PRELIMINARY THOUGHTS ON "PHONOLOGIZATION" WITHIN AN EXEMPLAR-BASED SPEECH PROCESSING SYSTEM

ROBERT KIRCHNER
University of Alberta

This paper discusses a particular problem attendant upon the hypothesis that phonological patterns emerge directly from considerations of phonetic functionality: the problem is that phonological patterns frequently display stability across tokens and contexts, where direct phonetic conditioning would predict variation. I suggest that this stabilization might emerge from an exemplar-based speech processing system, which, in essence, enforces the notion of analogical pattern extension.

This paper discusses a particular problem attendant upon the hypothesis that phonological patterns emerge directly from considerations of phonetic functionality, and tentatively explores a possible solution. The problem is that many sound patterns display a certain stability, across tokens and contexts, which is difficult to account for in terms of purely phonetic considerations; moreover, these stabilized patterns appear to refer to representations which are considerably more abstract and categorical than raw phonetic representations (see Anderson 1981; Hayes 1997). Typically, we observe a diachronic progression from an unstable, gradient pattern of variation to a stable, categorical alternation (henceforth *phonologization*, Hyman 1975). The standard theory merely stipulates this observation, in terms of distinct "phonological" and "phonetic" levels of representation, thereby giving up on, inter alia, an elegant account of the overwhelming commonalities between the two types of patterns. The standard treatment further forces a categorical choice between phonological and phonetic treatments, precluding a natural treatment of *partially* phonologized patterns of variation, e.g. English vowel reduction, which is relatively stable in certain classes of words, and certain phonetic contexts, though it is inseparable from a broader pattern of variable reduction (see Fidelholtz 1975).

Instead, I suggest that the mismatches between phonologized patterns and "mere phonetics" are precisely the properties which would emerge from a simple exemplar-based speech processing system: that is, a system which recognizes inputs, and generates outputs, by analogical

evaluation across a lexicon of distinct memory traces of each token of speech (see generally Johnson & Mullennix 1997). Ultimately, I propose to develop a computational simulation to test this claim, at which point this paper will be expanded to incorporate the results of this simulation.

An exemplar-based model has previously been applied to speech perception (Johnson 1997), as a way around the "lack of invariance" problem. That is, for any given lexical item, a fortiori for any individual phoneme, it has hitherto proved impossible to identify a set of phonetic cues which is invariantly present in the phonetic realization, across a range of speakers and phonetic contexts, which suffices to distinguish that item/phoneme from those with which it is in contrast.[1] Thus, some realizations of /tin/ can't be distinguished from realizations of /din/, of /tim/, of /tɪm/, etc. Rather, Johnson proposes, let the phonological representation of lexical items consist of a cohort of exemplars of previously perceived realizations of that item, with all phonetic details present. The input can be categorized as a new exemplar of some category (which may be understood as a lexical item, a semantic representation, or some lower-level linguistic type such as a phoneme) if it is sufficiently phonetically similar to the whole cohort of this category; and, in particular, if it is more similar to this cohort than to the cohort of any other category. Unlike the standard theory, however, no single property need be present in every member of the cohort. That is, instead of attempting to find invariant phonetic properties which distinguish among phonemes or lexical items, the task is to compute the similarity of the input to the various cohorts of exemplars within the lexicon. The results of Johnson's model (which addresses the limited problem of categorizing steady-state vowels across a range of male and female speakers), are encouraging: based on representations solely containing an F0 value, formant frequency values, and duration, vowels were identified with 80% accuracy; and the sex of the speaker was identified with 98% accuracy. Further motivation for this approach is found in a series of word identification experiments (Goldinger 1997). For low-frequency words, subjects were

[1] A frequently cited metaphor, which poignantly captures the difficulty of the problem, is Hockett's conveyor-belt of colored eggs (i.e. phonemes) which are smashed and smeared by rollers into a mess of yolk, albumen, and shell fragments (the phonetic realizations). The speech perception side of this interface must then reassemble the original sequence of eggs from this mess. Once the phonemes have been reconstituted, it is trivial to recognize the input as some (string of) lexical item(s), from which the meaning of the utterance can ultimately be computed. However, all the king's scientists have as yet been unable to put the eggs together again, despite concerted efforts over the past four decades.

found to perform significantly better when the stimulus and priming tokens involved the same voice, even when the stimulus was presented a full week after the priming. This relatively long-term effect on subjects' performance of same speaker in priming and stimulus tokens suggests that speaker-specific properties of tokens of words remain part of the representation of these words in long-term memory. (In the case of higher frequency words, however, the individual exemplar does not stand out from its neighbors to the same extent, because activation of the one also activates a much larger cohort of similar exemplars.[2]) (Exemplar-based models have also been applied to partially productive morphophonemic variation in Finnish, in Skousen 1989, 1992, and to English, in Skousen & Derwing 1994; however, this class of models uses an admittedly ad hoc metric of similarity, and, due to its "nearest neighbor" approach, fails to detect subtle but potentially significant similarities among diffuse cohorts.)

Notwithstanding this previous research, the deployment of this sort of model in phonological analysis represents a radical departure from many traditional assumptions of linguistic theory. In embarking upon this relatively untrodden path, I understand that I incur a debt, to explain to the linguistic community why this approach seems more promising, and how it might capture the basic empirical observations which we have previously captured in abstractionist, symbolic terms. Consider this paper, then, a first installment on this debt.

1. Phonetic Functionalism and its Discontents

1.1. Background

One of the oldest ideas in linguistic theory is the hypothesis that natural language sound patterns (whether viewed as diachronic developments, e.g. Osthoff & Brugmann 1878, or properties of a synchronic grammar, Grammont 1939, Stampe 1972) can be explained in terms of considerations of phonetic functionality. For example, it has been frequently observed that lenition processes, such as spirantization, typically occur in intervocalic position. This pattern would appear to have to do with articulatory effort minimization: more effort is presumably required, on average, to achieve closure when

[2] Similarly, the limitation of the effect to relatively recent tokens can be modeled by assuming a high base activation of recent exemplars, which decays over time to some minimum level, at which point the individual exemplar no longer stands out as well from the cohort as a whole. That is, the individual gets "lost in the crowd."

flanked by open segments than in other contexts; to avoid expending this extra effort, speakers tend to undershoot the closure target in this more effortful context, hence spirantization (Kirchner 1998, ch. 6). The elegance of this functionalist idea is compelling (assuming it can be properly fleshed out): the phonological pattern reduces to relatively well understood and well motivated principles of biomechanics, without any further, specifically linguistic, stipulations.

Historically, the problem with such a line of thought has been the difficulty of developing it into fully explicit analyses of the sound patterns of particular languages. Particularly within rule-based frameworks such as that of Chomsky & Halle 1968, it appears that particular sound systems must be formally characterized in terms of a set of rewrite rules. Given the necessity of these rewrite rules, the (ostensibly explanatory) functional phonetic principle is not actually doing any descriptive work. The explanatory principle may be, as Prince & Smolensky (1993) wryly put it, "inert but admired"; or it may be dismissed as naively Panglossian; but in either case, the principle is superfluous to the scientific task, namely the explicit analysis of phonological data.

The advent of Optimality Theory (Prince & Smolensky 1993) has changed this. Within this more recent framework, phonological patterns can be explicitly characterized in terms of interactions of soft constraints of extreme generality. Moreover, cross-linguistic variation can be made to follow from alternative rankings of a universal[3] set of constraints. The further step of identifying this universal constraint set, at least in part, with principles of phonetic functionality, has been taken in such recent work as Steriade 1993, 1995, 1996; Kaun 1994; Flemming 1995, 1997; Jun 1995; Silverman 1995; MacEachern 1996; Myers 1996; Beckman 1997; Hayes 1997; Kirchner 1997; Boersma 1998; Kirchner 1998; and Gordon (in progress).

1.2. Exemplification of the Problems

Some of the remaining obstacles to this theoretical goal are exemplified in an analysis of Tigrinya post-vocalic spirantization, from Kirchner 1998, ch. 9 (data from Kenstowicz 1982; see generally Kirchner 1998 for a more in-depth exposition of this effort-based approach to lenition):

[3] This is, at least, a goal to which the theory aspires, although particular OT analyses have sometimes fallen short of it, in the face of particularly unusual, often morphologized, phonological patterns. I discuss a solution to this problem in section 3.

(1)	a.	kətəma-xa	'town-2sg.m.'
		ʕarat-ka	'bed-2sg.m.'
		k'ətəl-ki	'kill-2sg.f. perfect'
		mɨrax-na	'calf-3sg.f.'
	b.	kəlbi	'dog'
		ʔa-xalɨb	'dogs'
		ʔɨti xalbi	'the dog'
	c.	k'ətəl-a	'kill-3pl.f. perfect'
		tɨ-χətl-i	'kill-2sg.f. imperfect'
	d.	fəkkəra	'boast'
		k'ətəl-na-kka	'we have killed you (masc.)'

As shown in (1) velar stops, except for tautomorphemic geminates (d), spirantize in post-vocalic position. (The geminate blocking, which is extraneous to our present concerns, is given an effort-based account in Kirchner 1998, ch. 5.) The post-vocalic environment is analyzed as the union of coda and intervocalic environments. Coda spirantization, in turn, is attributed to the impoverished perceptual cues, due to the unreleased realization of stops, in this position (the reasoning is that there is greater impetus to lenite in contexts where there is relatively little perceptual "bang" for the articulatory "buck".) This idea is formally expressed in terms of a context-sensitive faithfulness constraint, PRES(cont /released), which is inherently ranked above the more general PRES(cont) constraint:

(2)

	PRES(cont /released)	LAZY	PRES(cont)
mɨrak-na - mɨrakna		**!	
☞ mɨrak-na - mɨraxna		*	*
☞ ʕarat-ka - ʕaratka		**	
ʕarat-ka - ʕaratxa	*!	*	*

Intervocalic spirantization, as noted above, is attributed to the greater effort required to achieve a stop in this context. This idea is formally expressed by decomposing the scalar LAZY constraint (i.e. effort minimization) into a series of binary constraints, each penalizing some numerical effort threshold; a lenition-blocking constraint (in this case,

PRES(cont /released)) can then be ranked just below the effort threshold corresponding to the cost of achieving a stop in intervocalic position, here labeled *x*:

(3)

	LAZY$_x$	PRES(cont/ released)	LAZY *x*-1	PRES (cont)
mɨrak-na ~ mɨrakna			*!	
☞ mɨrak-na ~ mɨraxna				*
☞ ʕarat-ka ~ ʕaratka			*	
ʕarat-ka ~ ʕaratxa		*!		
kətəma-ka ~ kətəmaka	*!		*	
☞ kətəma-ka ~ kətəmaxa		*		

The problem with this analysis is that there is no particular effort threshold which can be stably equated with a context such as intervocalic position. A variety of pragmatic and phonetic conditions would presumably result in considerable variation in the effort cost of a given gesture (assuming that effort cost fairly directly reflects biomechanical notions such as energy). For example, the faster the speech rate, the greater the effort cost of achieving a given constriction: greater acceleration is required to achieve the target in a shorter amount of time. This is a two-edged sword. By tying lenition contexts such as intervocalic position to effort thresholds, we can achieve an elegant account of rate-sensitive lenition (an abundantly attested phonetic process, e.g. Lindblom 1983, Kohler 1991). But this particular Tigrinya pattern is not described as rate-sensitive: post-vocalic dorsal obstruents apparently spirantize even in slow, careful speech. If the effort threshold LAZY$_x$ is truly what conditions spirantization in this case, the spirantization pattern is incorrectly predicted to be variable, with higher probabilities when flanked by low vowels or in fast speech, and lower probabilities when adjacent to a lower-sonority segment. While it is possible, even probable, that many reports of ostensibly stable context-sensitive lenition merely reflect idealization of the sound patterns,[4] Michael Kenstowicz (p.c.) is clear that this is not the case in

[4] Such idealization, unfortunately, is encouraged by the tradition of collecting and publishing phonological data solely in terms of impressionistic phonetic transcriptions,

Tigrinya. We are apparently forced to concede that the conditioning of lenition, at least in Tigrinya, is only *quasi-phonetic*. This leads us to the second major problem. The conditioning of this spirantization alternation appears to refer to a relatively abstract, categorical property, i.e. presence vs. absence of a preceding vowel, rather than continuous phonetic values, such as the details of gestural timing and magnitude, which determine effort cost. Moreover, conditioning by a relatively small set of abstract, categorical distinctions appears to be a significant tendency (though not an absolute) among patterns involving stable, categorical structural changes generally.

The answer to this problem, I believe, is that stabilized, categorical, quasi-phonetic patterns can follow from interaction between truly phonetic constraints (such as LAZY) and an extremely general constraint, enforcing extension of lexical patterns; the latter constraint, in turn, is grounded in considerations of lexical learning and retrieval during speech processing, and emerges from an exemplar-based model of speech processing.

2. BASICS OF THE MODEL

It is perhaps easiest to discuss the *modus operandi* of this constraint in terms of a sketch of a neural net implementation of an exemplar-based speech processing system.

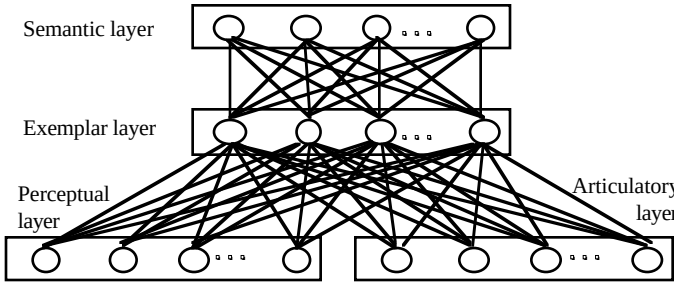

Figure 1. Neural net architecture implementing exemplar-based model

without recording and instrumentally measuring the speech properties in question. While phonetic transcription is certainly an indispensable tool for many aspects of linguistic analysis, it can give a misleading impression of categoricalness, due to the inherently categorical nature of discrete phonetic symbols.

A neural net is employed specifically for its capacity for readily computing multiple dimensions of similarity across a large number of items. As in most neural net models, the *activation level* of a node i is determined by the following equation:

$$(4) \quad a_i = f\left(\sum (a_j \cdot w_{ij})\right)$$

for all nodes j to which i is connected, where a_i and a_j are the activation levels of the nodes i and j respectively, and w_{ij} is the weight of the connection between i and j; the function f is some non-linear (typically logistic) transformation of the raw summation, making the activation level bimodal, i.e. tending towards either negligible or full activation. Substantive representational issues are discussed in section 2.3. Finally, it should be understood that a complete speech processing model would include phonetic and cognitive constraints, such as effort minimization, homophony avoidance, rhythmic well-formedness, and pragmatic felicity, interacting with the basic lexical storage and retrieval system shown in Figure 1.

2.1. *The Task of Word Recognition*

Word recognition involves the computation of an output semantic representation in response to an input perceptual representation (we can ignore the articulatory nodes for the moment). The input is represented in this model in terms of activation of particular perceptual nodes (for example, one of the nodes might be devoted to the detection of fricatives, becoming activated in response to aperiodic energy in the sound signal); and the output is represented as some pattern of activation over the semantic nodes. The computation takes the form of *spreading of activation* from the input nodes to the exemplars, and thence to the semantic nodes. It is the connection weights between pairs of nodes which determine precisely how the activation will spread. Thus, as in neural net models generally, learning takes the form of operations on connection weights. However, unlike conventional connectionist models, there are no "hidden units" (nodes which the system simply uses for purposes of internal computation, which do not admit of straightforward interpretation);[5] nor does the system need an external "teacher" to correct its errors; nor does the learning algorithm employ back-propagation of weight adjustments. The (simple Hebbian)

[5] Which, consequently, make it difficult to understand precisely how the system is arriving at the right (or wrong) answer.

learning algorithm can be summed up simply as "add another exemplar as you process the input." More precisely:

- Prior to any learning, all connection weights = 0.
- Let *i* be some as yet unused exemplar node (i.e. which currently has 0-weight connections to all other nodes). Set weights equal to 1 for all connections between *i* and the activated perceptual nodes (excitatory connections), and set weights equal to -1 for all other (unactivated) perceptual nodes (inhibitory connections). These connections constitute the perceptual representation associated with this exemplar.
- Activation spreads not only to *i*, but also to all exemplar nodes with (previously established) positive weights to the perceptual nodes which are currently activated. That is, previous exemplars are activated to the extent that they are perceptually similar to (i.e. share activated perceptual nodes with) the current input (the *analog probe*). Activation then spreads from the analog exemplars to all semantic nodes with (previously established) positive weights to the analog exemplars.
- In addition, semantic nodes may become activated (or inhibited) in response to the pragmatic context (for example, the learner is currently looking at a ball). This is crucial for learning semantic representations for new words, and for permitting recovery of words in the face of signal degradation, etc.
- Once the semantic pattern of activation has been computed, set weights equal to 1 for all connections between *i* and the activated semantic nodes, and set weights equal to -1 for all other (unactivated) semantic nodes. That is, establish a semantic representation for the new exemplar. Node *i* will now function as part of the analog probe for any similar inputs in the future.

2.2. The Task of Speech Production

Speech production (ultimately) involves the computation of an output plan of neuromuscular commands to the vocal tract from an input semantic representation, again mediated by the exemplar layer. But whereas every exemplar in an analog probe already has a perceptual representation, this is not necessarily the case on the articulatory side. The particular word which the speaker plans to produce may be one

which she has heard (or read[6]) but never produced before, in which case the cohort of exemplars of that word will have non-zero weights to the semantics and perception, but zero-weights to all articulatory nodes. Moreover, even if the speaker has previously produced a few tokens of this particular word, these particular tokens may be an inappropriate plan for pronunciation in the current context. For example, past tokens of the word *hemophiliac* may have involved quiet or normal speech productions, whereas the current context (e.g. a medical emergency) may require the word to be shouted.[7] What the simple semantics-to-articulation architecture is lacking, then, is more generalized, systematic knowledge of the mapping between perceptual targets and their articulatory realizations.

This lack can be filled, however, by including a *perceptual echo probe* in the production computation: that is, a secondary wave of activation, spreading from the perceptual layer to the articulatory layer.[8]

[6] In a fully developed system of this sort, we would ultimately need to include the visual system (which detects orthographic representations), with exemplar-mediated spreading of activation to the phonetic and semantic layers, embodying implicit knowledge of spelling-to-sound regularities.

[7] Shouted speech involves not only more forceful pulmonic egression, but non-trivial adjustments of the magnitude and duration of all the articulatory commands. For example, stop closure must be made with more force than under normal conditions, otherwise the greater air pressure might blow the articulators apart; approximants and vowels must be made with greater opening, otherwise the high-volume airflow could become turbulent.

[8] Infant babbling can thus be modeled as the creation of semantically null exemplars, recording experiences of, inter alia, articulation-to-perception mappings.

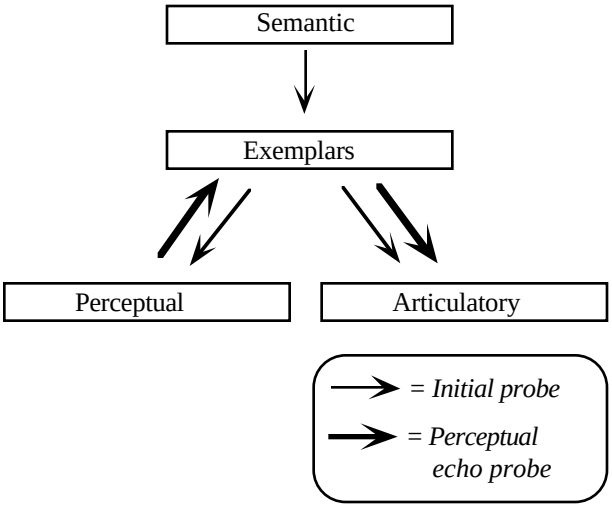

Figure 2: Path of the perceptual echo probe in speech production

That is, the input activation of the semantic nodes spreads (via the exemplars) not only to the articulatory layer (if there are any non-zero connections), but also to the perceptual layer, in effect creating a perceptual target for the production. The perceptual nodes then send back an echo to the exemplars, resulting in activation not only of the originally semantically activated exemplars (i.e. tokens of the target word), but also of all exemplars, in proportion to their perceptual similarity to the perceptual target. Finally, this echo of activation spreads from the exemplars to the articulatory layer, modifying the original articulatory plan (if any), in light of this richer base of experience of mapping from perceptual properties to articulatory commands.[9] It will be argued in section 3 that *phonologization is a consequence of this perceptual echo probe.* Finally, as in the recognition task, an empty (0-weight) exemplar node is assigned non-zero connections to all the peripheral nodes, representing the token of speech production which has just been computed, which can then be used in future production or recognition tasks.

2.3. Representational Issues

[9] I leave open the question of whether speech perception involves a similar articulatory echo probe. That is, hearers might make use of their own exemplars of production to supplement the actual sound signal detected by the perceptual nodes. This further assumption would bring the model closer, in certain respects, to the Motor Theory of speech perception.

The *modus operandi* of this model, then, is the emergence of linguistic types, in the course of perception and production, from simultaneous activation of cohorts of similar experiences.[10] It thus follows that the representations should be should be fairly "raw," without reification of linguistic types such as phoneme units; moreover, the representations must be broken down into the elements upon which similarity is to be computed (i.e. featural decomposition).[11]

- However, since evaluation of *semantic* similarity (beyond shared meaning within morphological paradigms) does not appear to be central to the basic phonologization problem I'm interested in testing, it seems safe to duck the (non-trivial) problem of positing a universal semantic feature set; rather, I assign a distinct node for each morpheme.
- The *articulatory* representation, in principle, consists of activation levels, over time, of particular muscle groups of the vocal tract; however, since no activation flows upward from the articulatory layer in the current model, the system never has to compute articulatorily-based similarity. I assume that the technique used for converting the continuous perceptual representation into discrete nodes in this model (see below) will extend to the articulatory representation as well.
- This leaves the perceptual representation, which is most crucial to our present concerns. In keeping with the "rawness" requirement, I assume this to be patterns of excitation of the basilar membrane, containing the same sort of information as a conventional acoustic spectrogram, but rescaled in perceptual units, Bark and phon, rather than Herz and decibels, and reflecting forward masking (i.e. relatively loud sounds partially obscure relatively quiet sounds which immediately follow them, due to a lag in neural recovery from excitation) (see generally Boersma 1998, ch. 4).

---

[10] A helpful metaphor for this notion of an emergent type is the superposition of photographic images of a human face (Goldinger 1997, attributing the metaphor to Semon (1909)). As more and more images are added, the details of particular faces blur into the background; what stands out are the generic features of the human face: a fuzzy outline of the head, with blurry regions corresponding to eyes, mouth, etc. In the composite photograph, however, the greater the inconsistencies, the muddier the result. But in a cohort of exemplars, the inconsistencies would cancel each other out, due to inhibitory connections; the property would then be activated only to the extent of the difference between the exemplars with that property and those without it; this results in a clearer emergent prototype than the photograph metaphor would suggest.

[11] It is sometimes asserted by phonologists that computation of a general similarity metric over pairs of whole representations is an intractable problem, in which case my approach is unfeasible. If this is correct, Optimality Theory is unfeasible as well, since the set of faithfulness constraints collectively perform an equivalent similarity evaluation.

To get from this auditory spectrographic representation to a layer of discrete nodes within a neural net, some sort of quantization is required. First, let us divide the spectrogram into timeslices of, say, 10 msec.

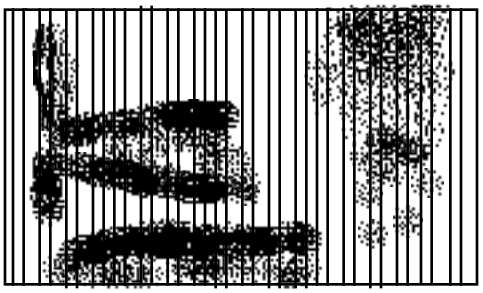

Figure 3: Auditory spectrogram divided into timeslices

Each timeslice can then be treated as an average of the spectral frequency and loudness values (periodic or aperiodic) during that timeslice (

b.

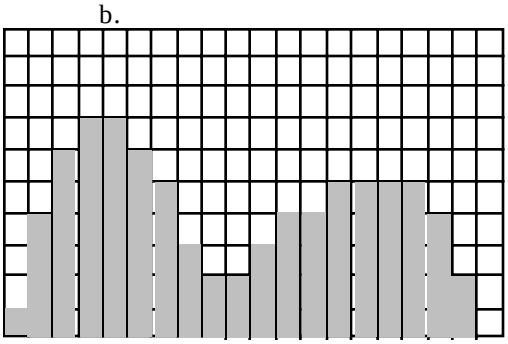

Figure 4a), which in turn can be quantized into a grid of feature and loudness ranges (b).

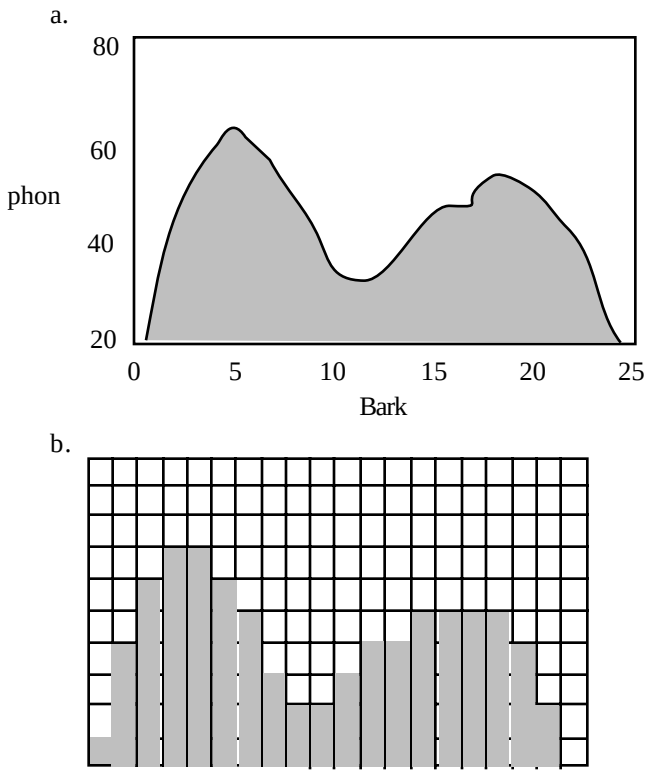

Figure 4. Quantization of timeslice spectrum into a grid of frequency and loudness ranges

This particular representational strategy is highly tentative; but I include this to give a concrete idea of how the problem might be approached. Now, we can think of each timeslice as a feature matrix, where the features are the frequency regions (e.g. 0-.9 Bark, 1-1.9 Bark, etc.) and the values of those features are the loudness ranges of periodic or aperiodic energy (e.g. [1-1.9 Bark(periodic loudness ≥ 25 phon)]). Moreover, we can have a static encoding of the temporal dimension, by having the actual nodes represent not the feature specifications themselves, but the temporal relations among these feature specifications, specifically *simultaneity* (for the specifications occurring within the same feature matrix), and *precedence* (i.e. the temporal order of a specification of a feature relative to other specifications of that

feature in other matrices). Thus, the perceptual layer would include nodes such as:

- Precedes([1-1.9 Bark(periodic loudness ≥ 25 phon)], [1-1.9 Bark(aperiodic loudness ≥ 15 phon)]), and
- Simultaneous([1-1.9 Bark(periodic loudness ≥ 25 phon)], [2-2.9 Bark(periodic loudness ≥ 20 phon)]).

Finally, the number of particular pairs of specifications for which these temporal relations hold true can be encoded in terms of the *activation level* of the nodes: the more instances, the higher the activation.[12]

Some further results of this representational scheme are that similarity with respect to inherently *salient* (i.e. louder and longer) properties will result in a higher similarity evaluation. In the case of duration, this is because the two representations will match with respect to a greater number of timeslices. In the case of loudness, this is because loud sounds match with respect to a greater number of loudness thresholds. Moreover, a representation of unlimited length can be encoded. And whereas comparison of phonetic similarity in the psychological literature often simplifies the problem, by considering only left-to-right one-to-one correspondence of segments (with disastrous results in (5b),

(5) a. p_1 $æ_2$ s_3
p_1 $æ_2$ s_3 t_4
(identical but for t_4)

b. p_1 $æ_2$ m_3
s_1 p_2 $æ_3$ m_4
(no similarity detected)

my model does not rely on particular alignments of the temporal units before similarity can be evaluated: the temporal relations of, for example, the features of [æ] to the features of [p] will be the same for both [pæm] and [spæm].

2.4. The "Head-Filling-Up" Problem

Each individual computation, in this model, is extremely simple; and because it is massively parallel, the computation should be extremely fast. The price paid for this speed is the need for massive storage.

[12] This is intended to address the [algal] vs. [algalgal] distinction of Pinker & Prince 1988.

Classic Generative Phonology assumed that this kind of storage was unfeasible (perhaps influenced by the severely limited memory available to computers during the decades of its heyday) and opted for maximally economical storage. But there is no serious psycholinguistic or neurological support for the Generative assumption that "mental space is at a premium." Nevertheless, with storage of every exemplar, one may reasonably inquire whether the "head filling up" becomes a concern. Recall, though, that for each additional exemplar added to the lexicon, only a single node is required, no matter how detailed the representations, because the phonetic and semantic content of the exemplar is represented in terms of connections to a finite number of nodes in the perceptual, articulatory, and semantic layers. Assuming that a human experiences three utterance/exemplars per minute, eighteen hours a day, over a lifetime of one hundred years, if we take seriously the equation of nodes with neurons, the number of neurons used up as speech exemplars is a mere 1,183,410, a small fraction of the roughly 10 billion neurons in the human brain. Finally, even if the "head filling up" were to remain a serious concern, there are techniques of modeling the crucial properties of an exemplar-based lexicon, without literally storing every exemplar. For example, Johnson (1997) describes a "covering map" strategy, which, in brief, does not store new exemplars of tokens which are essentially perceptually identical to some previous exemplar, but simply updates the token-frequency value of this region in perceptual space.

3. PHONOLOGIZATION AND THE PERCEPTUAL ECHO PROBE

Let us now return to the problem of Tigrinya post-vocalic spirantization. Assume an initial diachronic stage in which a set of purely phonetic constraints (essentially resembling the analysis sketched in section 1) results in a pattern of variable post-vocalic dorsal spirantization, with higher probabilities when flanked by low vowels or in fast speech, etc. A corpus of data reflecting this variable pattern is the input to our exemplar-based model. Let there be a particular lexical item (a cohort of exemplars with highly uniform semantic and phonetic representations) with a preponderance of post-vocalic dorsal stop rather than fricative realizations. In speech production, activation flows from the semantic units to this cohort of exemplars, and thence to the perceptual and articulatory representation associated with these exemplars. But then the perceptual echo probe kicks in, spreading activation to all exemplars, proportional to their perceptual similarity to the initially activated cohort. Now, assuming that similarity with

respect to containing a vowel + dorsal obstruent sequence results in a significant number of shared perceptual nodes, the perceptual echo probe will spread activation to all exemplars containing such sequences. Given the greater number of post-vocalic fricative exemplars, this will result in some amount of "pull" on the output towards a fricative realization as well. Particularly if the target lexical item is low-frequency (i.e. not having many exemplars of its own), this pull may be sufficient to override the (initially activated) non-continuant realization plan, forcing the output into line with the broader lexical pattern of post-vocalic dorsal fricatives. Outputs of the system are, in turn, stored as new exemplars.[13] Thus, even if the system initially generates outputs with a highly variable pattern, as the cycle of output generation and exemplar storage is iterated, the lexicon comes to contain a greater and greater proportion of exemplars reflecting post-vocalic spirantization, further reinforcing the strength of the pattern, until an unspirantized output becomes impossible, even in careful speech. The period during which the pattern gains strength corresponds to the phenomenon of *lexical diffusion*, or gradual extension of stability of the pattern, from high to low frequency words within the lexicon, and across the speech community.

In sum, the spread of activation in the exemplar-based speech processing system gives rise to an analogical pressure, enforcing extension of sound patterns to similar items. This analogical pressure is distinct from the phonetic constraints which initially induced the pattern (though LAZY, of course, remains active in speech production at all stages, alongside the analogical constraint). Consequently the analogical constraint may seize upon a functionally irrelevant, but highly correlated, property, such as the presence of a vowel. As this property does not vary across tokens, the pattern becomes stable. Thus, the typical conditioning of phonologized patterns by a relatively small set of categorical distinctions (i.e. the limited distinctive feature set of standard phonological theory) appears to follow from this model as well. These properties, being relatively salient, and being present across a preponderance of tokens, have a strong effect on the evaluation of similarity; consequently they give rise to strong similarity-based

[13] The diachronic progression I'm interested in modeling is presumably something that goes on within a speech community. Using a single speech processing system to simulate of this state of affairs amounts, then, to assuming a speech community of one member, who constantly talks to himself. A more realistic simulation (introducing a greater possibility of heterogeneity in the outcome), would involve building a number of these systems, and making them "talk" to one another (i.e one system generates an output, which the others then store as a new exemplar, and so on).

pressures on outputs to conform to patterns in which they play a role. The proposal is thus a resuscitation and modification, to synchronic ends, of the Neogrammarian distinction between phonetically-driven and analogically-driven linguistic regularities. Thenceforth, the phonologized pattern is replicated, being strongly instantiated in the exemplars of lexical items learned by successive generations of learner/speakers, until further phonetic changes, borrowings, etc. disrupt the regularity of the pattern, and its productiveness wanes, or ceases entirely, passing, so to speak, into the graveyard of diachrony.

Now let us consider the behavior of the exemplar-based model with regard to some phonological pattern which has become shot through with exceptions, but where these exceptions take the form of consistent subgeneralizations. The pattern of post-vocalic dorsal spirantization may then become too weak to be extended to new outputs in the general case. But if the pattern still remains strong, for example, with respect to some morphosyntactically defined class, the result will be the *morphologization* of what was previously a general phonological pattern (e.g. as in consonant mutation paradigms). Activation of the morphosyntactically defined class in this model is due to the combined effect of semantic (including morphosyntactic) and perceptual similarity. That is, initial activation of the particular lexical item results in some weaker activation of exemplars of the whole morphosyntactic class of which it is a member. In the perceptual echo probe, these exemplars receive further activation due to their perceptual similarity (i.e. they strongly instantiate the pattern of post-vocalic dorsal spirantization); consequently, they still may contribute sufficient pressure towards the spirantized realization just in case the output is a member of the morphosyntactic class. The entry of "unnatural," language-specific conditioning into phonological patterns[14] thus receives a natural treatment, without requiring the introduction of language-specific constraints to the constraint set. What is universal – what therefore belongs in the constraint set -- is the analogical pressure to extend patterns: the particular language-specific patterns follow from the application of this analogical constraint to the contents of a particular lexicon.

[14] This includes morphologized processes as discussed above, and "telescoping" of a series of natural sound changes, cf. Anderson's (1981) discussion of Icelandic palatalization.

4. On Capturing Certain Additional Notions of Standard Phonological Theory.

4.1. I/O Faithfulness

Recall that, in speech production, the model initially activates a semantically defined cohort of exemplars, which is then modified by the perceptual echo probe. This is strikingly analogous to the standard theory's notion of an underlying representation as the input to phonological computation. Moreover, the similarity-based activation of the perceptual echo probe is analogous to the enforcement of I/O similarity by the set of I/O faithfulness constraints.

4.2. Paradigmatic (O/O) Faithfulness and Morphological Productivity

On the other hand, it is not just the exemplars of a particular *word* which initially receive activation. All exemplars are activated, to the extent that they are semantically similar to the target expression. Among the semantically similar items to a morphologically complex word are its base, and other members of paradigms to which it belongs. Thus, for example, if the target semantic representation is [[condensation]], activated exemplars will include tokens of *condense* as well as well as of *condensation*. This gives rise to the possibility of paradigmatic faithfulness effects, such as base/derivative correspondence. Translating from OT analyses of cyclicity effects, we can surmise that it is the extra strength of the pattern instantiated in the base, of "a full vowel in the [dɛ̃n(s)] syllable," that allows [ˌkʰãndɛ̃n'seiʃən] to resist the more general lexical pattern of stressless vowel reduction.

We can also consider affixal paradigms in this light, such as the cohort of all plural forms ending in sibilant fricatives (standardly analyzed in terms of suffixation of the /-z/ morpheme): phonological similarity is much lower in this diffuse class, but the number of items in the paradigm is vastly larger, making it a very strong pattern. Thus, even if the lexicon does not already contain an exemplar of a plural form for some noun, e.g. /ˈɪnfowməˑʃɬ/ ('infomercial'), it can nevertheless generate a plural form [ˈɪnfowməˑʃɬz] as an effect of the strength of the final-sibilant pattern among plural nouns. The "family resemblance" phenomenon of irregular inflectional morphology, e.g. in the English strong verbs, can be handled with the same general pattern extension mechanism, as has been argued by proponents of "single-

system" morphological processing, e.g. Rumelhart & McClelland 1986, Bybee 1995, Albright 1999. Moreover, note that Kurylowicz' (1949) second law, i.e. the observation of typically more phonologically uniform realizations of morphological bases than of affixes, can now be seen as an effect of token frequency, under the assumption that bases, such as [k^hæt] typically have higher token frequency than derived forms such as [k^hæts]. (Greater token frequency, in this model, allows particular words to withstand the pressure to conform to broader lexical patterns.)

4.3. Treatment of Phonological Contrast

One desirable implication of OT's notion of *faithfulness* is that the phonological grammar can play an integral role in speech perception and production. For some phonetic property P, the extent to which this property resists variation/neutralization in production, and plays a dispositive role in word recognition (that is, its contrastive status) corresponds to the *ranking* of IDENT(P), relative to the other constraints in the grammar (see Kirchner 1997, 1998; Boersma 1998; cf. Smolensky 1996). In other words, ranking of IDENT(P) models the speaker/hearer's *attention* to P in speech processing.

The counterpart to this ranking in the exemplar-based model is *aggregation of exemplars*.[15] If property P is contrastive in the language, the lexicon will be partitioned fairly cleanly into sets of words (semantically uniform cohorts) which are either P or ¬P. For example, exemplars of *pat* will have a relatively uniform long VOT realization of the initial stop, and exemplars of *bat* will have a relatively uniform short-to-negative VOT. It follows that a new perceptual input [p^hæt] will strongly activate the *pat* cohort, whereas activation of the *bat* cohort will be much weaker, due to the inhibitory "long VOT" connections between the input and the *bat* cohort. That is, given this distribution of VOT among semantic cohorts, VOT will play a significant role in perception: *speakers/hearers attend to VOT distinctions as a function of such a phonological system.* On the other hand, if VOT is not contrastive, there will be no such partitioning of cohorts. The cohort of a word might contain, for example, 50% [p^hæt]

[15] In addition, the phonetic constraints, such as Lazy, must be ranked (or, since this model does not incorporate strict domination, weighted) relative to the lexical analogical constraint, and to one another. Also, note that the OT treatment conflates inherent salience asymmetries with language-specific attention asymmetries. In the exemplar-based model, salience asymmetries are reflected directly in the perceptual representation.

and 50% [bæt] exemplars. In response to a new perceptual input [p^{h}æt], the "long VOT" node will contribute nothing to the activation of this cohort, since the excitatory VOT connections the other [p^{h}æt] examplars will be canceled out by the inhibitory VOT connections to the [bæt] exemplars: *speakers/hearers disregard VOT distinctions as a function of such a phonological system.*

A remaining puzzle, though, is why sound systems maintain contrasts even in the absence of a minimal pair. For example, maintenance of the VOT distinction is necessary to distinguish *pat* from *bat*; but why are words like [p^{h}iæ̆now] ('piano') not found in free variation with [biæ̆now], since there is no word *biano*? The standard response (including Flemming 1995, Kirchner 1997), is to define contrast in terms of the set of *possible* words, which does include [biæ̆now]. Indeed, for Flemming, the set of possible words must be explicitly evaluated by the grammar, because the constraints refer directly to the maintenance of contrast, and to maintenance of perceptual distance between contrasting possible words. As Flemming notes, this distance-between-contrasting-possible-words approach is more appealing than the sort of perceptually-based fortition constraints posited in Kirchner 1998, because perceptually indistinct sounds are not marked per se (as the fortition-constraint approach forces us to treat them); rather, they are marked *as distinctions between contrastive forms*. Thus, for example, centralized vowels (which are indistinct in the F2 frequency (i.e. front/back) dimension) are only marked in vowel systems which employ a front/back contrast; in "vertical" vowel systems (with only height contrasts), such as Marshallese, centralized vowels appear to be unmarked.

But it is curious that such a fundamental role in phonological evaluation is played by the set of possible words, when knowledge of this set is so peripheral to the task of speech processing. Rather, it is knowledge of the set of *actual* words that is fundamental to perception and production. Moreover, there is reason to believe that judgments as to the well-formedness of possible words are based on extrapolation from the lexicon of actual words (cf. Frisch & Zawaydeh's (1997) finding that Arabic speakers' judgments of well-formedness for novel items (specifically, as regards OCP violations) are gradient, and correlate highly with the frequency of similar forms in the Arabic lexicon; cf. Ross-Zuraw 1998, with similar findings regarding the well-formedness of nonce forms involving nasal substitution in Tagalog). This result appears to be inconsistent with Flemming's (and all standard

phonological theories') assumption that the well-formedness of possible words is directly evaluated by the grammar, without reference to the words' actual presence in the lexicon.

The new insight afforded by the exemplar-based approach is that contrasts which are necessary to distinguish *actual* words can be generalized to words which lack minimal pairs, as an effect of analogical pattern extension. For example, given the English lexicon, with a large numbers of [p^h] vs. [b̥]~[b] minimal pairs of words (e.g. *pat*/*bat*), there is a strong pattern of labial stops (and more generally, stops at all places of articulation) being realized either with long VOT ([p^h]), or with short-to-negative VOT ([b̥]~[b]). The strength of this pattern is presumably sufficient to rule out [piæ̃now] (with intermediate VOT) as the typical realization of *piano*; and the pattern extension mechanism, in its capacity as I/O faithfulness enforcer (see section 4.1), forces particular lexical items, such as *piano*, to choose between the long and the short-to-negative VOT realizations: [biæ̃now] as a realization of *piano* would presumably be too dissimilar to the existing *piano* exemplars, and so *piano* is stably realized as [p^hiæ̃now]. Thus, the VOT contrast is generalized beyond items in which its maintenance is necessary to avoid homophony. On the other hand, in a sound system which lacks a VOT contrast, there are no minimal pairs for VOT, hence no pattern of polarized VOT values, hence any given lexical item can vary around intermediate VOT values (presumably whatever voicing configuration best satisfies LAZY in a given context), as observed in Keating 1990. Thus, the phenomena of contrast maintenance and perceptual dispersion of contrastive categories now derive from the more functionally plausible constraint of HOMOPHONY AVOIDANCE,[16] referring to actual words such as *pat* and *bat*. The set of atomized perceptual dispersion and contrast maintenance constraints are accordingly eliminated.

Further note that stop "voicing" contrasts frequently involve a combination of phonetic distinctions, viz. in closure voicing, VOT, and closure duration. Flemming observes that such cue multiplicity is

[16] Independent motivation for a principle of homophony avoidance, as distinct from contrast maintenance, is provided by Crosswhite 1997 and Wright 1996. Crosswhite reports that in Trigrad Bulgarian, reduction of suffixal vowels is blocked just in case the reduction would result in homophony. Similarly, Wright's study indicates that speakers have a greater tendency to hyperarticulate words from dense lexical "neighborhoods" (i.e. with many phonetically similar words, with which the target word is therefore readily confusable) than words from sparser neighborhoods. That is, perceptual distinctness can be relaxed, provided that this is not likely to result in confusion of actual words.

typical of phonological contrasts, improving their perceptual robustness. The puzzling point, though, is that this collection of cues has no unified auditory characterization. Rather, Flemming observes that the basis for particular combinations of cues lies in their typical cooccurrence due to articulatory considerations.[17] However, these ancillary consequences may become integral to the contrast, being maintained even in contexts where they are not articulatorily necessary (e.g. the closure and VOT cues associated with active devoicing, even in contexts where the stop would devoice passively). Flemming handles this phenomenon by stipulating particular combinations of cues in his perceptual distance constraints. But with an analogical pattern extension constraint, a more elegant story is possible. For articulatory reasons, particular values of, for example, closure voicing, VOT, and closure duration, typically cooccur in exemplars, giving rise to a strong pattern. The pattern is then fully generalized by analogical extension.

5. Conclusion

The appeal of this general approach is that it appears to permit a radical simplification of the constraint set, identified with the set of general, independently motivated functional considerations (including, of course, phonetics) which bear on the use of spoken language as a system of communication. To make good on this promise, the model must be computationally implemented, and its behavior tested.

References

Albright, A. (1999) Italian Verbs. Paper presented at LSA, January 1999, Los Angeles.

Anderson, S. (1981) Why Phonology Isn't Natural. *Linguistic Inquiry* 12:4, 493-539.

Beckman, J. (1997) Positional Faithfulness, Doctoral Dissertation, UMass-Amherst.

Boersma, P. (1998) Functional Phonology. Amsterdam: Landelijke Onderzoekschool Taalwetenschap.

Bybee, J. (1995) "Regular Morphology and the Lexicon," *Language and Cognitive Science* 10:425-55.

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

Derwing, B. & R. Skousen (1994) Productivity and the English Past Tense: Testing Skousen's Analogy Model. In S. Lima,

[17] For example, absence of closure voicing and positive VOT are both consequences of the glottal abduction gesture; and the longer closure duration is presumably attributable to the more fortis closure required to resist the greater oral air pressure occurring when the glottis is abducted.

R. Corrigan & G. Iverson (eds.) *The Reality of Linguistic Rules.* Amsterdam: J. Benjamins.

FIDELHOLTZ, J. (1975)

FLEMMING, E. (1995) Auditory Features in Phonology. Doctoral Dissertation, UCLA.

———. (1997) Phonetic Optimization, Compromise in Speech Production, paper presented at Hopkins OT Workshop/University of Maryland Mayfest 1997, Baltimore.

FRISCH, S. & ZAWAYDEH, B. (1997). Experimental evidence for abstract phonotactic constraints. *Research on spoken language*

GOLDINGER, S. (1997) Perception and Production in an Episodic Lexicon, in Johnson & Mullennix (1997).

GORDON, M. (in progress) Syllable Weight: Phonetics, Phonology and Typology, Doctoral Dissertation, UCLA.

GRAMMONT, M. (1933) *Traité de Phonetique*. Paris: Delgrave.

HAYES, B. (1997) Phonetically-Driven Phonology: The Role of Optimality Theory and Inductive Grounding. Ms. UCLA.

HYMAN, L. (1975) *Phonology: Theory & Analysis*. New York: Holt, Rinehart & Winston.

JOHNSON, K. & J. MULLENNIX (1997) *Talker Variability in Speech Processing.* San Diego: Academic Press.

JOHNSON, K. (1997) Speech Perception without Speaker Normalization, in Johnson & Mullennix (1997).

JUN, JONGHO. 1995. *Perceptual and Articulatory Factors in Place Assimilation: An Optimality-theoretic Approach*. PhD thesis, Los Angeles: UCLA.

KAUN, A. (1994) The Typology of Rounding Harmony. Doctoral Dissertation, UCLA.

KENSTOWICZ, M. (1982) Gemination and Spirantization in Tigrinya. Studies in the Linguistic Sciences 12:1, 103-122.

KIRCHNER, R. (1997) Contrastiveness and Faithfulness, Phonology 14:1, 83-111.

———. (1998) *An Effort-Based Approach to Consonant Lenition*. Doctoral Dissertation, UCLA.

KOHLER, K. (1991) The Phonetics/Phonology Issue in the Study of Articulatory Reduction, Phonetica 48, 180-192.

KURYLOWICZ, J. (1949) La Nature des Procès Dit 'Analogiques'. *Acta Linguistica* 5: 121-38.

LINDBLOM, B. (1983) Economy of Speech Gestures, in MacNeilage, P. (ed.), Speech Production, New York, Springer Verlag.

MACEACHERN, M. (1996) Laryngeal Similarity Effects in Quechua and Aymara, WCCFL 15.

MYERS, J. (1996) Canadian Raising and the Representation of Gradient Timing Relations, paper presented at MCWOP, University of Illinois, Urbana-Champaign.

OSTHOFF, H. & K. BRUGMANN (1878) *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen.*

PRINCE, A. & P. SMOLENSKY (1993) Optimality Theory: Constraint Interaction in Generative Grammar. Ms. Rutgers University, Johns Hopkins University.

processing: Progress report 21. Bloomington, IN: Speech Research Laboratory, Indiana University. 517-529.

ROSS-ZURAW, K. (1998) Knowledge of Lexical Regularities: Evidence from Tagalog nasal substitution. Paper presented at LSA, January 1999, Los Angeles.

RUMELHART, D. & J. MCCLELLAND (1986) "On learning the past tenses of English verbs," in J. McClelland and D. Rummelhart, eds., *Parallel Distributed Processing,* v. 2, Cambridge, MA: MIT Press.

SILVERMAN, D. (1995) Phasing and Recoverability. Doctoral Dissertation, UCLA.

SKOUSEN, R. (1989) *Analogical Modeling of Language.* Dordrecht: Kluwer.

———. (1992) *Analogy and Structure.* Dordrecht: Kluwer.

SMOLENSKY, P. (1996) On the Comprehension/Production Dilemma in Child Language. Ms. Johns Hopkins University.

STAMPE, D. (1972) *How I Spent My Summer Vacation*. Doctoral Dissertation, U. Chicago.

STERIADE, D. (1993) Positional Neutralization. Ms. UCLA.

———. (1995) Neutralization and the Expression of Contrast. Ms. UCLA.

———. 1996. Paradigm uniformity and the phonetics-phonology boundary. Paper presented at the fifth conference on Laboratory Phonology. Evanston, Illinois.