PERCEPTION: AUTOMATIC AND COGNITIVE PROCESSES

Terrance M. Nearey

University of Alberta, Edmonton AB, Canada, T6G 2E7

ABSTRACT

Speech perception is that process by which humans map acoustic waveforms onto strings of linguistic symbols. While accepting many of the premises about the complexity of signal-to-symbol mapping that have been so influential in Liberman and Mattingly's [9] motor theory of speech perception, it is argued that there exists an upper limit on that complexity imposed by perceptual mechanisms that map acoustic properties directly onto phonological units [15]. Evidence for this claim is presented together with its implications for the nature of cognitive processes in speech perception and their relative automaticity.

1. INTRODUCTION

The terms "automatic" versus "cognitive" may be seen to relate to a continuum of computational complexity. *Automatic* processes are likely to be associated with simple architectures with few free parameters, with a constrained, largely bottom-up data flow and with an overall "reflex-like" character. *Cognitive* processes may have a complex architecture with many free parameters, may possess a less constrained, strongly "top-down" data flow and may exhibit relatively "intelligent" behavior. These concepts are closely related to aspects of Fodor's modularity hypothesis [5].

Fodor proposes that there exists a highly flexible central processor representing the pinnacle of cognition. The central processor serves as a kind of "executive Sherlock", brilliantly integrating and evaluating information from a variety of sources. However, Sherlock's data doesn't come directly from the world at large; rather, it is bureaucratically passed upstream by a set of clever but narrow minded "forensic specialists", the input modules, who preprocess raw input in highly stylized ("work-to-rule") ways.

Fodor postulates the following sobering, if tongue-in-cheek, first law of cognitive science: "The more global (i.e., isotropic) a cognitive process is, the less anybody understands it." Isotropic is a term borrowed from the philosophy of science, meaning "facts relevant to the confirmation of a scientific hypothesis may be drawn from anywhere in the field of previously established truths." In Fodor's scheme, isotropism is a property only of the central processor, while input modules are much more constrained in their operation. Furthermore, he contends, it is precisely because they are of limited cognitive capacitiy that we are able to understand them at all. Input modules are viewed as computationally complex but specialized "cognitive reflexes" that constitute, in part, "the means whereby stupid processing systems manage to behave as though they were smart ones (p. 81)."

1.1 Motor (Gestural) Theories

Liberman and Mattingly (= "LM" [9]) adopt an overtly modularist perspective and they martial a wide variety of arguments in support of a special phonetic decoder as an input system in the Fodorian sense. Although their other arguments are important, I will be concerned only with the problem of the signal-to-symbol mapping, that is, to listeners' categorization of speech.

LM state their main premise as follows: "The first claim [of the motor theory] is that the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands, that call for movements of the articulators through certain linguistically significant configurations (p. 2)." They continue: "But the relationship between the gesture and the signal is not straightforward. The reason is that the timing of articulatory movements --the peripheral realizations of the gestures-- is not simply related to the ordering of gestures that is implied by the strings of symbols in phonetic transcription (p. 3)." Thus, in this framework, an articulatory space is essential in understanding the signal-to-symbol mapping.

To fix ideas, consider an example from Cooper et al. [3] involving classification of voiceless stop+vowel stimuli. An approximation of the decision space for this experiment is given in Fig. 1. It shows dominant response regions for the three voiceless stops /p/, /t/ and /k/ for a stimulus space consisting of a narrow band noise burst (characteristic of plosive release) followed by two-formant synthesized vowels. The authors emphasize that there are no absolutely invariant properties associated with perception of the stops, but only fairly complex relational properties. Such complexity, they contend, "requires the consonant-vowel combination as a minimal acoustic unit (p. 598)." The motor theory claims such intricate acoustic-to-phonetic patterns can be understood only by reference to the Rosetta stone of the underlying gestures. i.e., in this case, the coarticulation of a stop with its following vowel.

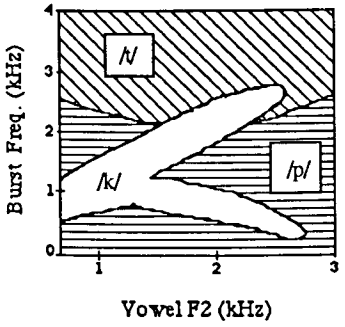

Figure. 1. Categorization of stop consonants (after [3]).

Fowler [6] and her colleagues also argue for the close relation between perception and articulation, but propose Gibsonian "direct perception" of the *actual* articulatory gestures of the moving vocal tract. Liberman and Mattingly are highly skeptical of this possibility, noting that, e.g., "given the many-to-one relation between vocal-tract configurations and acoustic signal, a purely analytic solution to the problem of recovering movements from the signal seems to be impossible." LM believe that it is "phonetic intentions", rather than actual peripheral events that count and they advocate analysis-by-synthesis (=ABS) decoding, whereby the phonetic module "has merely to determine which (if any) of a small number of gestures that might have been initiated at a particular instant could, in combination with gestures already in progress, account for the signal (p. 27)."

Klatt [8] agrees with LM's pessimistic assessment of the possibility of the recoverability of gesture from the acoustic waveform. But, having actually explored ABS for automatic speech recognition, he seems to conclude that it too is computationally intractable. Klatt nonetheless believes that "[p]roduction and perception are clearly closely tied in the sense that perceptual strategies must know a great deal about production options and their acoustic manifestations (p.178)." Though he voices hope for a more scientifically appealing approach in the future, he suggests that the most promising way to deal with context dependency is to pre-compile acoustic patterns of words in large networks that make no use of traditional phone-sized units.

The strong motor theory position of LM is best motivated in the context of a "full speed ahead and damn the torpedoes" model of speech production, as stereotyped in Hockett's "soft-boiled Easter egg plus wringer model." As MacNeilage [12] points out, this charicature bears a striking resemblence to actual models of speech production of the late 60's. MacNeilage's (and much subsequent) work has shown this view to be very mistaken. Rather, despite the persistence of residual peripheral variability, the motor system is capable of performing rather remarkable feats in

achieving relatively invariant peripheral manifestations of articulatory targets.

1.2 Auditory Theories

As the peripheral configurations associated with phonetic elements approach invariant targets, so to do their acoustic consequences. Many researchers believe that acoustic/auditory properties have a direct role in defining goals for speech production and perception. Blumstein and Stevens [1] and their colleagues argue for relatively invariant signal properties that actually motivate target articulations for stop consonants. Nearey [18] claims that acoustic properties of vowels exhibit demonstrably greater invariance across speakers than do articulatory manifestations. Diehl and Kleunder [4] compile arguments for the primacy of auditory rather than gestural considerations in speech perception. Finally, from diverse perspectives, researchers including Martinet, Lieberman, Lindblom and Ohala have insisted on emphasizing simultaneously articulatory and acoustic properties in understanding the long-term (diachronic) properties and even the evolution of language capacity.

2. SEGMENTS AS SYMBOLS

Nearey [15] argues that speech production and perception represent a compromise in complexity between articulatory and acoustic patterns. I elaborate here on that framework from a neo-Sapirian perspective. There are (at least) three domains involved in speech, two physical and one symbolic. The symbolic part consists of the sequence of language-specific phonological elements. Without loss of generality (we can change our minds later), assume those symbolic elements are "phoneme-size" units called segments. The two physical domains are the articulatory (gestural) and the acoustic (auditory). Speech production is the mapping from segments to signals and speech perception is the opposite (not to say inverse) mapping.

In *strong motor* theories (e.g., LM) it is assumed that a natural invariant relationship exists between gestures and segments, while the mapping from acoustics to segments can be arbitrarily complex. In *strong auditory* theories, the roles of acoustics and articulation are reversed. In *double-strong* theories (e.g. Blumstein and Stevens taken to the extreme), the relationship of both physical domains to segments is assumed to be natural and invariant.

I have long been impressed by the sophistication of arguments of the strong frameworks and of the scholarliness and sincerity of their proponents. In fact, they have each convinced me that the others are wrong. To resolve this, I have adopted a "symbolic segment model" that is *double-weak* (weak motor, weak auditory). Segments are symbols and are neither articulatory nor acoustic in character. There is no fully natural, simple invariant relationship between either gesture or signal and symbol. Left as it stands, this amounts to a retreat into radical structuralist arbitrariness, wherein almost anything can happen in phonetics. Yet, if the gesture- and/or sound-to-symbol relationship is tightened up in the extreme, we arrive back at one of the three strong systems of the preceding paragraph.

Instead, I assume that the relationships between gesture segment and signal approach an "equilibrium of complexity" [15], a compromise between efficiency in production and rapid decoding in perception. Both auditory and articulatory properties will have a profound long-term influence on phonological systems. However, for perception only "the weakest form of a motor theory [8] (p. 204)" holds, which involves merely "what has been learned about relations between speech-production capabilities and the resulting acoustic output." Conversely, a weak form of an "auditory theory of speech production" also holds: articulatory targets and permissable coarticulation are constrained by what limited perceptual structures can readily decode. The kinds of constraint I have in mind might be formulated as follows: 1) A relatively simple, but not fully transparent, family of articulatory patterns is associated with each symbol. 2) A relatively simple, but not fully transparent, family of acoustic patterns is also associated with each symbol. 3) The relevant families of patterns exhibit moderate within-category variation relatable to contextual factors its own physical domain (articulatory or auditory).

In Fodorian terms, this is a proposal about the limits of computational complexity of separate sub-modules for speech perception and production. Long term pressures force each to respect the other's limitations, while allowing each to exploit the other's flexibility. However, the real-time operation of each is independent of the other. Each is an encapsulated, relatively "stupid" processor that only "looks" like it knows about the internal workings of the other (cf. the discussion of lexical priming in [5]).

Though limitations of arbitririness are imposed by this "symbiosis of encapsulated modules", there is room for much variety in how different languages approach their equilibrium of complexity. This allows for a language-specific Sapirian "warping" of the possible phonetic space. Explicit modeling of how the putative perceptual sub-module might implement such a warping can shed light on the question of its computational complexity.

2.1 Segmental Filter Models

The modeling framework proposed below is a generalization of the pattern recognition system proposed by Nearey and Hogan [16] to account for language-specific warping of the cue-space for simple experimental situations. In this "segmental filter" model, speech perception is assumed to involve an essentially bottom up, "reflex-like" mapping between properties of acoustic waveform and phonological segments. It is assumed further that the following limit exists on this mapping: all the knowledge that the perceptual system has about the consequences of patterns of production can be embodied in a set of Gaussian "filters", one for each phonological segment, tuned by acoustic/auditory properties. (These models are formally related to Massaro's FLMP, see [15].)

2.1.1 The NAPP Model

Fig. 2 illustrates the general properties of a simplified network of Sapirian segmental units. This example is based on the Thai data of [11, 15, 16]. In panel (a), the output of three VOT-tuned stop filters (for /d/, /t/ and /t^h/) are shown. These filters produce output, reflecting the "typicalness" of a given stimulus considered as a member of each category.

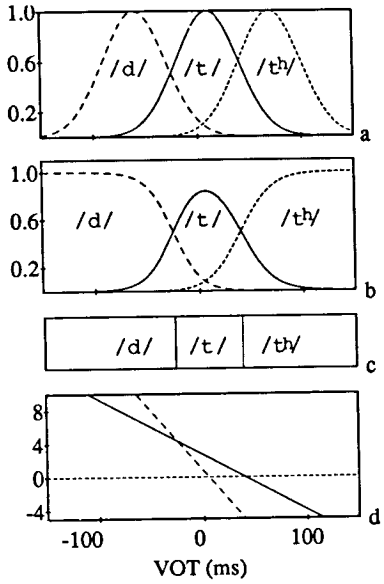

Figure 2. Segmental filters

In categorization, the probability of chosing a particular response, *r*, for a particular stimulus value, *s*, is a function of the relative distance from the "most typical value" for that category compared to the sum of the analogous distances from all three categories. Formally, this process can be separated into an *evaluation* function that determines the "fuzzy typicalness values" of Fig. 2(a) and a *choice* function that converts them into response probabilities. For a version of the "normal *a posteriori* probability" or NAPP model [15, 16], the evaluation function may be specified as follows:

$$f(r,s) = \frac{-.5\,[x(s)-m(r)]^2 + k(r)}{D(r)}, \quad (1)$$

where $x(s)$ is the stimulus value (here, VOT) for stimulus s, $m(r)$ is the mean VOT value for response category r (ranging over /d/, /t/ and /t^h/), $D(r)$ is the

standard deviation of VOT values in category *r* and *k(r)* is a normalizing parameter parameter that can accommodate a Bayesian *a priori probability* and can absorb a response bias term in a perceptual model. The models considered below involve homogeneous (homoschedastistic) Gaussian distributions, where the standard deviations (and correlations, in the multivariate case) are equal across all choice categories.. The choice function for the NAPP model is:

$$p(s,r) = \frac{\exp[f(r,s)]}{\sum_{r'} \exp[f(r',s)]}, \qquad (2)$$

where f(s,r) represents an evaluation function defined in (1) and the summation in the denominator is over all response categories.The result of the application of Equations (1) and (2) is shown in panel (b) of Fig. 2. which constitutes the *response surface* of the model and contains all the information about its stimulus-response mapping. In panel (c), the stimulus space is divided into three regions, each labeled with the dominant response in its range. The boundaries of such a *territorial map* or *decision space* are determined by the crossover points in the response surface of neighboring categories, which are in turn determined by the crossover points in the evaluation functions.

2.1.2 Logistic Models

For homogeneous Gaussians, the same response surface, (and decision space) can be formed by a set of three *linear logistic functions* as illustrated in panel (d) of Fig. 1. The equations for these lines are specified by:

$$f(r,s) = b(r) + a(r)\, x(s), \qquad (3)$$

where *b(r)* represents a bias term for the category *r*, which is *independent of the stimulus* value; while *a(r)* is a "stimulus-tuned effect." Such linear logistic models are choice-equivalent to Gaussian filter models. As discussed in detail in [15], logistics are readily estimable and can be generalized to characterize very complex decision spaces and response surfaces in a way that can be given interesting phonetic interpretations.

3.0 SEGMENTS VS. DIPHONES

A key property of segmental filter models is that they assume that stimulus properties are mediated in a fundamental way by phonological units of segment size. It has been demonstrated recently by Whalen [21] that response patterns from several experiments show that pure segmental models are not adequate to account for perceptual results. Nearey [15] shows that while Whalen's claim is true in a strict sense, only a minor modification of a "pure segmental" assumptions are motivated by available data.

Whalen sets out to test a claim by Mermelstein [14] concerning the independence of categorization of adjacent segments. Mermelstein's experiment involved simultaneous identification of both vowel and consonant in synthetic VC syllables when F1 and vowel duration were varied. The response categories ranged over the English words "bed, bet, bad, bat." Mermelstein's reults indicated that although vowel duration affected both vowel responses and consonant responses, *vowel and consonant judgments were made independently.*

In a series of analyses involving experiments with ambiguous VC and CV sequences, Whalen finds evidence counter to Mermelstein's claim, showing instead that the judgment of adjacent segments shows interdependencies consistent with a more complex decoding of production effects, in accord with a motor theoretic interpretation. Whalen's Experiment 3 involves categorization of fricative plus vowel sequences, spanning the choice set /si,su,ʃi, ʃu/. The kind of variation involved is typically described as coarticulatory, the most noticeable effect in production data being that frication noise for both fricatives has a lower low-frequency cutoff before /u/ than before /i/, presumably due to anticipatory coarticulation of lip rounding.

Based on previous experiments, Whalen notes that changes in vowel quality from /i/ to /u/ lead to fewer /ʃ/ and more /s/ responses for a given fricative noise. Conversely, changing a fricative context from /s/ to /ʃ/ causes more /i/ and fewer /u/ responses for vowels in an /i/-/u/ F2 continuum. In order to evaluate the contributions of physical versus phonological context, Whalen's Experiment 3 uses a two-parameter continuum spanning the four diphone choices. The parameters in question are: 1) F2 of a steady state vowel, ranging from 1386 to 1773 Hz in four steps; and 2) the frequency of a fricative pole (with a correlated zero located 1000 Hz below the pole) ranging from 2900 to 3100 Hz.

Roughly speaking, the fricative pole frequency, Pf, can be considered a "primary cue" for the /ʃ-s/ contrast, while F2 is the primary cue for /i-u/. However, Whalen's experiment shows that the /ʃi-si/ boundary along a Pf continuum differs from that of /ʃu-su/ in manner broadly in accord with production norms.

Three general varieties of effects of "vocoid" on "contoid"can be distinguished 1)The *physical value* of *F2* also directly affects or acts as a "secondary cue" for /ʃ-s/; 2). The */i-u/ judgment* affects the fricative response independently of the stimulus or 3) the acoustic properties directly affect the consonant and vowel choice in a manner that cannot be decomposed into effects like (1) and (2), so that a diphone is the smallest phonological unit that can be thought of as being directly tuned by acoustic properties. From a motor theory perspective, the last alternative represents a model that precompiles contextual variation into larger, more nearly invariant syllabic units.

3.1 Modeling

Nearey [15] presents a series of logistic analyses of Whalen's data which allow for the modeling of increasingly complex response surfaces using ANOVA-like factoring of terms. Specifically, it allows a decomposition of stimulus-response relationships in terms of 1) "stimulus-tuned" effects which cause changes in response probabilities as a function of changes in stimulus properties and 2) bias effects that are independent of stimulus properties. A further breakdown is possible in terms of the "size" of the phonological entity being considered, segments versus diphones. The factorization is represented by the terms in Table 1 (see [15]).

Table 1. Terms in logistic model.

Abbrev.	Term	Unit
Bias effects (stimulus-independent):		
V	bV(v)]	Vowel
C	bC(c)	Consonant
CV	bCV(c,v)	Diphone
F2-Tuned effects:		
VF	a1V(v) F2	Vowel
CF	a1C(c) F2	Consonant
CVF	a1CV(c,v) F2	Diphone
Fricative Pole-tuned effects:		
VP	a2V(v) Pf	Vowel
CP	a2C(c) Pf	Consonant
CVP	a2CV(c,v) Pf	Diphone

Various models can be constructed from these elements, producing decision spaces of varying complexity. For technical reasons, all models include the segmental bias terms V and C. A primary cue model would include the terms VF (vowel tuned by F2) and CP (consonant tuned by Pf). Its decision space is shown in Fig. 3. Note that the vowel boundary is independent of the fricative pole (parallel to the Pf axis) and the consonant boundary is similarly independent of F2.

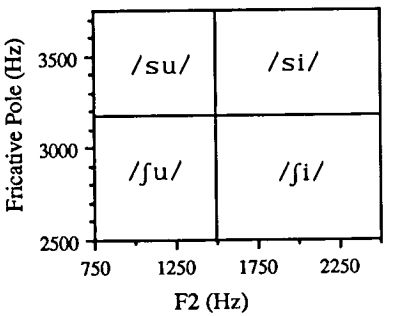

Figure 3. Primary cue model.

A secondary cue model could include the additional stimulus-tuned terms VP and CF and could lead to the decision space of Fig. 4. It remains a *pure segmental* model because *none* of the diphone terms of Table 1 are included; that is, although their *cues* overlap, the *symbolic* remain

segmental. Note that, while in some sense, this builds in context sensitivity that seems to have articulatory motivation, it does so "unintelligently" in that a generalized *acoustic* context effect can be directly incorporated as a (secondary) cue in the individual *consonant* filters, independent of phonological context, as in Mermelstein's original suggestion[14].

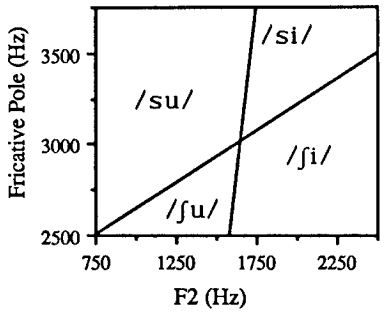

Figure 4. Pure segmental secondary cue model.

However, Nearey's analysis indicates, in accord with Whalen's claim, that *no* pure segmental model can adequately account for Whalen's empirical results.

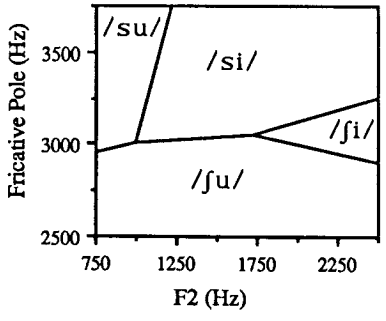

Figure 5. Hypothetical true diphone model.

On the other hand a true diphone model including *all* the terms in Table 1 can approximate more "intelligent" phonological context dependencies, achieving a decision space as complex as that of Fig. 5. But Nearey finds that such true diphone models are too powerful and, instead, an intermediate class of models, refered to as "transsegmentally biased segmental models" is completely adequate to account for Whalen's data.Such models can include all the terms of the secondary cue model plus the diphone bias terms (CV). However, the stimulus-tuned diphone terms (CVF, CVP) are *not* included. The decision space for the "best" model in Nearey's analysis is shown in Fig. 6.Formal properties of the model require that line segments separating syllables that share one segment must be parallel to each other, a restriction not shared by true diphone models. In fact, the model finally selected by Nearey

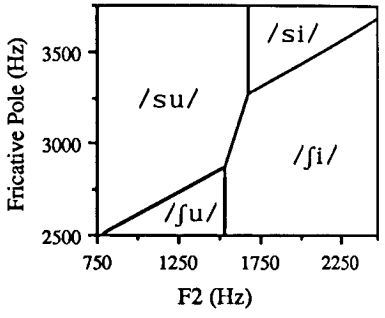

Figure 6. Restricted diphone-biased secondary cue model.

as best supported by the data is slightly simpler than the most complex possible biased segmental model, since the fricative-pole tuned vowel term (VP) is not included in this model. This is reflected by the fact that the /si-ʃi/ and /su-ʃu/ boundaries are parallel to the Pf axis in Fig. 6. Because of formal properties of the biased segmental models, these restrictions on parallelism of lines also would extend to *all other vowel conexts*, so that in a larger experiment, with more vowel responses, the same slope of the /s-ʃ/ boundary would be predicted within *all* vowel categories. In other words, the *relative efficacy* of the two cues (F2, Pf) in *changing* /s/ to /ʃ/ would be the same, independent of the following vowel.

3.2 Correction for coarticulation and allophonics

The above results have an interpretation in terms of a Fodorian "pseudo-smart" (one that is stupid, but looks smart) processor for coarticulation effects. Consider the following: Anticipatory lip-rounding makes /s/ more /ß/- like before /u/, while anticipatory spreading makes /ʃ/ more /s/ -like before /i/. That is, the vowel environments tend to produce "weaker cues" in those environments. But the diphone biases have the net effect of favoring the combinations with weakened contrasts, thus increasing their response areas. However, although useful, this is not a truly "intelligent correction", since it is *not cue sensitive*, but rather is a global bias on category pairs. This has implications for new experiments with more stimulus dimensions: namely that the response areas of the favored syllables would be in-creased along all stimulus axes vis-à-vis less favored ones, even those not affected directly by the main coarticulation effect in production. (So, /ʃi/ might "encroach" on /ʃɪ/ along the F1 axis, even though F1 was not involved in fricative vowel coarticulation).

In addition to its possible role as a "coarse correction" for coarticulation, Nearey notes [15] that many other experiments reported in the literature seem to be compatible with the restrictions of the biased segmental model and that there is as yet no clear experimental evidence to indicate that models as complex as true diphone models are ever required. Biased segment models can be viewed as a multilayer system. The first layer comprises a set of segmental filters wherein all stimulus tuning takes place, while higher-level units implement additive, stimulus-independent corrections for (passive) coarticulation, (preplanned) extrinsic allophony and phonotactic constraints. Such models also appear adequate to accommodate the kinds of "cognitive context effects" suggested by Ohala [19]. It also appears that the Ganong effect ([7]; see [15]) and the role lexical effects play in Lindblom's hypospeech [10] could be handled by lexical bias effects that do not interfere with the internal operation of the segmental filters.

4.LEXICAL ACCESS

Could informationally encapsulated segmental filters of the type described above really serve as the basis for lexical access? Marslen-Wilson ([13] has divided up the problem of "projecting sound into meaning" into two largely autonomous components: access and integration. Lexical access is viewed as *form-based* processing, whereby bottom-up phonetic information interacts with the lexicon to select a unique lexical item. This lexical candidate is then presented to a higher level "content-based" process of *integration*, wherein the newest lexical item is incorporated into the syntactic and semantic processing of the sentence. From the point of view of the existence of a submodular language processing system, the key conclusions are: "First, that sentential context does not function to override perceptual hypotheses based on the sensory input system. (p 19)." Second, that top-down effects (e.g., sentential context) "do not affect the basic perceptual processing of the sensory input." Some of the evidence for these conclusions is considered below.

Important work by Samuel indicates that there are very strong constraints on how syntactico-semantic information influences lower level processing. Samuel's work involves the use of a classical signal detection paradigm to investigate the decomposition of effects into what he refers to as *perceptual* and *post-perceptual* components. Subjects try to detect the difference between two kinds of distorted natural speech: one in which a phoneme has been replaced by noise and one in which noise has been added to the original phoneme. In a series of carefully designed experiments, Samuel varies a number of characteristics, including the phonetic nature of the segments distorted, lexical status (word versus pseudo-word) and sentential context.

Sentential context is shown to only affect listener's bias toward saying "added" (i.e. the phoneme is restored) in appropriate contexts, but the discriminability measure *d'* is not affected. That is, differences between "added noise" and

"replaced by noise" stimuli were equally salient to listeners, regardless of sentential context. They were simply globally more likely (biased) to say "restored" to stimuli in appropriate semantic contexts. However, in contrast to syntactico-semantic effects, Samuel's workindicates that *lexical status* (being a real word) may affect lower-level (phono-logical) processing, since *discriminability* for real words was less than for non-words. However, the work of Samuel and Ressler [20] confirms the finding (by Nusbaum and colleagues) that the lowered discriminability for words is strongly affected by attentional factors and may result mainly from subjects' inability to focus on segments within words. While more research is clearly needed, this result, coupled with the tractability of Ganong-effect in synthetic experiments, leaves open the possibility that an encapsulated set of segmental filters operating prior to lexical access.

5 EXTENSIONS AND PROBLEMS

While the biased segmental models seem to be compatible (so far) with a variety of results from the literature, there are at least a few cases that their simple linear boundaries cannot handle. The facts surrounding the famous case of place of articulation of stops appear to require something somewhat more complex. To the best of my knowledge, the Cooper et al. experiment represented by Fig. 1 manifests the most complex decision space ever found in phonetic research. The main pattern as characterized by the authors is roughly as follows: /t/ dominates when burst frequency is high; /k/ when its frequency is slightly above F2; and /p/ otherwise. While beyond the reach of homogeneous dispersion Gaussians, this general pattern can be achieved a Gaussian model which allows separate covariance matrices for each group (corresponding to a quadratic logistic model.; cf. Nearey and Shammass [17] for application of the related quadratic discriminant analysis to *transitions* in stop consonants). Finally, the minor mode of the /k/ region that occurs when the burst is just above the F1 in front vowels may require an additional wrinkle. Nonetheless, the general pattern of this decision space can be generated by segmental filters as described below: A single bivariate Gaussian in F2 and burst frequency is used to characterize each of the /p/ and /t/ distributions. However, /k/ requires a mixture of two bivariate Gaussians: one to characterize the F2 burst relationship and the other for F1 and burst. In fact, Fig. 1 was generated analytically in just this way. Although this pattern is more complex than those of the simple logistics of the previous example, it still represents a relatively elementary problem in pattern recognition. Note that this does not deny that aspects of the pattern are motivated in the long run by articulatory factors, only that real-time perceptual behavior does not need to compute articulatory or gestural properties to decode them. Since this may be as complex it ever gets, there seems good reason to continue to explore segmental filter approaches to speech perception.

There is, however, one very large problem that must be faced squarely in any such exploration: while the above model makes inroads on the traditional problem of invariance, it has ignored the problem of segmentation. First note that although the models are segmental at the symbolic level, they are manifestly not so at the acoustic level, since pervasive temporal overlap is allowed in the cue domains of neighboring segments. Though these segments are not necessarily phonemes ("major allophones" would do), I propose that the constraints of the acoustic-to-segment mapping be modified forms of Chomsky's conditions on the relation "systematic phones" to taxonomic phonemes[2]. (i) weak linearity: the centers of the window of relevance of the acoustic cues preserves the left-right order of the strings of segments. (ii) local determinacy: such windows are not arbitrarily wide; (iii) strict bottom-up mapping (replacing biuniqueness); (iv) higher-order invariance.

With respect to (iv), given temporal alignment of the window of relevance (and cue-extraction!), the claim is that the patterns are relatively invariant, usually mapping to simple linear decision spaces. This, however, is a very large "given." The plausibility of the scheme presented above, no matter how successful it may be for "toy" problems in the phonetics lab, is ultimately dependent on the ability to supply cognitively plausible models of signal alignment. In this regard, I think we have much to learn from the computational methods of time alignment being developed in the speech recognition community.

ACKNOWLEDGEMENTS

Work supported by SSHRC. Thanks to T. Welz for technical support.

REFERENCES

[1] BLUMSTEIN, S. & STEVENS, K. Acoustic invariance in speech production : Evidence from measurements of the spectral characteristics of stop consonants. JASA **66**: 1001-1017, 1979.

[2] CHOMSKY, N. "Current issues in linguistic theory." 1964 Mouton. The Hague.

[3] COOPER, F.S., DELATTRE, P., BORST, J. & GERSTMAN, L. Some experiments on the perception of synthetic speech sounds. JASA **24**: 597-606, 1952.

[4] DIEHL, R. & KLUENDER, K. On the objects of speech perception. Ecological Psychology. **1**: 121-144, 1989.

[5] FODOR, J. "The modularity of mind." 1983 MIT Press. Cambridge, MA.

[6] FOWLER, C. An event approach to the study of speech perception from a direct-realist perspective. J. Phonetics. **14**: 3-28, 1986.

[7] GANONG, W. Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance. **6**: 110-125, 1980.

[8] KLATT, D. "Review of selected models of speech perception." Lexical Representation and Process. Marslen-Wilson ed. 1989 MIT Press. Cambridge MA.

[9] LIBERMAN, A. & MATTIGLY, I. The motor theory of speech perception revised. Cognition. **21**: 1-36, 1985.

[10] LINDBLOM, B. "Explaining phonetic variation: a sketch of the H&H theory." Speech production and speech modeling. Hardcastle and Marchal ed. 1990 Kluwer Academic Publishers.

[11] LISKER, L. & ABRAMSON, A. "The voicing dimension: some experiments in comparative phonetics." Proceedings of the 6th International Congress of Phonetic Sciences, Prague. Hala, Romportl & Janota ed. 1970 Academia. Prague.

[12] MACNEILAGE, P. Motor control of the serial ordering of speech. Psychological Review. **77** (182-196): 182-196, 1970.

[13] MARSLEN-WILSON, W. "Access and integration: Projecting sound onto meaning." Lexical Representation and Process. Marslen-Wilson ed. 1989 MIT Press. Cambridge MA.

[14] MERMELSTEIN, P. On the relationship between vowel and consonant identification when cued by the same acoustic information. Percep. and Psychophys. **23**: 331-335, 1978.

[15] NEAREY, T. The segment as a unit of speech perception. J. Phonetics. **18**: 347-373, 1990.

[16] NEAREY, T. & HOGAN, J. "Phonological contrast in experimental phonetics: relating distributions of measurements production data to perceptual categorization curves." Experimental Phonology. Ohala & Jaeger ed. 1986 Academic Press. New York.

[17] NEAREY T. &. SHAMMASS, S. Formant transitions as partly distinctive invariant properties in the identification of voiced stops. Canadian Acoust. **15**: 17-24, 1987.

[18] NEAREY, T. M. On the physical interpretation of vowel quality: cinefluorographic and acoustic evidence. J. Phonetics. **8**:213-241, 1980.

[19] OHALA, J. "What's cognitive, what's not." Morrissey ed. in press Peter Lang Verlag. In press.

[20] SAMUEL, A. & RESSLER, W. Attention within auditory word perception: insights from the phonemic restoration illusion. Journal of Experimental Psychology: Human Perception and Performance. **12**(1): 70-79, 1986.

[21]WHALEN, D. Vowel and consonant judgments are not independent when cued by the same information. Percep. and Psychophys. **46**(3): 284-292, 1989.