; although, he expresses concern about what he calls "a premature commitment to a limited set of properties." His caution is well taken. However, even if one embraces the empiricist strategy in general, the empiricist approach should probably be the less attractive default condition invoked only after one has wrenched all the explanation one can from lawful constraints upon speech production and auditory perception.

Nearey's computational austerity is laudable with respect to virtues of good scientific theory, but it is questionable whether these models will be able to scale up gracefully from single-syllable examples to longer strings of fluent speech. Effects of coarticulation can be long lasting and quite varied, and the number of acoustic attributes that will have to be placed in the model will become large. Some coarticulatory effects cross-syllable boundaries. For example, the acoustic realization of [d] in the disyllable [alda] is quite different from [d] in [arda] (Dianora *et al.*, 1996). Labeling tasks have demonstrated that humans' identifications of /d/ in these syllables are sensitive to the characteristics of the preceding liquids (Mann, 1980). To account for these data, Nearey's models would need to include acoustic variables measured in the preceding syllable. *Diphone-bias* terms would have to operate across both preceding and following phonemes at a minimum. The number of attributes could increase dramatically to account for all coarticulatory and suprasegmental effects which may affect phonemic identification. As noted before, the logistic regression model can scale up easily to accommodate additional variables. However, such explosion of learned attributes may not be necessary.

Lotto and Kluender (1998) have described similar effects of preceding acoustic energy for nonspeech/speech hybrid stimuli, and Lotto *et al.* (1997b) have demonstrated that quail respond in the same pattern as humans for [alda] and [arda] stimuli, despite never having experience with coarticulatory covariation. It appears that many effects of covariation, even those occurring across syllables, can be accounted for by general auditory mechanisms. The problem is

not simply that Nearey's model downplays the importance of these general mechanisms, but that claims of learnability and simplicity are unnecessarily compromised as a result of failure to recognize the role of the auditory system. An approach that includes **both** aspects of auditory function and learning (e.g., Diehl and Kluender, 1987) would be more complete and has a realistic chance of accounting for most phenomena of speech perception. As noted above,<sup>6</sup> separating auditory processes and effects of experience may not even be possible.

### C. Audition and learning

With the suggestion that a complete model include both auditory and learning processes, a legitimate fear is that falsifiability becomes at risk. One way to characterize the foregoing is that part of the explanation of speech perception falls out of general auditory processes, and the remaining variance can be "mopped up" by learning processes. If the auditory hypothesis fails under test, the default position becomes the appeal to learning. In fact, Nearey describes his approach as having the potential to be viewed as a "fallback theoretical position" (1997, p. 3243). Stated this way, an auditory/learning model could be tough to falsify; however, it also could be true.

The response to concerns about falsifiability must be found in more specific definition of processes of audition and learning. What can be falsified are specific hypotheses about these processes. The occasional demonstration of similar patterns of response data for speech and nonspeech analogs (or, human and nonhuman subjects) may no longer be sufficient. Instead, hypotheses about specific auditory processes must be generated, and experiments that test specific predictions must be conducted. Instead of placing gestural theories at risk via analogy, specific auditory hypotheses must be placed at risk.

With respect to experience and learning, the same can be said. It is not very useful to hypothesize that learning plays some role. Of course it does. One type of information that will be critically important in the development of adequate models for perceptual experience with speech will be richer characterization of the input to the process. In this way, one's model can be constrained by ecology as well as computation. Next, learning processes must be given greater definition. Nearey (1997) demonstrates the power of one approach to this challenge, and he provides a fine example by making his model explicit and computationally simple. The model will be falsified and modified (likely by Nearey himself), and therein lies part of its strength.

### V. AUXILIARY ISSUES

Before conclusion, it may be useful to express some concerns with regard to attempts to reveal the processes through which a young child comes to exploit experienced regularities in becoming a proficient language user. Two seminal issues arise concerning learning and learnability. First, what is it that infants really learn? Do they learn about phonetic segments, syllables, or morphemes? It is beyond the present focus to explore each possibility. However, without

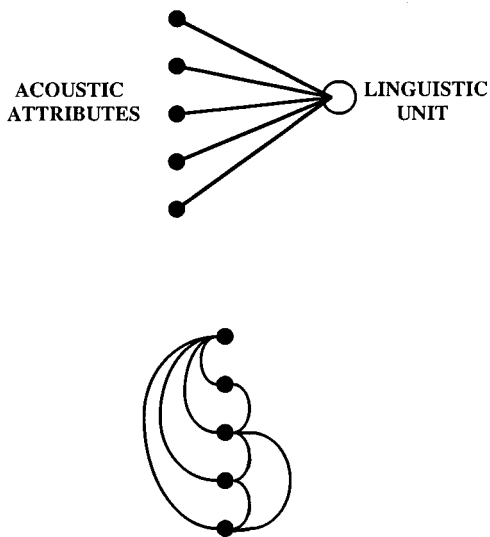


FIG. 2. The top figure illustrates multiple stimulus attributes each being correlated more or less with a particular phonetic unit. The bottom figure illustrates the concept of phonetic units defined as covariance between stimulus attributes.

dismissing phonetic units, one can imagine that something akin to phonetic dimensions come to play an increasingly important role as infants learn more functionally significant units such as morphemes. The traditional linguistic definition of phonemes is that they serve to efficiently distinguish morphemes. Because phonemes provide efficient descriptors of lexical space, could they emerge as dimensions of the developing lexical space instead of being taken as primitives (see, e.g., Lindblom *et al.*, 1984)?

The alternative, learning phonetic units *per se*, may require suggesting that lexical space begins its organization only after phonetic segments have been defined. Many reports concerning development of speech perception seem to portray the infant's task as one of learning phonemes as early functional units of language; however, explicit treatment of this question is rare. It is possible that neither lexical items nor phonetic units serve exclusively to structure the developing lexicon, and simultaneous organization seizing upon both sources of input regularities is not beyond possibility. In any case, this question is seminal for any computational or learning model aspiring to describe the use of experienced covariation in speech pattern recognition.

A closely related concern is that one could be misled by models that rely upon the correlation between particular stimulus attributes and putative phonetic units or symbols (Fig. 2, top). Nearey's linear logistic models seize precisely upon these correlations. Unless one is prepared to posit the innate existence of every phoneme from every language, and over 800 phonetic units is a lot (Kluender, 1994; Maddieson, 1984), one must consider an alternative. One alternative, favored here, is that correlations among stimulus attributes convey information (Fig. 2, bottom).<sup>10</sup> As a first approximation, one could characterize phonemes more as vectors through a covariance space. Studies of infant categorization of objects and events suggest that infants use correlated attributes to learn categories (Cohen and Strauss, 1979; Hu-

saim and Cohen, 1981; Younger, 1985; Younger and Cohen, 1983, 1985, 1986).

## VI. CONCLUSIONS

These questions concerning development extend beyond Nearey's present empiricist efforts, but they are central issues that should be of concern to all who find virtue in the empiricist approach to speech perception. Following the more circumspect nature of the second half of this commentary, the virtues of the double-weak approach merit brief repeat. Although there is some question regarding the degree to which Nearey has established novel middle ground between gestures and acoustics, there may be virtues to sidestepping a contentious issue about which efforts have shed more heat than light. Nearey has encouraged a balanced theory that may be more inclusive rather than exclusive. His approach implicitly employs constraints of ecology in as much as experienced probability distributions of attributes play a central role, and he encourages a welcome focus upon learnability, if not learning *per se*. Within the linear logistic framework, Nearey has constrained his models to embody admirable elegance and parsimony. Finally, Nearey's efforts may serve to illuminate seminal issues of ontogeny of phonetic perception.

## ACKNOWLEDGMENTS

The authors are grateful to Carol Fowler and Terrance Nearey for comments on an earlier version of this paper. This effort was supported by NSF Young Investigator Award DBS-9258482 to the first author.

<sup>1</sup>Not all theories that accept gestures to be the objects of speech perception require unique perceptual processes. Within the Gibsonian tradition, direct realism (e.g., Fowler, 1986) holds the objects of perception to be actual objects and events in the world. As a general theory of perception in force across modalities, direct realism takes gestures to be the objects of speech perception just as the theory would take physical structures (and not arrays of light of varying wavelengths and luminences) to be the proper objects of visual perception. In the broader view, direct realism deserves more attention than will be provided in this commentary. This is because Nearey (1997) does not consider direct realism to any great extent.

<sup>2</sup>Nearey (1997) briefly considers what he refers to as a "double-strong" theory of speech perception with strong relations between phonetic symbols and gestures and between phonetic symbols and auditory processes. Because Nearey argues that each of these putatively strong relations is relatively weak, this approach is doubly discouraged.

<sup>3</sup>Because one module is hypothesized to be responsible for both producing and perceiving speech, one may ask how one task could be more complex or special than the other. One way would be for one part, production of speech, to be much like motor control generally while the other part, perception of speech, is remarkable in that it is not accomplished in a fashion similar to perception of other objects and events in the world, instead hewing closely to motor control. Certainly, the emphasis of motor theory has been to distinguish perception of speech from more general models of perception.

<sup>4</sup>Because very recently there has been extended commentary upon Sussman's theorization (Sussman *et al.*, 1998), this aspect of Nearey's approach will not be considered in greater depth here.

<sup>5</sup>Describing vowels, particularly low vowels, as tense or lax is not unambiguous; however, for this discussion, the longer vowel /æ/ will be referred to as tense and the shorter vowel /ɛ/ as lax.

<sup>6</sup>In this commentary, auditory factors will be contrasted with experience and learning. This division of labor is too simplistic, and it is hoped that it is not misleading. Experience, particularly early in life, plays a considerable role in the development of sensory systems. Absent experience, nonpathological

perception does not exist. In addition, when something is the product of auditory-perceptual learning, it does not cease to be auditory because learning has occurred. Throughout this contribution, "auditory" can be taken to suggest general operating characteristics of auditory systems without consideration of experience within a particular domain (speech). "Learning" will refer to the processes by which performance comes to be in accord with systematic properties in the environment.

<sup>7</sup>Again, direct realism is distinct for other gesturalist approaches. This theory holds that perception of gestures is direct and no more complex or burdensome than perception of other objects or events.

<sup>8</sup>Brunswick did allow "psychological weights" to deviate somewhat from strict ecological validity.

<sup>9</sup>No inference is being made here with regard to the putative existence of prototypes. The present authors have argued elsewhere (Kluender *et al.*, 1998) that the theoretical construct of prototypes is unnecessary (see, e.g., Knapp and Anderson, 1984; Medin and Schaffer, 1978).

<sup>10</sup>These alternatives are rarely easy to distinguish for real-world objects and events such as speech because attributes that are relatively highly correlated with a category (perhaps phonetic) also tend to be correlated with one another.

- Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). "Computation of conditional probability statistics by 8-month-old infants," *Psych. Sci.* **9**, 321–324.
- Best, C. T. (1995). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience*, edited by W. Strange (York, Timonium, MD), pp. 171–204.
- Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English speaking adults and infants." *J. Exp. Psychol.* **14**, 345–360.
- Blodgett, H. C., Jeffress, L. A., and Taylor, R. W. (1958). "Relation of masked threshold to signal-duration for various phase-combinations," *Am. J. Psychol.* **71**, 283–290.
- Brunswick, E. (1937). "Psychology as a science of objective relations," *Philos. Sci.* **4**, 227–260.
- Brunswick, E. (1940). "Thing constancy as measured by correlation coefficients," *Psychol. Rev.* **47**, 69–78.
- Brunswick, E. (1944). "Distal focussing of perception: Size constancy in a representative sample of situations," *Psych. Mono.* **254**.
- Brunswick, E. (1955). "Representative design and probabilistic functionalism: A reply." *Psychol. Rev.* **62**, 236–242.
- Cohen, L. B., and Strauss, M. S. (1979). "Concept acquisition in the human infant," *Child Dev.* **7**, 419–424.
- Dianora, A., Hemphill, R., Hirata, Y., and Olson, K. (1996). "Effects of context and speaking rate on liquid-stop sequences: A reassessment of traditional acoustic cues," *J. Acoust. Soc. Am.* **100**, 2601.
- Diehl, R. L. (1987). "Auditory constraints on speech perception," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten, NATO ASI series, Series D, Behavioral and Social Sciences, No. 39 (Martinus-Nijhoff, Dordrecht, The Netherlands), pp. 210–219.
- Diehl, R. L., and Kluender, K. R. (1989a). "On the objects of speech perception," *Ecol. Psychol.* **1**, 121–144.
- Diehl, R. L., and Kluender, K. R. (1989b). "Reply to the commentators," *Ecol. Psychol.* **1**, 195–225.
- Diehl, R. L., Kluender, K. R., and Walsh, M. A. (1990). "Some auditory bases of speech perception and production," in *Advances in Speech, Hearing and Language Processing*, edited by W. A. Ainsworth (JAI, London), pp. 243–268.
- Fowler, C. A. (1986). "An event approach to the study of speech perception from a direct-realist perspective," *J. Phon.* **14**, 3–28.
- Fowler, C. A. (1989). "Real objects of speech perception: A commentary on Diehl and Kluender," *Ecol. Psychol.* **1**, 145–160.
- Greiser, D., and Kuhl, P. K. (1989). "Categorization of speech by infants: Support for speech-sound prototypes," *Dev. Psych.* **25**, 577–588.
- Husain, J. S., and Cohen, L. B. (1981). "Infant learning of ill-defined categories," *Merrill-Palmer Quart.* **27**, 443–456.
- Kim, M-R., Kluender, K. R., Lotto, A. J., and Read, C. (1994). "Perception of syllable-initial English stops by native Korean listeners," *J. Acoust. Soc. Am.* **95**, 2977.
- Kingston, J., and Diehl, R. L. (1994). "Phonetic knowledge," *Language* **70**, 419–454.
- Kingston, J., and Diehl, R. L. (1995). "Intermediate properties in the perception of distinctive feature values," in *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvanti (Cambridge U.P., Cambridge), pp. 7–27.
- Kluender, K. R. (1991). "Effects of first formant onset properties on voicing judgments result from processes not specific to humans," *J. Acoust. Soc. Am.* **90**, 83–96.
- Kluender, K. R. (1994). "Speech perception as a tractable problem in cognitive science," in *Handbook of Psycholinguistics*, edited by M. A. Gernsbacher (Academic, New York), pp. 173–217.
- Kluender, K. R. (1998). "Locus equations reveal learnability," *Behav. Brain Sci.* **21**, 273.
- Kluender, K. R., and Diehl, R. L. (1987). "Use of multiple speech dimensions in concept formation by Japanese quail," *J. Acoust. Soc. Am. Suppl.* **1** **82**, S84.
- Kluender, K. R., and Lotto, A. J. (1994). "Effects of first formant onset frequency on [-voice] judgments result from auditory processes not specific to speech," *J. Acoust. Soc. Am.* **95**, 1044–1052.
- Kluender, K. R., Diehl, R. L., and Killeen, P. R. (1987). "Japanese Quail can learn phonetic categories," *Science* **237**, 1195–1197.
- Kluender, K. R., Diehl, R. L., and Wright, B. A. (1988). "Vowel-length differences before voiced and voiceless consonants: An auditory explanation," *J. Phon.* **16**, 153–169.
- Kluender, K. R., Lotto, A. J., Holt, L. A., and Bloedel, S. L. (1998). "Role of experience for language-specific functional mappings of vowel sounds," *J. Acoust. Soc. Am.* **104**, 3568–3582.
- Knapp, A. G., and Anderson, J. A. (1984). "Theory of categorization based on distributed memory storage," *J. Exp. Psychol.* **10**, 616–637.
- Kuhl, P. K. (1991). "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories; Monkeys do not," *Percept. Psychophys.* **50**, 93–107.
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). "Linguistic experience alters phonetic perception in infants six-months of age," *Science* **255**, 606–608.
- Lieberman, A. M., and Mattingly, I. G. (1985). "The motor theory of speech perception revised," *Cognition* **21**, 1–36.
- Lieberman, A. M., and Mattingly, I. G. (1989). "A specialization for speech perception," *Science* **243**, 489–494.
- Lindblom, B., MacNeilage, P., and Studdert-Kennedy, M. (1984). "Self-organizing processes and the explanation of phonological universals," in *Explanations of Phonetic Universals*, edited by B. Butterworth, B. Comrie, and O. Dahl (Mouton, The Hague, The Netherlands).
- Lotto, A. J., and Kluender, K. R. (1998). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," *Percept. Psychophys.* **60**, 602–619.
- Lotto, A. J., Kluender, K. R., and Holt, L. L. (1997a). "Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*)," *J. Acoust. Soc. Am.* **102**, 1134–1140.
- Lotto, A. J., Kluender, K. R., and Holt, L. L. (1997b). "Effect of voice quality on the tense/lax distinction for English vowels," *Phonetica* **54**, 76–93.
- Maddieson, I. (1984). *Patterns of Sound* (Cambridge U.P., Cambridge).
- Mann, V. A. (1980). "Influence of preceding liquid on stop-consonant perception," *Percept. Psychophys.* **28**, 407–412.
- Massaro, D. W. (1987). *Speech Perception by Ear and Eye—A Paradigm for Psychological Inquiry* (Erlbaum, Hillsdale, NJ).
- Massaro, D. W., and Oden, G. C. (1980). "Evaluation and integration of acoustic features in speech perception," *J. Acoust. Soc. Am.* **67**, 996–1013.
- Medin, D. L., and Schaffer, M. M. (1978). "A context theory of classification learning," *Psychol. Rev.* **85**, 207–238.
- Mermelstein, P. (1978). "On the relationship between vowel and consonant identification when cued by the same acoustic information," *Percept. Psychophys.* **23**, 331–335.
- Nearey, T. M. (1990). "The segment as a unit of speech perception," *J. Phon.* **18**, 347–373.
- Nearey, T. M. (1992). "Context effects in a double-weak view of trading relations: Comments on Kingston and Diehl," in *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvanti (Cambridge U.P., Cambridge), pp. 28–40.
- Nearey, T. M. (1997). "Speech perception as pattern recognition," *J. Acoust. Soc. Am.* **101**, 3241–3254.
- Nearey, T. M., and Shammass, S. (1987). "Formant transitions as partly distinctive invariant properties in the identification of voiced stops," *Can. Acoust.* **15**, 17–24.
- Oden, G. C., and Massaro, D. W. (1978). "Integration of featural informa-

- tion in speech perception," *Psychol. Rev.* **85**, 172–191.
- Parker, E. M. (1988). "Auditory constraints on the perception of stop voicing: The influence of lower-tone frequency on judgements of tone-onset simultaneity," *J. Acoust. Soc. Am.* **83**, 1597–1607.
- Pastore, R. E. (1981). "Possible psychoacoustic factors in speech perception," in *Perspectives on the Study of Speech Perception*, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, NJ), pp. 165–205.
- Posner, M. I., and Keele, S. W. (1968). "On the genesis of abstract ideas," *J. Exp. Psychol.* **73**, 28–38.
- Postman, L., and Tolman, E. C. (1959). "Brunswik's probabilistic functionalism," in *Psychology: A Study of a Science. Volume 1. Sensory, Perceptual, and Physiological Formulations*, edited by S. Koch (McGraw-Hill, New York).
- Repp, B. H. (1979). "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," *Lang. Speech* **22**, 173–189.
- Repp, B. H., Liberman, A. M., Eccardt, T., and Pesetsky, D. (1978). "Perceptual integration of acoustic cues for stop, fricative, and affricate manner," *J. Exp. Psychol.* **4**, 621–637.
- Rosch, E. H. (1978). "Principles of categorization," in *Cognition and Categorization*, edited by E. Rosch and B. Lloyd (Erlbaum, Hillsdale, NJ).
- Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). "Statistical learning by 8-month-olds," *Science* **274**, 1926–1928.
- Sussman, H. M., Fruchter, D., and Cable, A. (1995). "Locus equations derived from compensatory articulation," *J. Acoust. Soc. Am.* **97**, 3112–3124.
- Sussman, H. M., Fruchter, D., Hilbert, J., and Sirosh (1998). "Linear correlates in the speech signal: The orderly output constraint," *Behav. Brain Sci.* **21**, 241–299.
- Werker, J. F., and Lalonde, C. E. (1988). "Cross-language speech perception: Initial capabilities and developmental change," *Dev. Psych.* **24**, 672–683.
- Werker, J. F., and Tees, R. C. (1984a). "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Inf. Behav. Dev.* **7**, 49–63.
- Werker, J. F., and Tees, R. C. (1984b). "Phonemic and phonetic factors in adult cross-language speech perception," *J. Acoust. Soc. Am.* **75**, 1866–1878.
- Younger, B. A. (1985). "The segregation of items into categories by 10-month-old infants," *Child Dev.* **56**, 1574–1583.
- Younger, B. A., and Cohen, L. B. (1983). "Infant perception of correlations among attributes," *Child Dev.* **54**, 858–867.
- Younger, B. A., and Cohen, L. B. (1985). "How infants form categories," in *The Psychology of Learning and Motivation: Advances in Research and Theory*, edited by G. Bower (Academic, New York), Vol. 19, pp. 211–247.
- Younger, B. A., and Cohen, L. B. (1986). "Developmental changes in infants' perception of correlations among attributes," *Child Dev.* **57**, 803–815.
- Zwicker, E., and Wright, H. N. (1963). "Temporal summation for tones in narrow-band noise," *J. Acoust. Soc. Am.* **35**, 691–695.
- Zwislocki, J. (1960). "Theory of temporal auditory summation," *J. Acoust. Soc. Am.* **32**, 1046–1060.