Speech perception as pattern recognition[a]

Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton, Alberta T6G 2E7, Canada

(Received 21 December 1995; accepted for publication 16 December 1996)

This work provides theoretical and empirical arguments in favor of an approach to phonetics that is called double-weak. It is so called because it assumes relatively weak constraints both on the articulatory gestures and on the auditory patterns that map phonological elements. This approach views speech production and perception as distinct but cooperative systems. Like the motor theory of speech perception, double-weak theory accepts that phonological units are modified by context in ways that are important to perception. It further agrees that many aspects of such context dependency have their origin in natural articulatory processes. However, double-weak theory sides with proponents of auditory theories of phonetics by accepting that the real-time objects of perception are well-defined auditory patterns. Because speakers find ways to obey ''orderly output conditions'' (Sussman *et al.*, 1995), listeners are able to successfully decode speech using relatively simple pattern-recognition mechanisms. It is suggested that this situation has arisen through a stylization of gestural patterns to accommodate real-time limits of the perceptual system. Results from a new perceptual experiment, involving a four-dimensional stimulus continuum and a 10-category /hVC/ response set, are shown to be largely compatible with this framework. © *1997 Acoustical Society of America.* [S0001-4966(97)05704-4]

PACS numbers: 43.10.Ln, 43.71.An, 43.71.Es [RAF]

INTRODUCTION

In this paper, I will sketch a framework called the *double-weak* theory of speech perception. Its main empirical hypothesis is that speech cues can be directly mapped onto phonological units of no larger than phoneme size. To put this theory in perspective, some underlying assumptions of current theories of phonetic specification will be examined. It is argued that most theories involve *a priori* constraints that may be too strong to stand the light of available data. The alternative advocated here requires only weaker versions of key assumptions of other theories. Many arguments presented below are summaries, elaborations, and minor revisions of views expressed elsewhere (Nearey, 1990, 1991, 1992, 1995). A summary of the theoretical and empirical work from the literature and results from a new experiment are presented. These generally support the main hypothesis of double-weak theory.

I. THEORETICAL BACKGROUND

Phonetics involves three logically distinct domains. The first of these is discrete and *symbolic*, consisting of one or more layers of phonological units (objects such as distinctive features, segments, moras, syllables, etc.). The other two domains are quasicontinuous and physical. They are the *gestural* (or articulatory) on the one hand and the *auditory* (or acoustic) on the other. The central issues of both phonology and phonetics involve deciding exactly what kinds of basic units and organizing principles exist in these three domains and how they relate to one another. The following discussion presupposes a ''default'' phonological organization of an utterance as a distinctive feature matrix (Chomsky and Halle, 1968). Given this representation, the central problem of phonetics concerns the way in which physical properties relate to distinctive features and phonetic segments.[1]

A. Strong theories

1. *Double-strong theory*

One important approach to this problem can be referred to as a *double-strong* theory of phonetic specification. Here the term *strong* is taken to imply a simple, robust, and transparent relation between physical and symbolic elements. Stevens and Blumstein's (1981) model is *double-strong*, in that it postulates (i) strong relations between symbols and gestures and (ii) strong relations between symbols and auditory properties. Note that there is a symmetry among objects in all three domains as represented in Fig. 1(a).

2. *Strong gestural and strong auditory theories*

Other theories are not so symmetrical: They postulate strong relations between *only one* of the physical domains and the symbolic level. *Strong-gestural* approaches may be exemplified by the motor theory of Liberman and Mattingly (1985). This postulates a strong, transparent relationship between symbols and gestures. There is a more complex and unidirectional path that relates gestures to auditory properties, as represented by the broken-shafted, single-headed arrow in Fig. 1(b). This reflects the complex, typically nonlinear mapping that psychomotor and biomechanical factors impose between natural units of the gestural domain and their acoustic consequences. The process of the inverse mapping from acoustics to symbols is not depicted. According to Liberman and Mattingly, this likely involves a complex analysis-by-synthesis process that generates internal auditory matches from symbolic hypotheses using an internal model

[a] ''Selected research articles'' are chosen occasionally by the Editor-in-Chief, that are judged (a) to have a subject of wide acoustical interest, and (b) to be written for understanding by broad acoustical readership.

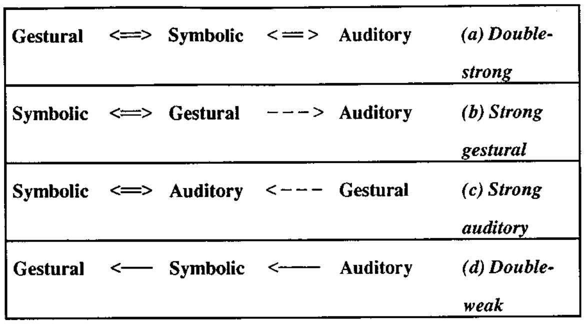
Gestural <==> Symbolic <==> Auditory (a) Double-strong
Symbolic <==> Gestural - - -> Auditory (b) Strong gestural
Symbolic <==> Auditory <- - - Gestural (c) Strong auditory
Gestural <— Symbolic <— Auditory (d) Double-weak

FIG. 1. Relations among domains in several theories. Double-shafted arrows indicate strong relations. Double-headed arrows indicate bidirectional relations. Broken-shafted arrows indicate highly complex, indirect relationships. Solid single shafted arrows indicate moderately complex, but highly systematic relations.

of the unidirectional mapping postulated. The direct realist theory of Fowler (1989) can probably also be represented fairly by Fig. 1(b). However, the inverse mapping from acoustics to gestures is *not* assumed to require analysis-by-synthesis decoding, but instead is said to involve ‘‘direct perception’’ of gestures by listeners.

Strong-auditory theories, exemplified most clearly by Diehl, Kingston, Kluender, and their colleagues (Diehl and Kluender, 1989; Kingston and Diehl, 1994, 1995) hold essentially the opposite position from that of the strong gesturalists. Relations are assumed to be strong between symbols and auditory properties but only weak and indirect between symbols and gestures [see Fig. 1(c)].

B. Criticism of strong theories

1. *Contra auditorists*

Any argument against either a strong auditory or a strong gesturalist approach is *a fortiori* an argument against a double-strong theory, so it suffices to review those two cases. The gesturalist criticism of auditorists has a long history. A main line of argumentation has been that the complex web of acoustic cues relevant to a particular phonetic distinction can only be understood in light of its articulatory source. An impressive collection of trading-relation and multiple cue experiments from the Haskins group have emphasized this theme (e.g., Liberman and Mattingly, 1985; Repp, 1982).

Consider the case of the voicing distinction in English stop+vowel syllables. Properties favoring voiced stops include negative or zero voice onset time (VOT), low $F1$ onset, and lowered $F0$. Those favoring voiceless stops include positive VOT, high $F1$ onset, and raised $F0$. These patterns are often interpreted by gesturalists as natural byproducts of a single glottal timing gesture. [See Kingston and Diehl (1994) for a critical review.]

2. *Contra gesturalists*

Auditorist criticism of gesturalist claims has a fairly long history in the case of vowels.[2] Ladefoged *et al.* (1972), Nearey (1980), and Johnson *et al.* (1993) have all argued that articulatory targets for the same vowel are quite varied across speakers. Since the corresponding acoustic output is more nearly invariant, they argue that linguistically relevant properties of vowels are acoustic or auditory rather than articulatory. Perkell *et al.* (1993) have recently shown a kind of trading relation between tongue and lip position in the production of /u/ that seems to be motivated by the acoustic synergy of distinct labial and velar constrictions. Some speakers appear to use varying articulatory means to achieve a relatively constant acoustic end in different phonetic contexts.

Kingston, Diehl, Kluender, and their colleagues (e.g., Kingston and Diehl, 1994, 1995; Diehl and Kluender, 1989) have recently launched a barrage of attacks on the gesturalist position. They have argued that a number of properties claimed to be the natural fallout of the biophysical interaction of gestures instead result from *deliberately controlled* actions. Specifically, they argue that many properties that covary in production are *actively managed* to produce auditory enhancement through the creation of a set of acoustic subproperties whose combination leads to derived perceptual properties (e.g., the *C:V ratio* or the *low-frequency property*) to which human auditory systems are particularly sensitive.

We thus have entirely contradictory claims from auditorists and gesturalists as to the organizational basis of important context dependent cues. As I have noted elsewhere (Nearey, 1991), I am genuinely impressed by the quality of the research by both the auditorists and the gesturalists that is critical of the other position. Each has convinced me that the others are wrong.

C. Simultaneous constraints in acoustic-articulatory space

Proponents of the double-strong theory are manifestly concerned with both articulation and audition. Although primarily defined in auditory terms, even the original Jakobsonian features (Jakobson *et al.*, 1963) were also supplied with straightforward (though sometimes not gesturally unique, as in the case of [±flat]) articulatory implementations. The development of the successor to this theory by Stevens and his colleagues has been directed toward finding a kind of harmony between stable articulatory implementations and robust auditory properties (Blumstein and Stevens, 1980; Stevens and Blumstein, 1981; Stevens and Keyser, 1989; Stevens, 1990). Stevens’ quantal theory suggests the motivation for the selection of a universal set of binary features rests on what might be considered ‘‘sweet spots’’ (Nearey, 1995) in the gestural by auditory space.

While simultaneous concerns for articulation and acoustics are natural to the double-strong approach, researchers in the other strong camps have also made concessions in that regard. Thus, the auditorists Kingston and Diehl (1994) argue that speakers ‘‘...optimize their phonetic behavior by both minimizing articulatory effort and maximizing [auditory] distinctiveness (p. 423).’’ Conversely, the gesturalist Fowler (1989) has stipulated that ‘‘perceptual constraints guide the development of sound inventories and of phonological processes in languages (1989:145).’’

Authors from both the single-strong schools have on occasion approvingly cited the research of Lindblom (e.g., 1986, 1990). Lindblom’s work partly involves the issue of real-time phenomena that affect immediate communication (e.g., his hyperspeech and hypospeech).[3] However, it also involves longer-term considerations such as the acquisition of sound contrasts and the truly secular issue of (sociocultural) evolution of phonological inventories. The distinction between long-term (diachronic), acquisitional, and real-time constraints may be crucial for understanding many aspects of linguistic behavior, including speech perception.

D. A double-weak approach to symbol-signal mapping

The *double-weak* approach outlined below can be viewed as a fallback theoretical position. In principle, the other theories (all of them, but double-strong theory in particular) would be preferable if they could be reconciled with the data. However, until there is compelling evidence that this can be done, it seems advisable to entertain other possibilities.

Figure 1(d) depicts the relation among domains in double-weak theory. Only two conditions are necessary for speech to operate as an effective communication system. First, a symbol sequence must be encoded into gestures. Second, the acoustic output of those gestures must provide the listener with auditory cues sufficient to decode the intended symbol sequence. Only subsets of symbol-to-gesture and sound-to-symbol mappings that meet this condition can be considered. On this account, gestures and auditory properties are linked only indirectly, through separate links to shared symbols. The arrows are unidirectional and are drawn with thin shafts in Fig. 1(d) to indicate that, unlike the strong theories, the relationships of properties to symbols is not necessarily straightforward. They are assumed only to be tractably systematic, to lie within the range of the feasible for the (possibly speech-specialized) auditory and motor control systems. Longer-term and secular trends effectively impose a communicative natural selection, ensuring that the phonology of a language remains within easy reach of the vast majority of speakers and listeners.

1. Double-weak phonetics from double-strong antecedents

It is instructive to imagine how a double-weak phonetic system might emerge from a double-strong one. Consider the case of the phonological opposition between voiced and voiceless stops postvocalically in prepausal position, i.e., in sequences such as /Vt/ or /Vd/. The textbook articulatory implementation of this opposition involves the presence (versus absence) of phonation during closure. This is a *double-strong* scenario in that both the gestures and their key acoustic consequences are readily identifiable. The portion of the waveform associated with the open vocal tract is readily distinguishable from the quasisinusoidal voice bar of the closed vocal tract section. As the vocal tract constriction becomes more radical and approaches complete closure, $F1$ will approach its low-frequency, closed tract asymptote (Fant, 1960).

Consider a set of hypothetical diachronic changes that could lead to a cue pattern more like that of modern English. For well-known aerodynamic reasons (Ohala, 1981; Kingston and Diehl, 1994), it is relatively difficult to produce voicing during complete vocal tract closure and some nearly heroic maneuvers may be necessary to maintain it for long. On the other hand, for the voiceless stops, glottal pulsing may not stop immediately when the vocal tract closes without additional adjustments (Kingston and Diehl, 1994). To maintain sufficient distinctiveness (Lindblom, 1990; Martinet, 1955; Kohler, 1984) some speakers may shut down voicing actively, e.g., by abducting the vocal folds a little early (or fully abducting them to form a glottal stop). This will keep voice bar out of the closure period. It will also cause a change in signal type at the time just prior the vocal tract closure. Strictly speaking, the result would be preaspiration (or glottalization). It may, however, be very close in its acoustic effect to silence. A necessary side effect of this gesture will be the truncation of the falling $F1$ pattern, as oral closure now occurs after $F1$ has become inaudible.

On the basis of results of Watson and his colleagues (Kewley-Port *et al.*, 1988; Watson and Kewley-Port, 1988; Watson and Foyle, 1985), it seems reasonable to suggest that the human auditory system is fully capable of tracking most of these modifications in good listening conditions. Assume also that listeners are capable of exploiting the correlation between apparent vowel duration, apparent closure duration, $F1$ termination, and the voiceless character of the following closure interval to support the phonological contrast between the final stops.

From a strong-gesturalist perspective, all the consequences of the change in glottal timing would be attributed by listeners to just that. Under a double-weak account, however, the cues have a potentially independent status and may be seized upon by other speakers *as proper signals of the consonantal opposition.* This, in turn, opens the door for some of these cues to be approximated *by means quite distinct from glottal timing.* Thus, the most salient effects of preaspiration (or glottalization) might be approximated by a longer (unvoiced) oral closure period. Such programmed changes in gestural patterns might further shift the ''functional load'' of cues so that the presence of voice bar during closure is not as important as it was earlier and speakers may be able to get away with hardly producing any at all.

Other noticeable properties that may have started as accidents of production could also be enhanced. Vocoid shortening is a likely articulatorily byproduct of early devoicing and of closure enhancement in the voiceless environment. However, the duration of the voiced closure could also be shortened for /Vd/ (and the V actively lengthened) to enhance the emerging differences in temporal signature between the /Vt/ and /Vd/ patterns.

Although some speech-specific perceptual mechanisms might be involved, it seems more reasonable to assume that such a reweighting of cues could take place through a general process of auditory-perceptual learning. Cue evolution is viewed as diachronic in the traditional sense. As such, children and perhaps older speakers who are retuning their dialect under sociolinguistic influences (Labov, 1972) might be expected to bear special responsibility in these changes. This resembles Ohala's account of listener-oriented sound charge (Ohala and Shriberg, 1990). However, unlike the cases considered by Ohala, all the changes just discussed are strictly subphonemic, operating in effect only on cue weights. There is no quantal reinterpretation at the symbolic level, only the specific gestural and acoustic properties that map the symbols have changed. At some later stage, if voice bar were eliminated during /d/-closure and if (perhaps later still) $F1$ termination differences were eliminated, phonologists would likely reinterpret the differences in terms of distinctive consonantal or vocalic quantity (with the other redundantly varying), rather than as a voicing distinction. Language acquires might arrive at a similar solution. At this point, an Ohala-style phonological *hypocorrection* might be said to have occurred.

2. Double-weak versus strong auditorist accounts

This account differs from that of strong-auditorists primarily in that it concedes that a principle source of the kinds of covariation in cue patterns is articulatory behavior. The complex pattern of $F1$, closure duration, and vowel duration can plausibly be attributed to an initial change in glottal timing. While auditorists like Kingston and Diehl have not entirely ruled out articulatory influences on the choice of covarying patterns, they have suggested that the primary source of covarying cue patterns are what they have termed *intermediate perceptual properties* (Kingston and Diehl, 1995). Two such properties plausibly involved in the case just discussed are the *V:C ratio* (which integrates information about vocalic and consonantal duration) and the *low-frequency property* (which integrates information about $F1$ offset and voice bar during closure). Such derived properties are presumed to be ‘‘hard-wired’’ in human (and possibly other) auditory systems, but unlike the quantal features of Stevens, they are not themselves part of a universal speech-specific inventory. Rather, Kingston and Diehl (1994) appear to assume that the derived auditory properties can be combined in various language-specific mixtures into a universal set of distinctive features.

As in Kingston and Diehl’s account, a double-weak approach assumes that temporally distributed relational features (such as the V:C ratio which can span a few hundreds of milliseconds) can be incorporated into localized phonological oppositions (such as the voicing distinction in stops). It also admits of language-specific weights. However, these weights are assumed to be applied to more primitive perceptual cues, i.e., to relatively direct perceptual correlates of acoustic properties like vocoid duration, voice-bar duration, etc. In this regard, it resembles the model of Lindau and Ladefoged (1986).

E. Pattern recognition of stylized output

A double-weak account may admit even weaker and more complex relations between phonological elements and physical properties than those described in Lindau and Ladefoged (1986), since it also allows considerable flexibility in the time alignment of properties and sharing of information among segmental elements. The theory as sketched above seems *a priori* to have little chance of offering any insight into universals of human language. [See Blumstein (1986).] From this perspective, it is precariously close to a neostructuralist position which posits that languages can have phonologies that can vary arbitrarily from each other. Although the work of Ladefoged (e.g., 1990) and others offers strong evidence that phonologies may differ from each other in ways not contemplated by some distinctive feature philosophies, the variation observed does *not* appear to be unbounded. Without additional constraints, we might expect the development of wildly varying phonologies across languages. However, additional real-time demands of production and perception may limit the relation of symbol to signal more stringently.

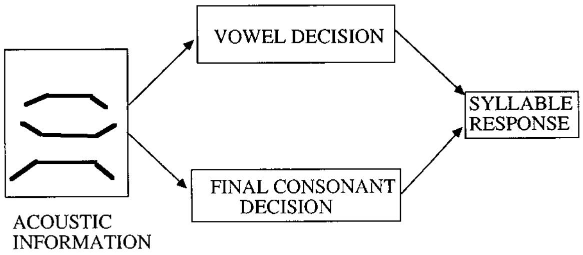

FIG. 2. Mermelstein’s 1978 model for VC judgments. Adapted from Mermelstein (1978), Fig. 3, with permission from Psychonomic Society Press.

What I suggest is that phonologies must allow for coarticulatory influence in production that eases the real-time demands on the speaker to move articulators too far too fast. On the other hand, coarticulatory influences may be actively managed to accommodate perceptual mechanisms of rather limited and specific computational capacity. This suggests that coarticulation is *stylized* to produce stylized covariance patterns in speech output. While the *direction* of the covariance has articulatory motivation, the details of the final patterns are constrained by what Sussman *et al.* (1995) have recently called ‘‘orderly output conditions.’’ Such stylized output patterns are amenable to decoding by relatively simple pattern-recognition techniques. If so, their simplicity should be revealed in the analysis of appropriately designed speech perception experiments.

******Highlight colors reset here******

II. FACTORING PHONOLOGICAL OPPOSITIONS

A. Cue sharing in VC syllables

Mermelstein (1978) presented an analysis of experimental results that lead to surprising conclusions. Mermelstein’s experiment involved VC syllables, where the vowel ranged over /ɛ/ and /æ/ and the consonant over /t/ and /d/. The stimuli involved manipulation of $F1$ of the steady-state vocoid and vocoid duration. Mermelstein’s analysis suggested that both vowel decisions and consonant are influenced by both $F1$ and vocoid duration, but *every aspect of the vowel decision is independent of consonant decision and vice versa.* (See also Allen, 1994.) This is represented schematically in Fig. 2.

The kinds of questions Mermelstein raised in his experiment can be addressed directly using logistic regression analysis, which shares many characteristics with the analysis of covariance (Haberman, 1979; McCullagh and Nelder, 1989). Two general classes of effects can be isolated in these models, as shown in Table I. These are bias effects and

TABLE I. Effects in a logistic regression model for Mermelstein's 1978 experiment.

Bias effects (Response only effects. Stimulus-independent effects)
V=vowel bias main effects
C=consonant bias main effects
V×C=diphone bias effects
Stimulus-tuned effects (Response-by-stimulus interactions)
F1-tuned effects:
V× $F1$=$F1$-tuned vowel effects (ɛ/æ distinctions)
C×$F1$=$F1$-tuned consonant effects (t/d)
V×C×$F1$=$F1$-tuned diphone effects
Vocoid Duration-tuned effects:
V×vocoid duration=duration-tuned vowel effects
C×Vocoid duration=vocoid duration-tuned vowel effects
V×C×vocoid duration=vocoid duration-tuned diphone effects

stimulus-tuned effects. The latter may be further broken down for each stimulus dimension. (For a more detailed discussion, see Nearey, 1990.)

A very simple (simpler than Mermelstein's) model that allows no property-sharing can be contemplated. Here, the $F1$ variation is related to vowel choices only, while the consonant decision is related to vocoid duration only. This can be called a *primary cue model*.[4] It would contain only the following terms: V, V×$F1$, C, C×vocoid duration. A labeled partitioning of the pattern space consistent with such a model is shown in Fig. 3(A). Such *territorial maps* can be generated in ways detailed by Nearey (1990, 1992) from the coefficients of the logistic regression models. Note that the vowel category boundary depends on (or ''is tuned by'') $F1$ but is independent of vocoid duration. Similarly, the consonant boundary is tuned by duration, but is orthogonal to $F1$.

Mermelstein's model in Fig. 2 could be termed a *secondary cue model*, where duration serves as a secondary cue to the vowel distinction and $F1$ serves as a secondary cue to the consonant distinction. This model is characterized by a logistic model including the following effects: V, V×$F1$, V×vocoid duration; C, C×$F1$, C×vocoid duration. The territorial map for such a model is shown in Fig. 3(B). Notice here that there is only a single line separating /Vt/ and /Vd/ responses and a single line separating /ɛC/ and /æC/ responses. Since neither of the lines are orthogonal to either of the axes, both consonants and vowels can be said to be *tuned* by both stimulus properties.

Whalen (1989) reported a replication and extension of Mermelstein's experiment and provided an analysis that indicated that Mermelstein's results did not generalize to larger scale experiments. Nearey (1990) reanalyzed Whalen's data using logistic regression analysis and agreed with Whalen that Mermelstein's model was not adequate. However, Nearey's reanalysis showed that Mermelstein's model required only slight modification to achieve very good agreement with the data. This modification was the addition of *diphone bias* effects (V×C). A key property of such models is that syllables differing by a single phoneme are constrained to have *parallel boundaries*: The /ɛt–æt/ boundary must be parallel to the /ɛd–æd/ line, and the /ɛt–ɛd/ boundary must be parallel to the /æt–æd/. The magnitude of the diphone bias effects required by Whalen's data are actually

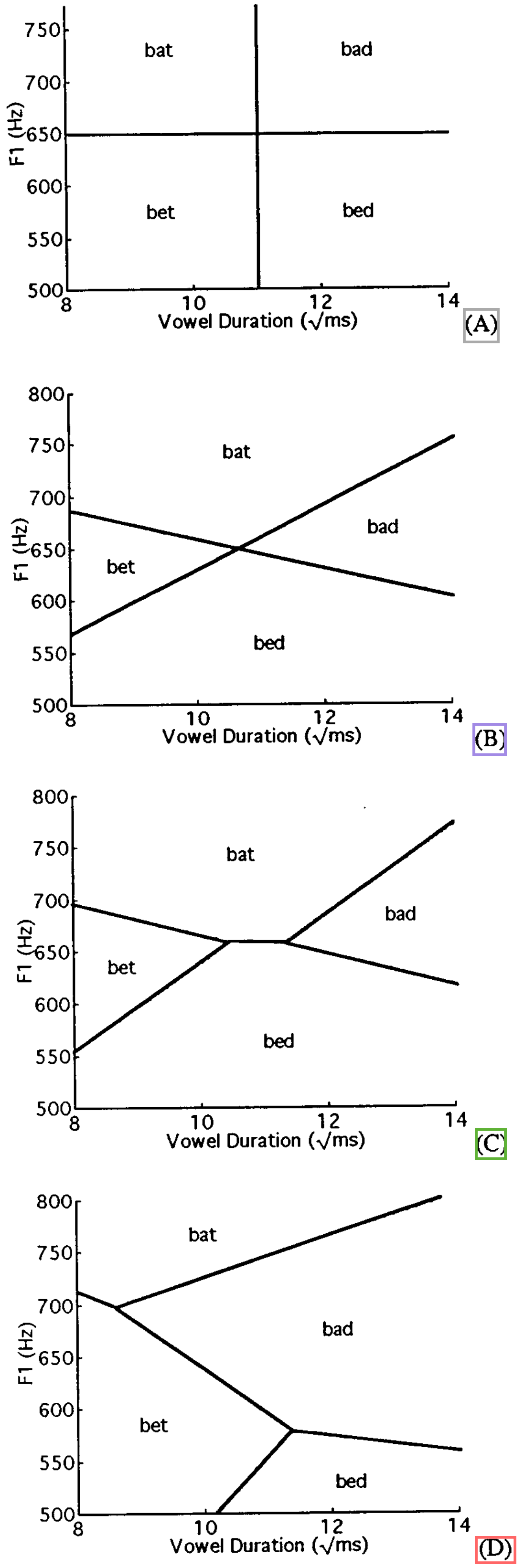

FIG. 3. Territorial maps for primary cue model (A); Mermelstein-like secondary cue model (B); secondary cue model with diphone bias terms (C); and full diphone model including stimulus-tuned diphone terms (D).

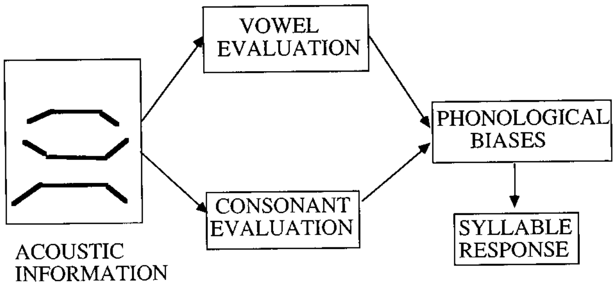

FIG. 4. Modification of Mermelstein's model to accommodate Nearey's (1990) diphone biases.

quite small. Their main consequences can be seen by comparing Fig. 3(C), which depicts the territorial map implied by Nearey's selected model, to the simpler secondary cue model of Fig. 3(B). Conceptually, the necessary modification to Mermelstein's block diagram is illustrated in Fig. 4. Note that consonant evaluation and vowel evaluation are modified by phonological (e.g., diphone) biases before a syllable decision is made.

What might diphone bias effects represent? They may serve several logically distinct functions. First, they can play the same role in implementing phonotactic constraints as Massaro and Cohen's (1983) contextual features do in their fuzzy logical models (which are computationally very similar to the logistics discussed here). Second, they may ''absorb'' differential lexical bias effects, whereby certain words (or nonsense syllables) are preferred by listeners independent of stimulus properties. Finally, they may serve simply as *fudge factors* that make the model fit the data better. In this example, the net effect of the diphone bias term is to increase the area of ''bat'' and ''bed'' responses (symmetrically, as it turns out for technical reasons) at the expense of the /ɛt/ and /æd/ categories.

Why might this be helpful? Peterson and Lehiste (1960) report that duration is about 40% greater for /æ/ than for /ɛ/. Similarly, vowels before /d/ are between 40% and 50% longer than those before /t/. Thus, the duration difference associated with the consonants is about the same as that associated with the vowels so that the duration pattern of the four syllables is /ɛt/</ɛd/≈/æt/</æd/. In the extreme cases, the duration characteristics associated with the vowels and consonants are synergistic, both short for /ɛt/ and both long for /æd/. The middle duration region represents a kind of conflict between duration patterns of the consonants and vowels involved. An optimal decision rule should reflect some kind of correlation between consonant and vowel decisions in the middle range, where only /ɛd/ and /æt/ are likely. For example, assume that other cues were relatively neutral with respect to /-t/ and /-d/ and that the vocoid duration was in the ambiguous range. Then, if the vowel sounds like /ɛ/, a listener could expect to reduce errors by choosing /-d/, because /ɛt/ is very unlikely in this duration range. There are a number of possible ways in which such interdependence could be implemented, as discussed by Nearey (1990, 1992). One of the simplest is through adjustment (''fudging'') of diphone biases to approximate the following rule: Bias decisions slightly in favor of the ''conflicting-cue'' /ɛd/ and /æt/ syllables over the others. Although it may appear inelegant, such a rule may provide a simple approximate solution to an otherwise difficult problem, as it can be applied automatically and globally. (In cases where cues are not conflicting, they will simply override the mild biases in favor of the conflicting cue categories.)

It might be objected that the effect of the above analysis is not very different from simply precompiling all the stimulus differences among the four syllables into four distinct diphone patterns. Nearey (1990) argues that if such precompilation were necessary, then radically more complex patterns could be expected to occur in perception. For example, a pattern like that in Fig. 3(d) would be admissible, where there is no requirement of parallelism on the boundary lines. This corresponds to a *complete diphone* model, which includes *stimulus-tuned diphone terms* corresponding to V×C ×vocoid duration and V×C× $F1$ interactions. Each cue is in effect allowed to have an arbitrary weight for each diphone element. Nearey (1990, 1992) has argued that the *absence of such complex patterns* is significant and may reflect the operation of relatively limited perceptual mechanisms that structure phonological systems. However, it is possible that the parallelism observed in this example is merely accidental. Perhaps if we looked at a larger problem, more complex patterns, such as that illustrated in Fig. 3(D), might yet arise. The following, larger experiment may provide more suitable testing ground for evaluating models.

III. EXPERIMENT I: PERCEPTION OF A LARGE /hVC/ CONTINUUM

A. Method

1. Stimuli

A set of 480 /hVC/ stimuli was synthesized, where V ranged over the five Canadian English back and central vowels /u,o,ʊ,ɒ,ʌ/ and C range over /t/ and /d/. The stimuli were synthesized at 10 kHz on an implementation of the Klatt80 synthesizer (Jamieson *et al.*, 1989; Klatt, 1980). They began with a 60-ms /h/-like fricative with formant frequencies of the following vocoid, but excited only by the aspiration source. This was followed by a variable steady-state vocoid with $F3$ fixed at 2350 Hz and $F4$ and $F5$ fixed at 3300 and 3750 Hz, respectively. The vocoid had a falling $F0$ contour from 125 to 100 Hz over the course of its duration. This was followed by 52-ms transition in $F2$ and $F3$ appropriate for a coronal stop. The $F2$ values moved from the steady state of the vocoid to a preclosure target of 1040 Hz+50% of the steady-state value. $F3$ moved from the steady state $F3$ of the vowel to 2700 Hz. $F1$ remained at its steady-state frequency during the $F2$–$F3$ transitions.[5] A weak, nonvoiced coronal stop burst was synthesized at 90 ms after the offset of the vocoid, with any interval between the voicing offset (of the vocoid or of the voice bar) being filled with silence.

The variable stimulus factors were arrayed in a four-factor, fully crossed design. (1) Vocoid duration controlled the duration of the steady-state vocalic part of the signal. This duration ranged from 90 to 218 ms in 32-ms steps. (2) Voice bar duration controlled a period of quasisinusoidal voicing (using the Klatt AVS parameter and $F0$ of 100 Hz).

This followed the vocoid and ranged from 0 to 90 ms in 30-ms steps. (3) $F1$ controlled the first formant frequency of the vocoid and ranged from 330 to 730 Hz in eight steps. (4) $F2$ separation controlled the difference $F2$ minus $F1$ of the vocoid and ranged from 425 to 700 Hz in three steps.

2. Subjects

Fifteen paid subjects, all of whom had a small amount of phonetic training, were recruited as listeners. Each categorized each stimulus five times, for a total of 2400 responses per subject. Responses were gathered for each subject over several sittings, usually on different days.

3. Presentation

The stimuli were presented under computer control on a PC using a 12 bit D-A converter and an antialiasing filter with a 5-kHz cutoff. Signals were then amplified and played through a small loudspeaker mounted on a table in front of the listener. The listener was seated in a sound-attenuated booth and responded using a computer mouse. The PC screen was positioned outside the booth and was visible through a double-gazed window. There were 10 response boxes labeled with the words ‘‘who’d,’’ ‘‘hood,’’ ‘‘hoed,’’ ‘‘hawed,’’ and the pseudoword ‘‘hud.’’ Pseudophonetic transcriptions were also provided using ASCII characters and a sheet listing the IPA transcription of the words was available for consulting in the listening booth. Subjects were run one at a time, with separate randomizations for each subject.

B. Results and discussion

1. Background

It is reasonable to divide the cues into primary and secondary for consonants and vowels, based on prior findings, including those of Nearey (1990). Primary cues are those that are expected to have a large effect on the probability of responses to the category in question. Thus $F1$ and $F2$ separation are both expected to strongly affect vowel responses (lower $F1$ favoring higher vowels, lower $F2$ backer or rounder vowels). Similarly, voice bar duration and vocoid duration are expected to strongly affect consonant response probabilities (longer voice bar and longer vocoid duration favoring /d/).

On the basis of Mermelstein’s and Whalen’s experiments (among others), vocoid duration is expected to have an effect on vowel responses (though probably somewhat weaker than $F1$ or $F2$) and is classed as an anticipated *secondary* vowel cue. Similarly $F1$ [both at the termination of the vocalic section and earlier in the steady-state vocoid (Summers, 1987, 1988)] is also expected to affect consonant choice (lower $F1$ favoring /d/). Two other cue relations of the same formal complexity (stimulus-by-segment interaction in analysis of covariance notation) as the primary cues are included for completeness. These will be referred to as *minor cues*. They are voice bar as a vowel cue and $F2$ as a consonant cue. A summary of these effects and their abbreviated labels are presented in Table II. Bias terms (whose labels involve V and C terms only, and no stimulus effects) are also indicated there.

TABLE II. Breakdown of effects in logistic models for experiment I. All effects listed are allowed by Nearey’s diphone-biased segmental models. If the VC diphone bias term is not included, models are compatible with Mermelstein’s independent segment models.

Biases:
Segmental V=vowel; C=consonant
Diphone: VC=vowel×consonant
Anticipated primary cue effects:
Primary vowel effects: VF=V× $F1$, VS=V×$F2$ separation
Primary Consonant effects: CB=C×voice bar, CD=C×vocoid duration
Anticipated secondary cue effects:
Secondary vowel effects: VD=V×vocoid duration
Secondary consonant effects: CF=C×$F1$
Possible minor cue effects:
Vowel effects: VB=V×voice bar duration
Consonant effects: CS=C×$F2$ separation

A model consistent with that proposed by Mermelstein (1978) could contain all the effects in Table II, except for the diphone bias terms, VC. The presence of the latter effects implies that vowel and consonant decisions are *not fully independent*. Nearey (1990) allows VC effects in what are termed diphone-biased segmental models. For the present experiment, inclusion of these terms results in a modest complication of the model requiring only an additional four degrees of freedom beyond a pure secondary-cue Mermelstein-like model.

There are, however, a number of effects (not shown in Table II) that Nearey’s (1990) hypothesis specifically disallows. These are stimulus-tuned diphone terms, represented by stimulus factors multiplying V×C diphone interactions. Thus a term like VCF would represent a vowel×consonant ×$F1$ interaction.

Some proposals in the literature imply just such interactions. Fisher and Ohde (1990) and Summers (1988) suggest that the effectiveness of $F1$ (both at its termination and during steady state) as a cue to a voicing may be influenced by the overall $F1$level of the preceding vowel. Thus, a small change in $F1$ frequency might have a larger effect on the relative probabilities of /t/–/d/ choices in /hɒ-/ contexts where the average $F1$ level is higher than in /hu-/ contexts where it is lower. Such cue-weight differences would lead to non-null V×C×$F1$ effects in the modeling scheme investigated here. Similarly, arguments by Kingston and Diehl (1994, 1995) imply the existence of noticeable V×C×voice bar duration interactions. These issues will be pursued in the analysis below.

2. Logistic regression analysis

Table III focuses on the relative goodness of fit of several models. It follows Nearey’s (1990) analysis of Whalen’s ‘‘bad–bet’’ experiment.[6] Column I gives a descriptive label and column II shows the effects included in the model or those added to the model in the previous row. Thus, model 3 includes all the terms in model 2 (V, C, VF, VS, CB, CD) plus the V×vocoid duration and C×$F1$ terms (VD, CF). Column III reports the deviance statistic, G^2. This is the ‘‘lack of fit’’ measure (the smaller the value, the better the fit) typically used in logistic and log linear modeling (McCullagh and Nelder, 1989). Column V reports rms error

TABLE III. Goodness-of-fit measures of selected logistic models for experiment I.

I Model label	II Effects included	III Deviance (G^2)	IV Residual degrees of freedom	V rms error	VI Reduction of residual deviance
1. Bias only model	V C	98 284.18	4315	21.96%	⋯
2. Primary cues segmental	V C VF VS CB CD	16 099.39	4305	9.73%	83.6%
3. Secondary cue	2+ VD CF	5943.07	4300	4.67%	63.1%
4. Diphone biased secondary cue	3+ VC	5687.45	4296	4.53%	4.3%
5. Diphone biased full cue	4+ VB CS	5659.82	4291	4.51%	0.5%
6. Full diphone	5+ VCF VCS VCD VCB	5273.08	4275	4.18%	6.8%

(in percentage points of response probability) of predicted versus observed values. Column VI reports the percentage of reduction residual deviance (column III) from the previous line of the table. Thus the change in deviance from the primary cue model of row 2 to the secondary cue model of row 3 represents a change of $(16\,099.39-5943.07)/16\,099.39=63.1\%$.

Model 1 serves to provide a baseline error measure. It is a random guessing model with no stimulus effects. It predicts a fixed probability of choosing each vowel and each consonant category regardless of the stimulus. Model 2 is the primary cue model, where consonants and vowels have nonoverlapping cues: $F1$ and $F2$ separation tune vowel decisions and voice bar duration and vocoid duration tune consonant decisions. The rms error rate has decreased by more than 12 percentage points and the residual deviance is reduced by more than 83%.

The move to the secondary cue model (model 3) reduces the absolute rms error to less than 5%. Here, in addition to the primary cues of the previous models, $F1$ is allowed to tune consonant decisions (CF) as well as vowels (VF) and vowel duration is allowed to tune vowel decisions (VD) as well as consonants (CD). This represents a purely separable Mermelstein-like model where consonant and vowel decisions are entirely independent, though $F1$ and vowel duration cues are shared. The addition of the secondary cue terms reduces the rms error by another five percentage points and reduces the residual deviance from model 2 by more than 63%. There is relatively little error left to explain, with only about 6% of the residual deviance from model 1 remaining.

The addition of the Nearey (1990) diphone bias terms VC in model 4 reduces the rms error by less than two tenths of a percentage point. The VC terms may represent ‘‘fudge factors’’ that might ‘‘boost’’ probabilities of categories with conflicting cues. They may also absorb lexical bias effects (some of the /hVC/ syllables are common words, some are not). The reduction in rms values is certainly not impressive. However, the model is already fitting rather well in absolute terms and the reduction of the residual deviance is a little more than 4% on only four additional degrees of freedom.

The addition of the two ‘‘minor cues’’ in model 5 (V ×voice bar duration and C×$F2$ separation) improves the fit very little, less than 1% additional deviance being accounted for with five additional degrees of freedom in the model.

Finally, the move from model 5 to the (complete diphone) model 6 results in a reduction of less than four tenths of a percentage point in the rms error. This model allows each diphone choice to find an arbitrary weight on each of the stimulus properties. Thus, it would be possible, if the patterns in the data warranted, for the model to attach large weights to $F1$ for the /ɒt/–/ɒd/ distinction while attaching a zero weight to $F1$ for the /ut/–/ud/ distinction. This additional flexibility contributes about a 7% reduction of the residual deviance. This seems a relatively modest gain given the 16 additional degrees of freedom used by the model.

3. Statistical hypothesis tests

Two types of statistical analysis are described below. The first involves quasilikelihood procedure described by McCullagh and Nelder (1989) and used by Nearey (1990). This focuses on the change in goodness-of-fit of successive models in Table III.[7] According to this procedure, the improvement from model 1 to 2 is highly significant [$F(10,4305)=6265.9,\ p<0.000\,01$], as is that from model 2 to 3 [$F(5,4300)=212.36,\ p<0.000\,01$] and that from model 3 to 4 [$F(4,4296)=6.475,\ p<0.000\,04$]. The addition of the minor cue terms in model 5 is not significant [$F(5,4291)=0.552,\ p<0.5$]. However, the addition of the stimulus-tuned diphone terms in model 6 is significant [$F(16,4275)=2.281,\ p<0.003$].

The second set of tests assesses the reliability of logistic coefficients across subjects. The most complex model (model 6) is first fit to the data of each subject individually. The coefficients from these analyses are then subjected to the second-stage multivariate tests described in Gumpertz and Pantula (1989).[8] The results of these tests are summarized in Table IV.

The general pattern of results agrees with the quasilikelihood analysis. The primary cue (VF, VS, CB, CD) and secondary cue (CF, VD) effects are all highly significant. The minor cue effects (VB and CS) are not significant. The diphone bias effects (VC) are significant. Three of the four diphone-tuned stimulus effects (VCF, VCD, and VCB) are

TABLE IV. Random effects F tests for families in model 6 of Table III.

	Family	F	df	p
1	V	87.199	4, 11	0.000 00
2	C	58.227	1, 14	0.000 00
3	VC	6.783	4, 11	0.005 27
4	VF	102.639	4, 11	0.000 00
5	CF	91.704	1, 14	0.000 00
6	VCF	6.899	4, 11	0.004 95
7	VS	79.881	4, 11	0.000 00
8	CS	1.434	1, 14	0.251 04
9	VCS	1.468	4, 11	0.277 23
10	VD	48.427	4, 11	0.000 00
11	CD	76.086	1, 14	0.000 00
12	VCD	4.261	4, 11	0.025 25
13	VB	1.705	4, 11	0.218 72
14	CB	106.104	1, 14	0.000 00
15	VCB	8.646	4, 11	0.002 08

significant, while a fourth (VCS) is not.[9] The presence of reliable diphone bias effects is not surprising (Nearey, 1990) and has in fact been ‘‘reconciled’’ within double-weak theory (Nearey, 1992). The significance of diphone-tuned stimulus effects is another matter and deserves further attention.

4. Graphic analysis of logistic coefficients

Considerable insight into the nature of stimulus-response relations can be obtained by a study of coefficients of a logistic model using diphones as the basic symbolic unit. The coefficients from this model can be displayed in a manner similar to interaction plots in analysis of variance (ANOVA), as shown in Figs. 5–9. The y axis shows the value of the coefficients. The x axis represents the vowel categories roughly in order of their IPA vowel height. The two lines represent /-Vt/ and /-Vd/ syllables. Note that if consonant and vowel choices were totally independent as in Mermelstein’s original proposal, *all* of these plots would show only parallel lines. Nearey’s (1990) diphone biased models require parallel lines in all panels except Fig. 5, which displays bias coefficients.[10]

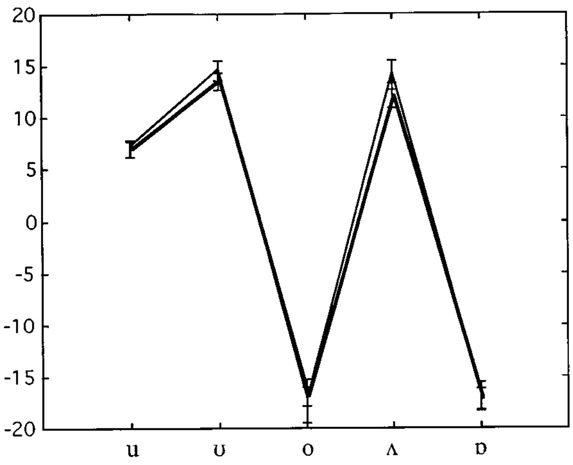

FIG. 6. Mean $F2$ separation coefficients (and standard errors across listeners) corresponding to model 6 of Table III.

For Figs. 6–9, the ordering of the coefficient values roughly reflects the expected ordering of the categories along the stimulus dimension in question. Thus, in Fig. 8 the coefficients associated with the vowel categories increase roughly in order of expected $F1$ associated with the categories. Similarly, the coefficients for /-Vt/ categories and /-Vd/ are ordered in the expected way (higher values associated with /-Vt/).

Figure 6 shows that the $F2$-separation value has essentially no effect on consonant judgments, while the vowels differ substantially. This agrees with the statistical results in rows 7–9 of Table IV, where only VS effects are significant, while VC and VCS are not. The pattern of coefficients indicates that vowels /o/ and /ɒ/ are favored by small $F2-F1$ differences and the others by larger separations. These patterns are roughly in line with production data patterns (Nearey and Assmann, 1986).

Figure 7 shows that vowel duration coefficients are noticeably different for both consonants and vowels. This accords with the significance of the VD and CD terms in rows 10–12 of Table IV. These lines appear to be roughly parallel.

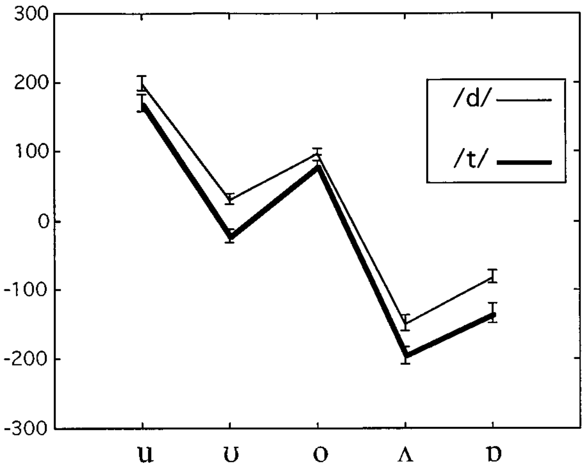

FIG. 5. Mean bias coefficients (and standard errors across listeners) corresponding to model 6 of Table III.

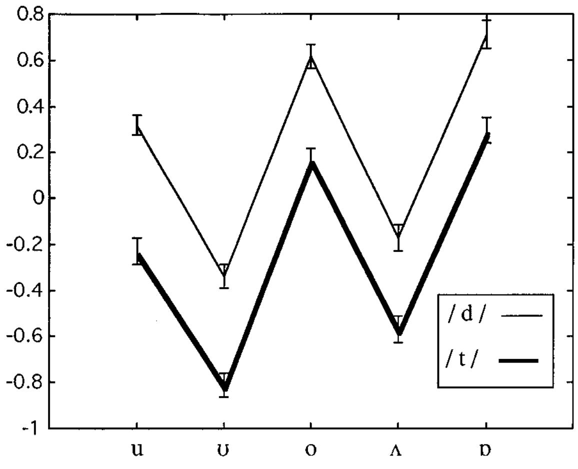

FIG. 7. Mean vocoid duration coefficients (and standard errors across listeners) corresponding to model 6 of Table III.

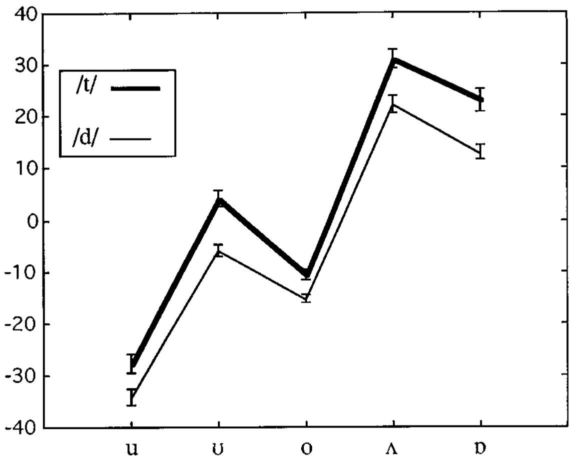

FIG. 8. Mean $F1$ coefficients (and standard errors across listeners) corresponding to model 6 of Table III.

This has the same interpretation as in similar ANOVA means plots: There are stimulus-tuned main effects for vowels and for consonants (corresponding to VD and CD terms in the models discussed above) but little evidence of important VC interactions. Nevertheless, the VCD term is nominally significant in Table IV. There appears to be a slight trend for wider separations of the /-t/ and /-d/ lines for vowels associated with lower $F1$'s. Since the deviations from parallelism are quite small and since there is no clear basis from alternate theories for the pattern observed, they will not be discussed further.[11]

On the other hand, the VCF interaction in Table IV is significant and the $F1$-tuned effects in Fig. 8 show reasonably clear visual evidence for lack of parallelism of the /-d/ and /-t/ lines. Furthermore, there is also some discussion in the literature relevant to interactions. It has been suggested by Fisher and Ohde (1990) and by Summers (1988) that the effectiveness of $F1$ as a cue to voicing may be influenced by $F1$ level. Thus for high $F1$vowels, variation in $F1$ within vowels might exert a fairly large influence on consonant judgment, compared to low $F1$ vowels. In Fig. 8, the effectiveness of $F1$ in separating the consonants for different vowel choices is reflected by the separation of the corresponding /-Vt/ and /-Vd/ points. If attention is limited to only the lowest and highest $F1$ vowels (/u/ and /ɒ/), one might conclude that there is some support for Fisher and Ohde's hypothesis. However, the overall pattern indicates that $F1$ is relatively less effective for the vowels /u/ and /o/, while /ʊ/ [which has comparable $F1$ characteristics to /o/ in this dialect; see Nearey and Assmann (1986)] shows a larger difference, roughly comparable to the two highest $F1$ vowels. It is possible that some more complex interaction is taking place, e.g., that when both $F1$ and $F2$ are relatively low, $F1$ has a smaller effect. However, I know of no theoretical basis for such a finding. Evidence for lack of parallelism is still not particularly strong ($F1$ noticeably affects both vowel and consonant choice throughout) and it would seem difficult to argue that the relatively subtle deviations have great theoretical import without clearer support from other experiments.

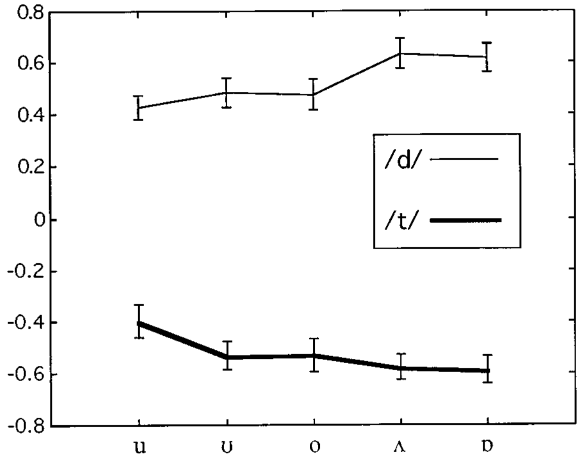

FIG. 9. Mean voice bar duration coefficients (and standard errors across listeners) corresponding to model 6 of Table III.

Finally, consider Fig. 9, which displays the effect of voice bar. A potential source of diphone interaction for this case can be derived from an analysis of VCV's by Kingston and Diehl (1995). They suggest that voice bar affects voicing judgments by contributing to the low frequency property in all $F1$ contexts. However, they also argue that the [±voice] feature is implemented in English by both the low-frequency property and by V:C ratio. Consequently, the relation of this property to voicing judgments should be nonlinearly affected by $F1$ offset. When $F1$ offsets are low and spectrally contiguous with voice bar, the perceived V:C ratio is increased because the apparent gap duration is shorter (and the apparent vowel duration longer). This implies that consonant decisions should be more strongly affected by voice bar for low-$F1$ vowels like /u/ (where voice bar contributes to both the low-frequency property and to V:C ratio) than for the high $F1$ vowels like /ɒ/. Such differential effectiveness of voice bar duration could be expected show up in Fig. 9 as a wider separation of the /-Vt/ and /-Vd/ lines for the low $F1$ vowels and narrower separation for the higher $F1$ vowels. In fact, a slight trend in the opposite direction is found. Again, I know of no theory that would predict such a pattern.[12]

In sum, it is important to note that there are some statistical and graphic indications that Nearey's (1990) diphone-biased segmental models are not entirely adequate to the data from this experiment. Nevertheless, the overall goodness-of-fit of a straightforward extension of Mermelstein's (1978) model is remarkably good (model 3 in Table III, with about 5% rms error). The degree of divergence from parallelism in Figs. 6–9 is modest and the direction of deviations noted does not appear to relate very well to any previously articulated theory.[13]

IV. GENERAL DISCUSSION

A. Pattern-recognition models for more complex data

Every language learner is faced with a *fait accompli* of what Kuhl (1992) has aptly called ''the ambient language.'' A key task for the language learner is the construction of an appropriate pattern-recognition scheme for the problem at hand. Analysis of the stimulus-to-response mapping for mature listeners in laboratory studies suggests that the computational complexity of the adult recognition scheme is rather stringently constrained. If transmission of symbolic information from speaker to hearer is to approach optimality (i.e., if speakers produce patterns that can usually be recognized), similarly stringent constraints must apply to output patterns. The models discussed thus far in this paper are all linear in their stimulus effects. This implies that the decision regions for each category in territorial maps (e.g., Fig. 3) are linearly separable, i.e., delimited by straight lines.[14] While many speech patterns may be compatible with this restriction, there appears to be at least one notable exception. This involves the cue patterns associated with place of articulation in stop consonants.

Nearey and Shammass (1987) published work on phenomena that are now referred to as ''locus-equations'' (Sussman *et al.*, 1995; Sussman *et al.*, 1993; Sussman *et al.*, 1991). We measured $F2$ and $F3$ of voiced stop+vowel syllables as early as possible after the stop burst ($F2_i$ and $F3_i$) and at 60 ms into the vocalic portion ($F2_v$ and $F3_v$) of the syllable. Results from 10 speakers showed that formant frequencies clustered quite closely about a regression line for each consonant. However, for $F2$ the /d/ line crosses both of the others, clearly indicating a lack of linear separability. Furthermore, statistical tests showed that the patterns did not have equal covariance matrices for the three consonants and that quadratic, rather than linear, discriminant methods were called for. Using these, we obtained approximately 72% correct partition of the training data into the three classes.[15] Although such recognition models involve territorial maps more complex than those required for experiment I, it is important to note that classification takes place without reference to stimulus characteristics of specific CV diphones. Rather, for each consonant a single mean vector and a single covariance matrix are estimated *without regard to the phonological identity of the vowel*.

Sussman *et al.* (1993) have studied locus equations in a number of different languages, finding reliable linear patterns within each language. They have also shown that language-specific differences associated with secondary-articulation factors, such as pharyngealization, have reliable effects on these patterns. Sussman and colleagues have further argued that the resulting linear patterns are not simply an accident of production, but rather that they likely represent a deliberate ''steering'' of coarticulation effects to produce well-defined covariance patterns. [See also Nearey (1992).]

The degree of statistical separability of the three stop classes from information represented in locus equations is clearly not enough to account for very high identification rates for stop+vowel syllables by humans. However, other cues, including the shape of release spectra, are known to provide substantial information about place for stops (Stevens and Blumstein, 1981; Lahiri *et al.*, 1984; Forrest *et al.*, 1988). Nossair and Zahorian (1991) have demonstrated a speaker-independent automatic recognition system that achieves very high identification rates (93.7% correct on an independent test set) for the six English stops in 11 vowel environments on 15 speakers (males, females, and children). Nossair and Zahorian used quadratic discrimination methods (based on separate mean vectors and separate covariance matrices for each consonant) using measures derived from the first 60 ms of stop+vowel+C syllables. The performance of the best algorithms they studied approached that of a panel of human listeners (96.6%) on the same syllables. (Listeners showed performance of only 89.9% when they heard only the first 50 ms of the stimuli, a situation more comparable to the computer modeling.) There was a reasonably good correspondence between error matrices for listeners and Nossair and Zahorian's best pattern recognizer when errors were pooled over speakers and vowels. There was also some correlation of error rates for individual talkers.

B. Strong and weak theories of phonological contrast

Proponents of various strong theories might well object that the complexity of the signal representations and relatively large number of parameters required in Nossair and Zahorian's model makes them unlikely candidates as perceptual models. Nevertheless, their results are compatible with an extended version of Sussman's *orderly output conditions*. Their algorithm effectively establishes an upper bound on the computational complexity required for a creditable solution of this long long-standing perceptual puzzle.

The presence of such an upper bound should encourage the search for a better understanding of the patterns in terms of more primitive phonological and psychophysical elements, including classical distinctive features and Kingston and Diehl's intermediate perceptual properties. However, we should also consider the possibility that acquisition of phonological contrasts involves extensive auditory-perceptual learning. Listeners may effectively gather statistics about the auditory signatures of phonological symbols from the ambient language. It is not beyond imagining that mastery of such patterns implies the equivalent of learning mean and covariance patterns for a few dozen symbols in a space of a few dozen perceptual parameters.

The degree of perceptual learning and the complexity of cue structures implied by such a program may seem wrongheaded to advocates of strong theories. Double-strong theorists clearly favor temporally localized, quantal, universal cues. Yet some definitions of features, such as the ''low frequency property'' (coined by Blumstein and Stevens), do take account of some lack of absolute synchronization. Furthermore, some cues proposed by Lahiri *et al.* (1984) to separate labials from coronals involve relatively complex relational properties of temporally distributed spectral patterns, information that seems hardly more restricted in scope than that used in the locus-equations models. Auditorists like Kingston and Diehl (1994) also acknowledge temporally distributed cue information involving overtly relational properties. However, they favor an inventory of universal derived auditory properties which are functions of multiple subproperties which stand in an ''interlocking network of mutual enhancement relations'' to the derived properties.

Both auditorists and double-strong theorists allow for some measure of perceptual learning (see, e.g., Stevens and Blumstein, 1981) and the differences among the alternatives might best be viewed as a matter of degree. Dealing with a limited inventory of universal distinctive features or of intermediate perceptual properties would pose no problems in principle for the kind of perceptual modeling I am advocating. Indeed, if they could be established firmly, they would aid perceptual modeling by limiting the degrees of freedom of the cue weights or statistical distributions to be estimated. However, a premature commitment to a limited set of properties may jeopardize modeling accuracy and, in a rush to universals, we may miss some very important facts about how language works.

In light of this danger, it seems appropriate that at least some of us pursue what may seem to strong-theorists to be a crassly empirical approach to phonological contrast. Yet, something similar has been already been advocated by Kohler (1981) and by Ladefoged (1990; Lindau and Ladefoged, 1986). There is considerable evidence from cross-language studies of production (see, e.g., Lindau and Ladefoged, 1986; Ladefoged, 1990; Sussman *et al.*, 1993) that phonetic patterns of language differ from each other in subtle and complex ways. There is also evidence from perception that speakers of different languages differ in the weights they assign to different cues (Crowther and Mann, 1992; Munro, 1992; Kuhl *et al.*, 1992; Kohler, 1981; Flege *et al.*, 1994; Scholes, 1967). Until a stronger theory is able to present a far more compelling case than any has so far, it may be prudent to take such differences at their face value, allowing language-specific patterns to be incorporated directly into perceptual models of phonological contrast. In due course, the language-specific empirical generalizations summarized by such models may serve as grist for the mill of a stronger and more convincing theoretical approach yet to be discovered.

ACKNOWLEDGMENTS

Portions of this paper were presented at the 129th meeting of the Acoustical Society of America, June 1995 in Washington DC. This work was supported by a grant from SSHRC.

[1] See Nearey (1995) and Massaro and Oden (1980) for discussion of this assumption. A number of other important assumptions are also being made but not defended. Two of the more salient are: (1) The segmentation of the input signals and measurement of stimulus properties before they are presented to a perceptual model. (2) What we learn in the laboratory generalizes to more complex speech communication situations.

[2] For a more general discussion of the role of articulation in speech, see McGowan and Faber (1996) and related papers in the same issue.

[3] Lindblom's (1990) research on hypospeech suggests that a substantial part of the variation associated with weaker prosodic contexts may represent *true information loss* at the phonetic level that can be tolerated because of higher-level redundancies. See also Allen (1994).

[4] The distinction between ''primary cue'' and ''secondary cue'' is probably just a matter of degree. Cues with higher weights are relatively primary.

[5] $F1$ offset was not varied as a separate factor because of the large number of stimuli involved. Compromise values were considered in initial planning stages, but stimuli without final $F1$ transitions sound quite acceptable (although a weak thump can be heard in stimuli with high $F1$'s if one listens carefully). Fortuitously, these stimuli are very well suited for comparison with some of Kingston and Diehl's observations about possible interactions of voice bar with $F1$ level. See Nearey (1995) and note 12.

[6] Stimulus dimensions were transformed by taking logs of frequencies and square roots of durations before analysis. Experience with data analysis for vowels indicates that log frequency works about as well as any other proposed tonotopic scales (e.g., ERB, Bark), all of which work better than linear (Hz) measures. Square roots of durations were also used by Nearey (1990) for analysis of Whalen's data because all models considered showed substantially better fit when this was done. Transformations of raw treatment values are routinely used to improve fit in probit analysis, where they are called ''dose metameters'' (Finney, 1971).

[7] The quasilikelihood tests involve F ratios using deviance statistics from Table III and a heterogeneity coefficient estimated from Pearson X^2 statistics. The tests of increase of goodness-of-fit are analogous to ''extra sum of square tests'' in regression models. Some caution seems in order in interpreting the significance levels, since this approach makes rather restrictive assumptions about components of error that may not be appropriate for repeated measures data.

[8] Families of coefficients associated with the first column of Table IV are tested separately against the null hypothesis that each listener's coefficients are drawn from a multivariant normal population with a zero mean. The assumptions of these relatively simple tests appear to be plausible. However, newer, more complex techniques for analyzing generalized linear mixed models (GLMM's) may prove to be more powerful. Although GLMM's have received considerable attention in the recent statistics literature (e.g., Breslow and Clayton, 1993), no software using GLMM techniques capable of handling a problem the size of experiment I appears to be available.

[9] A conservative approach to the testing of multiple families employing Bonferroni correction would set the per-family significance level at 0.003 for an experiment-wise error rate of 0.05. By this conservative criterion, only the VCB interaction and the primary (VF, VS, CD, CB) and secondary cue (VD, FS) terms reach significance.

[10] The diphone bias effects are not discussed here. Diphone biases in model 6 (or any other model that includes V×C×stimulus interactions) are not interpretable for the same reasons that group mean differences are not interpretable in analysis of covariance when there are separate slopes allowed for distinct groups (Snedecor and Cochran, 1967).

[11] There may be some relation to Kingston and Diehl's (1995) suggestion that lower $F1$'s can affect apparent gap duration when voice bar is present, but it is not clear to me how this would be reflected as VCD effects in the current analysis.

[12] Kingston and Diehl (personal communication) have provided me with data from an experiment they devised to replicate a subset of the stimuli described here. Analyses in both our laboratories agree that their experiment *does* show evidence for interaction patterns of the type they predict. The reason for the discrepancies is currently under investigation at both sites.

[13] There are a number of simplifying assumptions involved in the specific subclass of logistic regression models used here. If incorrect, these could result in lack of fit and could possibly lead to spurious diphone interactions analogous to what is sometimes called ''removable nonadditivity'' in ANOVA models. For example, the log of frequency and square root of duration transformations were chosen in advance of the analysis. Exploration of some other choices of metameter (see note 6) indicated that a linear time scale (ms) for voice bar duration leads to a slightly better overall fit for model 6. In this case, the VCB interaction, which showed the largest F-ratio of all the stimulus-tuned diphone terms in Table IV, was no longer significant by a random coefficients regression test $[F(4,11)=2.578, p<0.09$. In addition, though still significant, the F-ratios for the other two significant stimulus-by-diphone interactions were also decreased: for VCD, $F(4,11)=3.715$, $p<0.04$; for VCF, $F(4,11)=6.311$, $p<0.007$. The VCS interaction remained nonsignificant, $F(4,11)=1.481$, $p<0.27]$.

[14] A set of output conditions sufficient to allow *optimal* recognition by diphone biased secondary cue models with linear boundaries [e.g., Fig. 3(c)] is discussed in Nearey (1992).

[15] Quadratic boundaries for perceptual data can be studied with logistic regression by including square and cross products of stimulus effects as additional variables.

Allen, J. (**1994**). ''How do humans process and recognize speech?,'' IEEE Trans. Speech Audio Process. **2,** 567–577.

Blumstein, S. (**1986**). ''Comment on Lindau and Ladefoged,'' in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillside, NJ), pp. 465–478.

Blumstein, S., and Stevens, K. (**1980**). ''Perceptual invariance and onset spectra for stop consonants in different vowel environments,'' J. Acoust. Soc. Am. **67,** 648–662.

Breslow, N. E., and Clayton, D. G. (**1983**). ‘‘Approximate inference in generalized linear mixed models,’’ J. Am. Stat. Assoc. **88**, 9–25.

Chomsky, N., and Halle, M. (**1968**). *The Sound Pattern of English* (Harper, New York).

Crowther, C., and Mann, V. (**1992**). ‘‘Native language factors affecting use of vocalic cues to final consonant voicing in English,’’ J. Acoust. Soc. Am. **82,** 711–722.

Diehl, R. L., and Kluender, K. R. (**1989**). ‘‘On the objects of speech perception,’’ Ecol. Psychol. **1**, 121–144.

Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Finney, D. J. (**1971**). *Probit Analysis* (Cambridge U.P., Cambridge, England), 3rd. ed.

Fisher, R. M., and Ohde, R. N. (**1990**). ‘‘Spectral and duration properties of front vowels as cues to final stop-consonant voicing,’’ J. Acoust. Soc. Am. **88,** 1250–1259.

Flege, J. E., Munro, M. J., and Fox, R. A. (**1994**). ‘‘Auditory and categorical effects on cross-language vowel perception,’’ J. Acoust. Soc. Am. **95,** 3623–3641.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. (**1988**). ‘‘Statistical analysis of word-initial voiceless obstruents: Preliminary data,’’ J. Acoust. Soc. Am. **84,** 115–123.

Fowler, C. (**1989**). ‘‘Real objects of speech perception: a commentary on Diehl and Kluender,’’ Ecol. Psychol. **1**, 145–169.

Gumpertz, M., and Pantula, S. (**1989**). ‘‘A simple approach to inference in random coefficient models,’’ Am. Stat. **43**, 203–210.

Haberman, S. J. (**1979**). *Analysis of Qualitative Data, Volume 2* (Academic, New York).

Jakobson, R., Fant, G., and Halle, M. (**1963**). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates* (MIT, Cambridge, MA).

Jamieson, D. G., Nearey, T. M., and Ramji, K. (**1989**). ‘‘CSRE: a speech research environment,’’ Can. Acoust. **17**, 23–25.

Johnson, K., Ladefoged, P., and Lindau, M. (**1993**). ‘‘Individual differences in vowel production,’’ J. Acoust. Soc. Am. **94,** 701–714.

Kewley-Port, D., Watson, C., and Foyle, D. (**1988**). ‘‘Auditory temporal acuity in relation to category boundaries; speech and nonspeech stimuli,’’ J. Acoust. Soc. Am. **83,** 1133–1145.

Kingston, J., and Diehl, R. (**1994**). ‘‘Phonetic knowledge,’’ Language **70**, 419–454.

Kingston, J., and Diehl, R. (**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, England), pp. 7–27.

Kingston, J., and Diehl, R. (**1996**). Personal communication.

Klatt, D. (**1981**). ‘‘Software for a cascade/parallel formant synthesizer,’’ J. Acoust. Soc. Am. **67,** 971–995.

Kohler, K. J. (**1981**). ‘‘Contrastive phonology and the acquisition of phonetic skills,’’ Phonetica **38**, 213–226.

Kohler, K. J. (**1984**). ‘‘Phonetic explanation in phonology: the feature fortis/lenis,’’ Phonetica **41**, 150–174.

Kuhl, P. (**1992**). ‘‘Infants’ perception and representation of speech: development of a new theory,’’ in *Proceedings ICSLP 92*, edited by J. Ohala, T. Nearey, B. Derwing, M. Hodge, and G. Wiebe (University of Alberta, Edmonton), pp. 449–455.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (**1992**). ‘‘Linguistic experience alters phonetic perception in infants by 6 months of age,’’ Science **255,** 606–608.

Labov, W. (**1972**). *Sociolinguistic Patterns* (University of Pennsylvania, Philadelphia).

Ladefoged, P. (**1990**). ‘‘Some reflections on the IPA,’’ J. Phon. **18**, 335–346.

Ladefoged, P., DeClerk, J., Lindau, M., and Papçun, G. (**1972**). ‘‘An auditory-motor theory of speech production,’’ UCLA Working Papers in Phonetics **22**, 48–75.

Lahiri, A., Gewrith, L., and Blumstein, S. (**1984**). ‘‘A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study,’’ J. Acoust. Soc. Am. **76,** 391–404.

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

Lindau, M., and Ladefoged, P. (**1986**). ‘‘Variability of feature specifications,’’ in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 465–478.

Lindblom, B. (**1986**). ‘‘Phonetic universals in vowel systems,’’ in *Experimental Phonology*, edited by J. Ohala and J. Jaeger (Academic Press, Orlando).

Lindblom, B. (**1990**). ‘‘Explaining phonetic variation: a sketch of the H and H theory,’’ in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Ḿarchal (Kluwer, Amsterdam), pp. 403–439.

Martinet, A. (**1955**). *Économie des changements phonétiques* (A. Francke, Bern).

Massaro, D., and Cohen, M. (**1983**). ‘‘Phonological context in speech perception,’’ Percept. Psychophys. **34,** 338–348.

Massaro, D., and Oden, G. (**1980**). ‘‘Evaluation and integration of acoustic features in speech perception,’’ J. Acoust. Soc. Am. **67,** 996–1013.

McCullagh, P., and Nelder, J. A. (**1989**). *Generalized Linear Models* (Chapman and Hall, London).

McGowan, R. S., and Faber, A. (**1996**). ‘‘Introduction to papers on speech recognition and perception from an articulatory point of view,’’ J. Acoust. Soc. Am. **99,** 1680–1682.

Mermelstein, P. (**1978**). ‘‘On the relationship between vowel and consonant identification when cued by the same acoustic information,’’ Percept. Psychophys. **23,** 331–335.

Munro, M. J. (**1992**). ‘‘Perception and production of English vowels by native speakers of Arabic,’’ Ph.D. thesis, University of Alberta.

Nearey, T. M., (**1980**). ‘‘On the physical interpretation of vowel quality: cinefluorographic and acoustic evidence,’’ J. Phonetics **8**, 213–241.

Nearey, T. M. (**1990**). ‘‘The segment as a unit of speech perception,’’ J. Phon. **18**, 347–373.

Nearey, T. M. (**1991**). ‘‘Perception: Automatic and cognitive processes,’’ in *Proceedings of the XII International Congress of Phonetic Sciences* (Publications de L'Université de Provence, Aix-en-Provence), Vol. 1, pp. 40–49.

Nearey, T. M. (**1992**). ‘‘Context effects in a double-weak theory of speech perception,’’ Lang. Speech **35**, 153–172.

Nearey, T. M. (**1995**). ‘‘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, England), pp. 28–40.

Nearey, T., and Assmann, P. (**1986**). ‘‘Modeling the role of inherent spectral change in vowel identification,’’ J. Acoust. Soc. Am. **80,** 1297–1308.

Nearey, T., and Shammass, S. (**1987**). ‘‘Formant transitions as partly distinctive invariant properties in the identification of voiced stops,’’ Can. Acoust. **15**, 17–24.

Nossair, Z. B., and Zahorian, S. A. (**1991**). ‘‘Dynamic spectral shape features as acoustic correlates for initial stop consonants,’’ J. Acoust. Soc. Am. **89,** 2978.

Ohala, J. (**1981**). ‘‘The listener as a source of sound change,’’ in *Papers from the Parasession on Language and Behavior*, edited by C. S. Masek, R. A. Hendrick, and M. F. Miller (Chicago Linguistic Society, Chicago), pp. 178–203.

Ohala, J., and Shriberg, E. (**1990**). ‘‘Hypercorrection in speech perception,’’ in *Proceedings of the 1990 International Conference on Spoken Language Processing* (Acoustical Society of Japan, Kobe), pp. 405–407.

Perkell, J. S., Mathies, M. L., Svirsky, M. A., and Jordan, M. (**1993**). ‘‘Trading relations between tongue-body raising and lip rounding in production of the vowel /u/,’’ J. Acoust. Soc. Am. **93,** 2948–2961.

Peterson, G. E., and Lehiste, I. (**1960**). ‘‘Duration of syllable nuclei in English,’’ J. Acoust. Soc. Am. **32,** 693–703.

Repp, B. (**1982**). ‘‘Phonetic trading relations and contexts effects: new evidence for a phonetic mode of perception,’’ Psychol. Bull. **92,** 81–110.

Scholes, R. (**1967**). ‘‘Phoneme categorization of synthetic vocalic stimuli by speakers of Japanese, Spanish, Persian and American English,’’ Lang. Speech **10**, 46–68.

Snedecor, G. W., and Cochran, W. G. (**1967**). *Statistical Methods* (Iowa State U. P., Ames, IA).

Stevens, K. (**1990**). ‘‘On the quantal nature of speech,’’ J. Phon. **17**, 3–45.

Stevens, K. N., and Blumstein, S. (**1981**). ‘‘The search for invariant acoustic correlates of phonetic features,’’ in *Perspectives on the Study of Speech*, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, NJ), pp. 1–38.

Stevens, K. N., and Keyser, S. J. (**1989**). ‘‘Primary features and their enhancement in consonants,’’ Language **65**, 81–106.

Summers, W. V. (**1987**). ‘‘Effects of stress and final-consonant voicing on vowel production,’’ J. Acoust. Soc. Am. **82,** 847–863.

Summers, W. V. (**1988**). ‘‘ $F1$ structure provides information for final-consonant voicing,’’ J. Acoust. Soc. Am. **84,** 485–492.

Sussman, H. M., Hoemeke, K. A., and Ahmed, F. S. (**1993**). ‘‘A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation,'' J. Acoust. Soc. Am. **94,** 1256–1268.

Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (**1991**). ''An investigation of locus equations as a source of relational invariance for stop place categorization,'' J. Acoust. Soc. Am. **90,** 1256–1268.

Sussman, H., Fruchter, D., and Cable, A. (**1995**). ''Locus equations derived from compensatory articulation,'' J. Acoust. Soc. Am. **97,** 3112–3124.

Watson, C. J., and Foyle, D. C. (**1985**). ''Central factors in the discrimination and identification of complex sounds,'' J. Acoust. Soc. Am. **78,** 375–379.

Watson, C., and Kewley-Port, D. (**1988**). ''Some remarks on Pastore (1988),'' J. Acoust. Soc. Am. **84,** 2266–2270.

Whalen, D. (**1989**). ''Vowel and consonant judgments are not independent when cued by the same information,'' Percept. Psychophys. **46,** 284–292.