Assembling Phonological Representations

Christo Kirov
New York University
cnk218@nyu.edu

Adamantios Gafos
New York University & Haskins Laboratories
ag63@nyu.edu

December 13, 2007

1 Introduction

When speaking words, a person must retrieve the phonological representation of a target lexical item by assembling a set of parameter values that specify the required vocal tract action. In the cognitive psychology of speech, chronometric studies have shown that the time to retrieve words depends on a number of factors. For instance, measuring picture-naming latencies in word production has shown that words with many phonological neighbors (many similar words) are faster to retrieve than words with fewer neighbors (Vitevitch, 2002b).

In the first part of this paper, we present a dynamical model for phonological planning. The model traces the time-course of planning from the semantic selection of a lexical item to the selection of a sequence of phonemes, viewed here as phonetic targets, to be produced. By design, it incorporates information about lexical factors such as lexical frequency and phonological neighborhood density, and thus makes predictions about how these factors affect speech planning. Motivated by an explicit focus on the often-ignored temporal dimension of speech planning, the proposed model is based on the mathematical formalism of nonlinear dynamical systems. In particular, a discrete formulation of the dynamic field theory of Erlhagen and Schöner (2001) was used. The model is characterized by a few abstract principles that apply irrespective of its three internal levels of representation, word units, phoneme units and feature units. Nevertheless, it allows us to make reasonably accurate predictions about the time-course of phonological assembly, and it provides an intuitive and formally explicit understanding of key concepts implicated in the experimental literature (e.g. competition and facilitation).

In the second part of the paper, we extend our model to the domain of diachronic changes in phonological representations. This is the context of sound change where word representations evolve at a much slower time scale than that of word production. Diachronic changes in phonological representations accumulate gradually during repeated production-perception loops, that is, through the impact of a perceived word on the internal representation and subsequent production of that word. To capture these changes, we capitalize on the continuity of parameter values at the featural level. Our specific aim here is to illustrate our model by contrasting it with another recent view exemplified with an exemplar model of lenition, which also embraces continuity in its representational parameters.

The two parts share the concern of providing a formal basis of change in phonological representations using basic concepts from the mathematics of dynamical systems.

2 Synchronic Dimension: the time-course of speech planning

There exist formally explicit models with components devoted to the control and execution of speech movements (Guenther, 1995; Saltzman & Munhall, 1989; Browman & Goldstein, 1990). Our focus in this section is, instead, on the speech planning process, which takes place before movement execution. Thus, the output of the planning model we develop here can be seen as the input to the later phases of speech production, the neuromuscular and organic phases, during which a speech plan is articulated.

Our theoretical objective is to provide a formally explicit link between traditional approaches which emphasize the linguistically-significant dimensions of control for phonology, and cognitive psychology approaches which emphasize the process of accessing the lexicon and assembling the phonological representation of a word.

2.1 Model Architecture

The proposed model consists of a hierarchical dynamical system. Its hierarchical nature is reflected in its two connected levels of representation, word nodes and phoneme nodes, graphically depicted in Figure 1. Its dynamical nature is reflected in the mathematical tools used to model the workings of the system within its different levels as well as the relation between these levels. Specifically, each level is formalized by a dynamical system, and the different levels are coupled. That is, the dynamical system that governs decisions at the word level ties into the dynamical system that governs decisions at the phoneme level. In mathematical terms, a function of the **output**, $A(t)$, of one system serves as the **input** to the lower system. This coupling is shown in the following set of coupled equations.

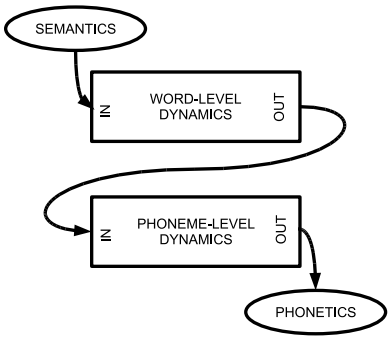

Figure 1: Input/output relationships, showing cascaded coupling between word level and phoneme level dynamics.

$$\tau dA_{word}(t) = -A_{word}(t) + r + input(t) + interaction(t) + noise \quad (1)$$

$$\tau dA_{phoneme}(t) = -A_{phoneme}(t) + r + f(A_{word}(t)) + interaction(t) + noise \quad (2)$$

τ is a constant that controls the rate at which the dynamics defined by an equation progress (a large τ means slow change over time). $A_{word}(t)$ is the activation of some word node at time t, and

$dA_{word}(t)$ is the change in activation at t. Similarly, $A_{phoneme}(t)$ is the activation of some phoneme node, and $dA_{phoneme}(t)$ is the change is activation. r is a resting activation which a given node will not fall below. $input(t)$ represents activation that flows into some word node from another system, such as semantic selection. $f(A_{word}(t))$, is similar to the $input(t)$ term for word nodes. It depicts the coupling between word and phoneme level dynamics since phoneme nodes receive input activation from word nodes. $interaction(t)$ represents activation (both excitatory and inhibitory), that spreads between the nodes on a particular level of dynamics (word nodes *interact* with each other, as do phoneme nodes). Finally, all dynamics are influenced by some amount of *noise*.

We now turn to describe the contents of the phoneme and word nodes in turn. Phoneme nodes are to be viewed as stored phonetic targets. As such, they contain a featural specification based on a set of standard features used to describe contrasts in the English phoneme inventory (Bailey & Ulrike, 2005). For consonants, these features were place (labial, dental, coronal, velar, or glottal), manner (stop, fricative, affricate, lateral, glide, rhotic, or nasal), voicing (voiced or unvoiced), and sonority (sonorant or obstruent). For vowels, the relevant features were height (high, mid, or low), frontness (front, central, or back), rounding (round or unround), and tenseness (tense or lax). Each phoneme node also has a resting level of activation directly related to the log of the phoneme's frequency of occurrence. In addition, each phoneme is linked to a set of neighbors, or phonemes with similar featural specifications. Each link to a neighbor in this set is weighted according to the feature-based "distance" between the node and the neighbor. This distance is based on the number of different features between two phonemes. For example, [p] is defined as [labial, stop, voiceless, obstruent] and [b] is defined as [labial, stop, voiced, obstruent]. The two differ by just one feature out of four, and thus the distance between them is 0.25.

Each word node also contains a resting activation based on the log of its frequency of occurrence. The phonological content of a word node is specified in an ordered list of phonemes that make up the word. For example, the word node representing "good" contains the phoneme list [[g][ʊ][d]]. Like each phoneme node, each word node is connected to a set of phonologically similar neighboring nodes. Following experimental convention, a neighbor is defined as a word differing by at most one phoneme. For example, "good" contains links to "goad" and "got".

In the proposed model, the connection between the word layer and the phoneme layer is one-directional. Activation flows from words to phonemes, but not vice versa. In addition, there are lateral connections within a layer. Neighbors are directly connected to each other.

Thanks to the information stored in word and phoneme nodes, we can use the model to examine the effects of various lexical factors on speech planning. Words in the model differ along the following dimensions: word frequency (instantiated as resting activation), neighborhood density (the number of neighboring nodes), and neighborhood frequency (the average resting activation over all a word's neighbors).

The proposed model also adds a third major structural component besides phoneme and word nodes: the production queue. The queue is an internal abstraction over phonological organization. It consists of a series of phoneme slots. Each phoneme has a channel in the queue, and can have a different activation at each slot. A graphical representation of the queue is shown in Figure 2. In general, activation across the queue's slots evolves in parallel, rather than in a serial slot-by-slot fashion. Eventually, some external cue dictates that it is time to articulate the contents of the queue. At this point, the executed phonetic output for a queue slot is a weighted mixture of the phonetic plans of the phoneme nodes at that slot.

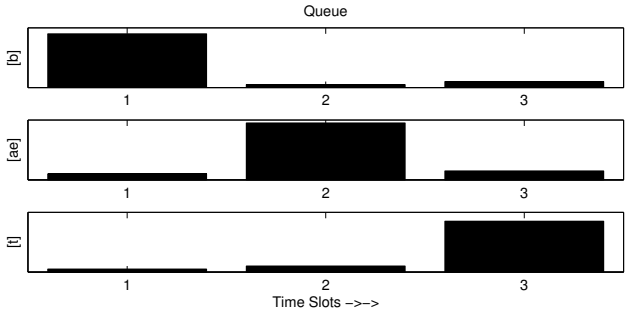

Figure 2: The queue with [b],[ae],and[t] channels shown, as they might look late in the planning stages of "bat".

2.2 Model Dynamics

Studying speech planning as a cognitive process occurring in time broadly motivates a dynamical approach to modeling. A *dynamical model* is a formal system whose internal state changes in a controlled and mathematically explicit way over time. The workings of the proposed model are based on a discrete version of a dynamical formalism called *Dynamic Field Theory* (Erlhagen & Schöner, 2002).

The activation of a word node in the model is governed by the following generalized differential equation:

$$\tau dA(t) = -A(t) + r + input(t) + interaction(t) + noise \tag{3}$$

where τ scales the rate at which the dynamical system evolves, $A(t)$ is the activation of the node at time t, $dA(t)$ is the change in activation of the node at time t, r is the resting activation of the node, $input(t)$ is any activation the node receives from an outside source (i.e. a node on a different layer) at time t, $interaction(t)$ represents the relationship between a node and its neighbors at time t, and $noise$ is a Gaussian random variable inducing stochastic behavior.

The equation can be broken down into simpler components to better understand how it functions. The core component $\tau dA(t) = -A(t) + r$ is an instance of exponential decay dynamics. In the absence of any input or interaction, the activation of a node will simply decay down to its resting level, r, as shown in Figure 3. If a node starts at resting activation, it will remain there forever. In the terminology of dynamical systems, the starting activation of a node is known as an *initial condition*, and the activation which it ends up at, in this case the resting activation, is known as an *attractor*. If the input term, $input(t)$, is non-zero, then the system will move towards a point equivalent to its resting activation plus the input term. The speed of the process is modulated using the τ term.

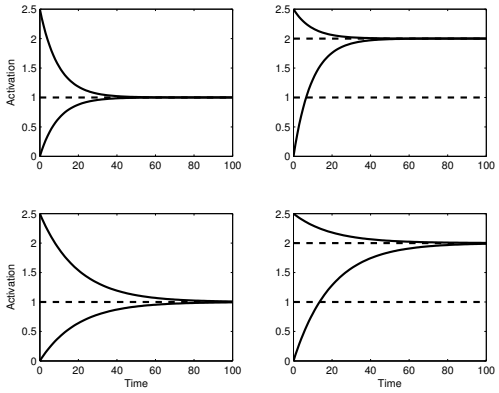

Figure 3: Top left: In the absence of input, node activation converges to resting level $r = 1$ (dashed line) ($\tau = 10$). Top right: With added input $input(t) = 1$, activation converges to resting level $r = 1$ plus input (top dashed line)($\tau = 10$). Bottom left: In the absence of input, node activation converges to resting level $r = 1$ ($\tau = 20$). Bottom left: With added input $input(t) = 1$, activation converges to resting level $r = 1$ plus input ($\tau = 20$).

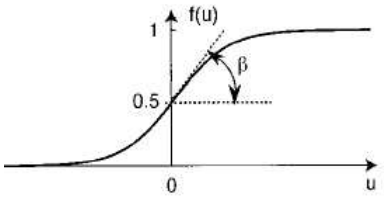

Figure 4: Threshold function $f(u)$. Threshold is the inflection point of the function (in this diagram 0). The slope β represents the "softness" of the threshold.

Input to a word node in the proposed model comes from the semantics. The speech planning process begins when one or more word nodes begin to receive input activation from the semantic system (i.e. they are selected for meaning). This input causes the activation of a word node to creep up over time.

The interaction term, $interaction(t)$, comes into play when a node's activation passes a certain threshold set by a sigmoid (s-shaped) or step function, as shown in Figure 4. Thus, its effects can be called nonlinear. Once the threshold is surpassed, the function of the interaction is to enhance the activation of a word and its close phonological neighbors while inhibiting the activation of other words. This functionality is graphically represented in Figure 5 from Schöner (2001). The mutual enhancement aspect of the interaction mirrors the support a word receives from its neighbors in other models of phonological planning, such as that of Dell & Gordon (2003). The crucial differences are twofold. In our model, the same interaction term provides for both enhancement and inhibition.

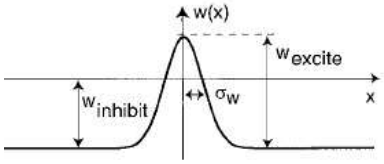

Figure 5: Schematic of interaction effect. All neighbors are inhibited by $w_{inhibit}$. However, close neighbors (those no farther than σ_w distance apart) receive mutual activation of w_{excite}, which means they receive a net positive activation from the interaction ($w_{excite} - w_{inhibit} > 0$).

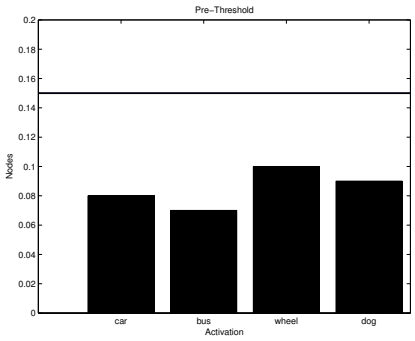

Figure 6: All nodes have started at roughly the same resting activation and are creeping towards the threshold while receiving input.

Moreover, the interaction is nonlinear in nature as shown in Figures 6 through 8.

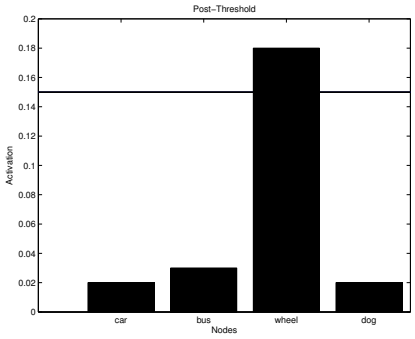

Figure 7: One node has crossed the threshold, and so pushes up on itself and down on its competitors.

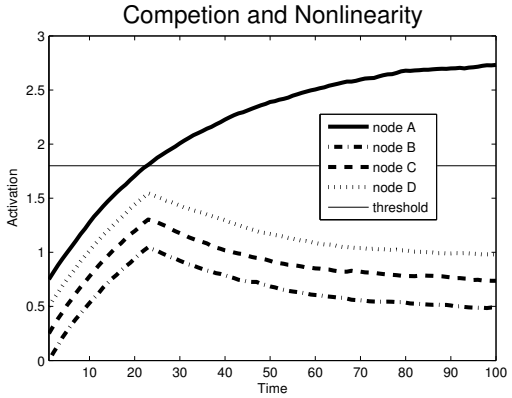

Figure 8: Competition among candidates in a nonlinear system. The activation trajectories of 4 nodes starting at different resting activation levels, all with equivalent input. The node with the highest resting level reaches the threshold first. Once this happens, it stabilizes its own activation and inhibits its competitors.

Phoneme nodes are governed by the same dynamics as word nodes, using the following equation:

$$\tau dA(t,x) = -A(t,x) + r + input(t,x) + interaction(t,x) + noise \tag{4}$$

The x parameters in the equations refer to slots in the queue. Each slot is governed by its own dynamical system.

Phoneme nodes receive input from their parent word nodes. At each timestep of the process, word nodes spread their own activation to their associated phoneme nodes in the appropriate timing slots of the queue. For example, "dog" will spread activation to [d] in slot 1, [ɑ] in slot 2, and [g] in slot 3. Its similar word "log" will spread activation to [l] in slot 1, [ɑ] in slot 2, and [g] in slot 3. In principle, several words can be active in parallel, and this can be seen as competition between word nodes to have their own phonemes be selected for each queue slot.

The interaction term is also nearly identical for phoneme nodes as it is for word nodes. It includes the same enhancement/inhibition function shown in Figure 5. This means that when one phoneme node crosses the activation threshold, it will quickly inhibit all of its competitors, as seen in Figure 8.

2.3 The Model in Action

Given the dynamics above, we expect a typical planning process to proceed as follows. First, one or more word nodes will receive input activation from the semantic system, causing their own activation to creep up towards the activation threshold from their resting levels. As soon as some node passes the threshold, the interaction term will take effect. This means that the words in the phonological neighborhood to which the word belongs will begin to mutually enhance their own activations, while suppressing phonologically unrelated words. It is important that the semantic input to the target word be high enough to overcome any advantages the target's neighbors may have in terms of higher resting activation or neighborhood frequency. At this point, the dynamics have effectively decided which word to produce.

Meanwhile, phoneme nodes will have been slowly increasing in activation at each slot of the queue from the input activation passed to them by their parent word nodes. Eventually, some phoneme in each slot will pass the activation threshold, and inhibit its competitors. Once this happens in every queue slot, the system will have decided which phonemes to produce. Equivalently, the system will have decided which phonetic plans should be implemented at execution phases of speech production.

The goal of the model is to illuminate how this planning process develops over time. Thus, the output of the model takes the form of certain metrics plotted against simulation time. The first metric is phonemic error. Phonemic error is the number of phonemes nodes that are incorrectly favored (i.e. that don't match the phonemes of the target word) at a given point in time. The "favored" phoneme at some slot in the queue at some point in time is just the most activated phoneme node at that time. Since each word consists of a discrete number of phonemes, phonemic error is always an integer value.

Typically, phonemic error will decrease as simulation time progresses until eventually reaching a minimum value. For a given production, this minimum value will not necessarily be zero, even if given an arbitrarily long simulation time. This is due to the nonlinear nature of the model dynamics. If the wrong phoneme happens to cross the decision threshold first, it could suppress the correct phoneme indefinitely through the inhibitory effects of the interaction kernel. We measure the time it takes to settle on a phonemic plan as the time step at which the minimum phonemic error is achieved over the course of a simulation run.

The second metric is entropy summed up over all the queue slots. Entropy in a certain queue slot is measured as the information entropy over the distribution of phoneme node activations in that slot. Even though some phoneme may be favored at a particular point in time, that does not mean its phonemic competitors are completely inactive. If the correct phoneme is favored, but many other nodes have roughly similar activations, it is assumed that the phonetic plan actually executed by the articulatory system will be some mixture of the specifications of all the highly activated phonemes, not just the target. As in information theory, maximum entropy, and thus maximum uncertainty, occurs when all nodes have equivalent activation. Low entropy, on the other hand, refers to a situation in which one node is very active, and the rest are essentially inactive. In such a situation, we can confidently say that the implemented phonetic plan will closely resemble

the plan associated with the most highly activated phoneme.

As an example of model output, the following graph shows the above metrics for the planning of the word "cat". It can be seen that the phonemic error drops quickly indicating that the correct phonemes become the most activated early in the planning process (solid line). However, it takes some time for the correct phonemes to reach threshold activation and suppress all other competitors (dotted line).

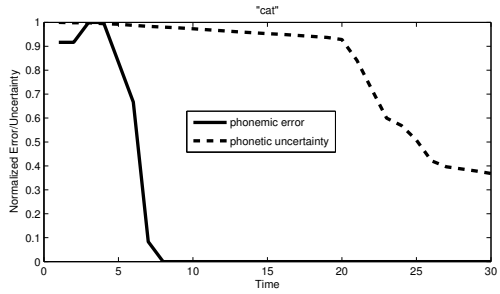

Figure 9: Ten averaged productions of "cat" showing different time scales for phonemic decision and subsequent certainty.

2.4 Simulations

In order to test the model, an artificial lexicon was constructed with the goal of creating a fairly realistic data set conforming to the distributional properties of a real lexicon.

The set of phonemes used to construct each lexical item was derived from the phonemes used by the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). It consists of 35 consonants and non-diphthong vowels from the full set of CMU symbols.

The lexicon itself consists of 1197 actual English words derived from the English Lexicon Project (http://elexicon.wustl.edu/). All words are monosyllabic, monomorphemic, and have a CVC structure. Only words defined by the CMU Pronouncing Dictionary that can be transcribed using the set of available phonemes were used.

The resting activations of word and phoneme nodes were derived from the logs of the Hyperspace Analogue to Language (HAL) word frequencies provided by the English Lexicon Project. The frequencies are based on the HAL corpus, which consists of approximately 131 million words gathered across 3,000 Usenet newsgroups during February 1995 (Lund & Burgess, 1996).

It has been shown that certain lexical factors affect the microchronic time-course of speech production. In particular, more frequent lexical items are produced more quickly and accurately than less frequent lexical items (Vitevitch & Sommers, 2003). In addition, lexical items with more similar sounding neighbors, or words in a dense phonological neighborhood, are produced more quickly and accurately than lexical items in a sparse phonological neighborhood (Vitevitch, 2002a; Dell & Gordon, 2003).

We assessed the ability of our proposed model to capture these results by simulating the planning of a series of words from the test lexicon, with 10 simulation runs per word. Appendix A shows the words used. The choice of words, as well as their categorization into high or low frequency and dense or sparse neighborhood, was taken from Vitevitch & Sommers (2003).

Figure 10 shows the average minimum phonemic error for high versus low frequency words, and the average time required to settle upon the minimum error. In accordance with the experimental results, high frequency words have a lower average minimum error and require less time to settle upon a phonemic plan than low frequency words.

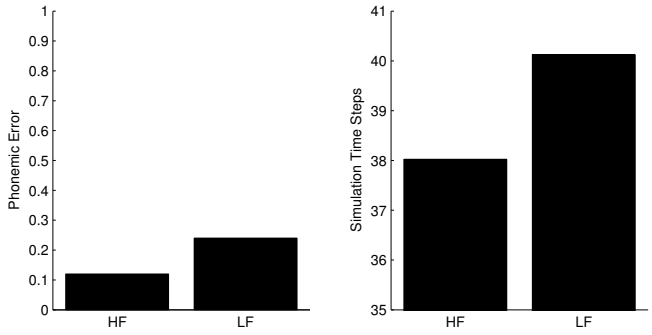

Figure 10: High versus low frequency.

Figure 11 shows the average minimum phonemic error for dense versus sparse neighborhood words, and the average time required to settle upon the minimum error. In accordance with the experimental results, words from a dense neighborhood have a lower average minimum error and require less time to settle upon a phonemic plan than words from a sparse neighborhood.

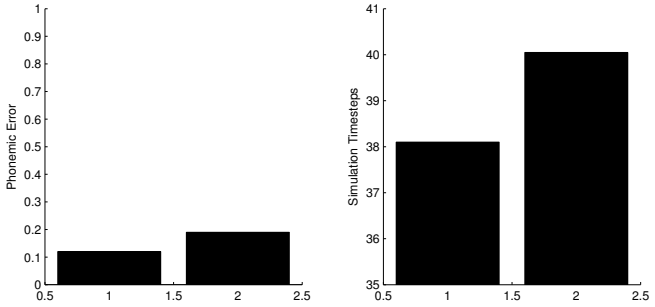

Figure 11: Dense versus sparse neighborhood.

3 Diachronic Dimension: the case of Lenition

The dynamical formalism employed in the previous section can serve as a basis for exploring the evolution of lexical representations at slower time-scales. We turn to this domain in this section, with the specific aim of capturing certain key properties of an unfolding lenition process.

Lenition is usually defined as a "weakening" sound change. One diachronic example is Grimm's law, according to which Proto-Indoeuropean voiceless stops became Germanic voiceless fricatives

(e.g. PIE *[t] > Gmc *[θ]). Grimm's law can be seen as part of a "weakening chain" of PIE sounds: d > t > θ > h > ∅ (Cser, 2003). More examples of historical sound changes assumed to be the result of diachronic lenition processes in different languages are given below (also from Cser, 2003). In each case, a stop has turned to a fricative similar in place of articulation to the original stop.

(1) a. Southern Italian dialects: [b d g]→[v ð ɣ] intervocalically.
b. Greek (Koine): [p^h t^h k^h]→[f θ x] except after obstruents.
c. Proto-Gaelic: [t k]→[θ x] intervocalically.
d. Hungarian: [p]→[f] word initially.

To fix some context, consider any single transition between two states of a lenition process, say, starting with a stop [b] and resulting in a fricative [v], [b] > [v]. At a broad level, one can describe two kinds of approaches to this kind of transition. The symbolic approach, as exemplified by Kiparsky's classic paper on linguistic universals and sound change (Kiparsky, 1968), studies the internal composition of the individual stages (e.g. feature matrices at each stage) and makes inferences about the nature of the grammar and the representations. The continuity of sound change, that is, how the representation of the lexical item containing a [b] changes in time to one containing an [v], is not studied. This is in part due to the theoretical assumption that representations are discrete. That is, there is no symbol corresponding to an intermediate degree of stricture between that of a stop and a fricative. In the dynamical approach, the transition process between the stages is studied at the same time as the sequence of stages. In what follows, we instantiate a small, yet core part of a dynamical alternative to the symbolic model of sound change.

3.1 An Exemplar Model of Lenition

It is useful to describe the main aspects of our model by contrasting it with another model proposed recently by Pierrehumbert. This is a model of sound change aimed at accounting for certain generalizations about lenition, extrapolated from observations of synchronic variation or sound changes in progress. The model proposed in Pierrehumbert (2001) has two attractive properties. It offers a way to represent the fine phonetic substance of linguistic categories, and it provides a handle on the effect of lexical frequency in the course of an unfolding lenition process.

In Pierrehumbert's discussion of lenition, it is assumed that the production side of a lenition process is characterized by the following set of properties.

Properties of Lenition

1. Each word displays a certain amount of variability in production.
2. The effect of word frequency on lenition rates is gradient.
3. The effect of word frequency on lenition rates should be observable within the speech of individuals; it is not an artifact of averaging data across the different generations which make up a speech community.
4. The effect of word frequency on lenition rates should be observable both synchronically (by comparing the pronunciation of words of different frequency) and diachronically (by examining the evolution of word pronunciations over the years within the speech of individuals.)
5. The phonetic variability of a category should decrease over time, a phenomenon known as entrenchment. The actual impact of entrenchment on lenition is not clear, and Pierrehumbert does not cite any data specific to entrenchment for this particular diachronic effect. In fact, while a sound change is in progress, it seems equally intuitive (in the absense of any data to the contrary) that a wider, rather than narrower range of pronunciations is available to the speaker. Pierrehumbert uses the example of a child's productions of a category becoming less variable over time, but this may only apply to stable categories, rather than ones undergoing diachronic change. It may also be orthogonal to the child's phonetic representations, and rather be due to an initial lack of biomechanical control. For these reasons, therefore, our own model is not designed to guarantee entrenchment while sound change is taking place, but does show entrenchment effects for diachronically stable categories.

Table 1: Properties of Lenition

The frequency related properties are based on previous work by Bybee who claims that at least some lenition processes apply variably based on word frequency (Bybee, 2003). Examples include schwa reduction (e.g. *memory* tends to be pronounced [mɛmri]) and t/d-deletion (e.g. *told* tends to be pronounced [tol]). Once a lenition process has begun, Bybee's claim amounts to saying that words with high frequency will weaken more quickly over time than rare words. Consequently, lenition effects can be seen both synchronically and diachronically. Synchronically, a more frequent word will be produced more lenited (with more undershoot) than a less frequent word in the current speech of a single person. Diachronically, all words in a language will weaken across all speakers, albeit at different rates.

What are the minimal prerequisites in accounting for the lenition properties above? First, it is clear that individuals must be capable of storing phonetic detail within each lexical item. We also need a mechanism for gradiently changing the lexical representations over time. To do this, the perceptual system must be capable of making fine phonetic distinctions, so that the information carried by these distinctions can reach the currently spoken item in the lexicon.

Pierrehumbert's exemplar-based model of lenition gives explicit formal content to each of these prerequisites (Pierrehumbert, 2001). The model is built on a few key ideas, which can be described in brief terms. Specifically, in the exemplar-based model, a given linguistic category is stored in a

space whose axes define the parameters of the category. In Pierrehumbert (2001), it is suggested that vowels, for example, might be stored in an F1/F2 formant space. This space is quantized into discrete cells based on perceptual limits. Each cell is considered to be a bin for perceptual experiences, and Pierrehumbert views each bin to be a unique potential exemplar. When the system receives an input, it places it in the appropriate bin. All items in a bin are assumed to be identical as far as the perceptual system is concerned, and the more items in a particular bin, the greater the activation of the bin is. All bins start out empty and are not associated with any exemplars that have actually been produced and/or perceived (memory begins as a *tabula rasa*). When a bin is filled, this is equivalent to the storage of an exemplar. The new exemplar is given a categorical label based on the labels of other nearby exemplars. This scheme limits the actual memory used by exemplars. There is a limited number of discrete bins, and each bin only stores an activation value proportional to the number of exemplar instances that fall into it. Thus, not all the exemplar instances need to be stored. A decay process decreases the activation of an exemplar bin over time, corresponding to memory decay. Figure 12, taken from Pierrehumbert (2001), shows the F2 space discretized into categorically labeled bins.

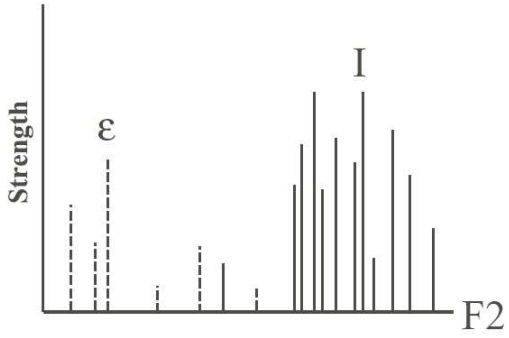

Figure 12: Exemplar bins with varying activations.

The set of exemplars with a particular category label constitutes an extensional approximation of a probability distribution for that category over the storage space. Given coordinates in the storage space, the distribution would provide the likelihood of a token at those coordinates belonging to some category (e.g. how likely is it that the token is an /a/). During production, a particular exemplar from memory is chosen to be produced, where the likelihood of being chosen depends on how activated the exemplar is. The chosen exemplar is shifted by a bias in the direction of lenition. This bias reflects the synchronic phonetic motivation for lenition. This includes at least a tendency to undershoot the degree of oral constriction in contexts favoring gestural weakening, e.g. in non-stressed syllables, syllable codas, or intervocalically. For relevant discussion see Beckman *et al.* (1992) and Wright (1994). To account for entrenchment (see Table 1(5)), Pierrehumbert extends this production model by averaging over a randomly selected area of exemplars to generate a produced candidate. Since the set of exemplars defines a probability distribution (in an extensional sense), weighing the average by each exemplar's probability results in a production candidate pushed toward the center of the distribution.

The exemplar scheme described in this section derives the five properties of lenition discussed

earlier as follows. Variability in production is directly accounted for since production is modeled as an average of the exemplar neighborhood centered around a randomly selected exemplar from the entire set stored in the system. Each lexical item has its own exemplars, and each production/perception loop causes the addition of a new exemplar to the set. This new exemplar is more lenited than the speaker originally intended due to biases in production, so the distribution of exemplars skews over time. In a given period of time, the number of production/perception loops an item goes through is proportional to its frequency. Thus, the amount of lenition associated with a given item shows gradient variation according to the item's frequency (Dell, 2000). As all processes directly described by the exemplar model occur within a single individual, lenition is clearly observable within the speech of individuals. Diachronically, lenition will proceed at a faster rate for more frequent items because they go through more production/perception loops in a given time frame. The synchronic consequence of this is that at a point in time, more frequent items will be more lenited in the speech of an individual than less frequent items. Finally, entrenchment is a consequence of averaging over several neighboring exemplars during production, shifting the resulting production towards the mean of the distribution described by all the exemplars.

In sum, the exemplar-based model offers a direct way to represent the fine phonetic substance of linguistic categories. The model also offers a way to capture the assumed effects of frequency on the unfolding lenition process. Pierrehumbert further claims that the exemplar model is the only type of model that can properly handle the above conception of lenition (Pierrehumbert, 2001, 147). In what follows, we will propose a model of lenition that does not depend on exemplar theory. At a broad level, a salient difference between exemplar-based models and the model to be proposed is that in the latter there is no need for storing an arbitrarily large amount of exemplars.

3.2 A Dynamical Model of Lenition

A central component of our model is the spatio-temporal nature of its representations. Take a lexical item containing a tongue tip gesture as that for /d/. We can think of the specification of the speech movements associated with this gesture as a process of assigning values to a number of behavioral parameters. In well-developed models that include a speech production component, these parameters include constriction location and constriction degree (Guenther, 1995; Saltzman & Munhall, 1989; Browman & Goldstein, 1990). A key idea in our model is that each such parameter is not specified exactly but rather by a distribution depicting the continuity of its phonetic detail.

Although our model does not commit to any specific phonological feature set or any particular model for the control and execution of movement, to illustrate our proposal more explicitly let us assume the representational parameters of *Articulatory Phonology* (Saltzman & Munhall, 1989; Browman & Goldstein, 1990). Thus, let us assume that lexical items must at some level take the form of gestural scores. A gestural score, for current purposes, is simply a sequence of gestures (we put aside the intergestural temporal relations that also must be specified as part of a full gestural score). For example, the sequence /das/ consists of three oral gestures - a tongue tip gesture for /d/, a tongue dorsum gesture for /a/, and a tongue tip gesture for /s/. Gestures are specified by target descriptors for the *vocal tract variables* of *Constriction Location* (CL) and *Constriction Degree* (CD), parameters defining the target vocal tract state. For example, /d/ and /s/ have the CL target descriptor {alveolar}. The CD descriptor of /d/ is {closure} and for /s/ it is {critical}. These descriptors correspond to actual numerical values. For instance, in the tongue tip gesture of a /d/, {alveolar} corresponds to 56 degrees (where 90 degrees is vertical and would correspond to a midpalatal constriction) and {closure} corresponds to a value of 0 mm (in fact, a negative value

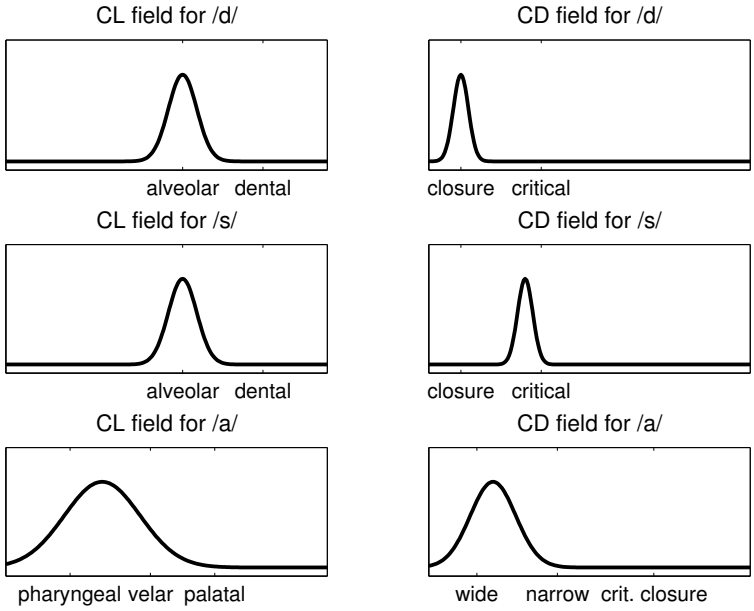

Figure 13: Component fields of /d/, /s/, and /a/. y-axis represents activation. /d/ and /s/ have nearly identical CL fields, as they are both alveolars, but they differ in CD.

-3.5 mm is used to model surface compression).

In our model, each parameter is not specified by a unique numerical value as above, but rather by a continuous activation field over a range of values for the parameter. The field captures among other things a distribution of activation over the space of possible parameter values so that a range of more activated parameter values is more likely to be used in the actual execution of the movement than a range of less activated parameter values. The parameter fields then resemble distributions over the continuous details of vocal tract variables. A lexical item therefore is a gestural score where the parameters of each gesture are represented by their own fields. Schematic fields corresponding to the (oral) gestures of the consonants in /das/ are given in Figure 13.

Formally, parameters are manipulated using the familiar dynamical law from dynamic field theory (Erlhagen & Schöner, 2002). Thus, the basic dynamics governing each field are described by:

$$\tau dp(x,t) = -p(x,t) + h + input(x,t) + noise \tag{5}$$

where p is the field in memory (a function of a continuous variable x), h is the field's resting activation , τ is a constant corresponding to the rate of decay of the field (i.e. the rate of memory decay), and $input(x,t)$ is a field representing time dependent external input to the system (i.e. perceived token) in the form of a localized activation spike. Unlike the discretized version of this equation used in section 2, x is a continuous variable (a point along a field rather than an individual node). We can conceive of every point along the x-axis of the field as governed by its own exponential decay equation.

Fields are spatio-temporal in nature. Thus specifying the value of a gestural parameter is a

spatio-temporal process in our model. We describe each of these aspects, spatial and temporal, in turn. The spatial aspect of the gestural specification process corresponds to picking a value to produce from any of the fields in Figure 13, e.g. choosing a value for Constriction Location for /d/ and /s/ from within the range of values corresponding to the [alveolar] category. This is done by sampling the Constriction Location field, much as we might sample a probability distribution. Since each field encodes variability within the user's experience, we are likely to select reasonable parameter values for production. A demonstration of this is shown in Figure 14. The noisy character of the specification process allows for variation in the value ultimately specified, but as the series of simulations in Figure 14 verifies the selected values cluster reliably around the maximally activated point of the field.

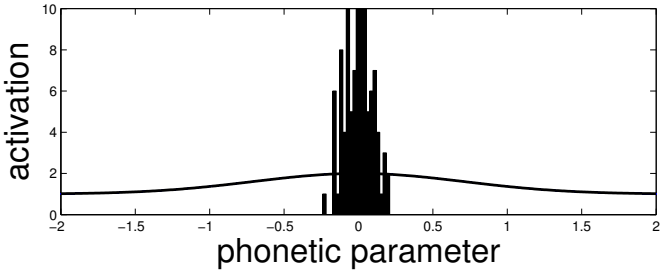

Figure 14: Variability in production. Histogram of selected values over 100 simulations of gestural specification on field with: $h_0 = 1$, $d = 5$, $\tau = 10$, $\omega = 0.05$, and a positive δ value of 1.1. Histogram overlaid on top of field to show clustering of selected values near the field maximum.

The specification process presented here is similar but not identical to the sampling of a probability distribution. Fields have unique properties that make them useful for modeling memory. Unlike distributions, fields need not be normalized to an area under the curve of one. The key addition here is the concept of *activation*. Fields can vary from one another in total activation while keeping within the same limits of parameter values. Because of this added notion of activation, the specification process is more biased towards the maximally activated point in the field (i.e. the mean of the distribution) than a true random sampling would be. This leads to an entrenchment effect for categories not undergoing change. This behavior is shown in Figure 15. In addition, fields have a resting activation level (a lower-limit on activation). This level slowly tends to zero over time, but increases every time the field is accessed during production or perception. Thus, lexical items whose fields are accessed more frequently have higher resting activation levels than lexical items whose component fields are accessed less frequently. Finally, much as memory wanes over time, activation along a field decays if not reinforced by input.

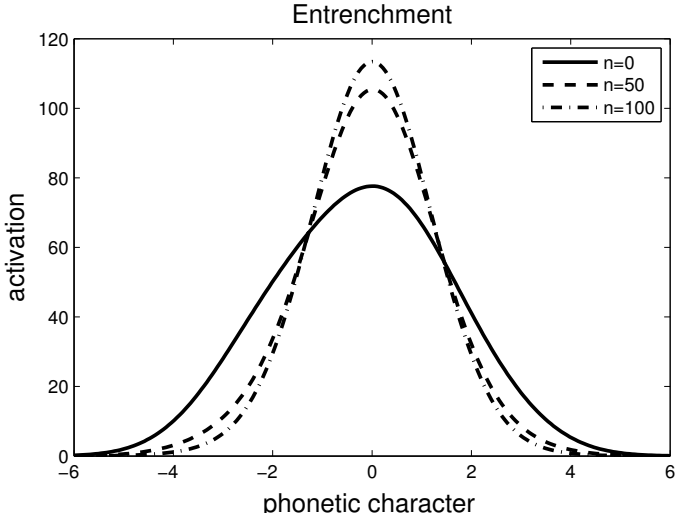

Figure 15: Output of entrenchment simulation. The x-axis represents a phonetic dimension (e.g. constriction degree). The field defining the distribution of this parameter is shown at various points in time. As time progresses, the field becomes narrower.

The other crucial aspect of the specification process is its time-course. Formalizing gestural parameters with fields adds a time-course dimension to the gestural specification process. If a lexical representation contains a /d/, the CD and CL parameters for this /d/ are not statically assigned to their (language- or speaker- specific) canonical values, e.g., CL = [alveolar]. Rather, assigning values to these parameters is a time-dependent process, captured as the evolution of a dynamical system over time. In short, lexical representations are not static units. This allows us to derive predictions about the time-course of choosing or specifying different gestural parameters.

The specification process begins by a temporary increase in the resting activation of the field, i.e. pushing the field up, caused by an intent to produce a particular lexical item (which includes a gesture ultimately specified for a parameter represented by this field). Resting activation increases steadily but noisily until some part of the field crosses a decision threshold and becomes the parameter value used in production. This scheme ensures that the areas of maximum activation are likely to cross the decision threshold first. After a decision has been made, resting activation returns to its pre-production level. The following equation represents this process mathematically:

$$\tau dh/dt = -h + h_0 + \delta(d, max(p)) * h + \omega * noise \tag{6}$$

where h is the resting activation during production, τ is a time scaling parameter, h_0 is the pre-production resting activation, $\delta(d, max(p))$ is a nonlinear sigmoid or step function over the distance between the decision threshold d and the maximum activation of field p, and $\omega * noise$ is scaled gaussian noise. While the distance is positive (the decision threshold has not yet been breached), the δ function is also positive and greater than 1, overpowering the $-h$ term and causing a gradual increase in the resting activation h. When the decision threshold is breached, the δ function becomes 0, and remains clamped at 0 regardless of the subsequent field state, allowing the $-h$ term to bring activation back to h_0.

The gestural specification process is affected by the pre-production resting activation of the field, in that a field with high resting activation is already "presampled", and thus automatically closer to the decision threshold. This leads to faster decisions for more activated fields, and by extension more frequent parameter values. The relevant simulations are described below. Figure 16 shows representative initial fields, and Figure 17 shows the progression of the featural specification process over time. We see that given two fields identical in all respects except for resting activation, the field with the higher resting activation reaches the decision threshold first.

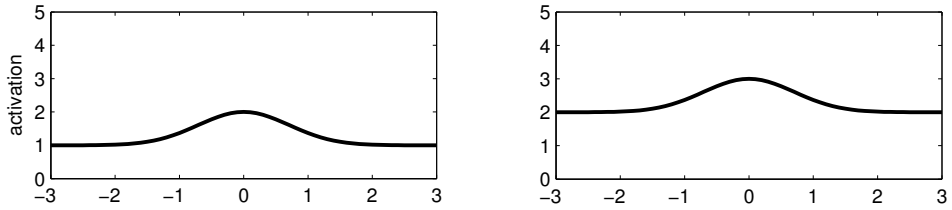

Figure 16: The two fields are identical except for resting activation: $h_0 = 1$ (left), $h_0 = 2$ (right). The x-axis is arbitrary.

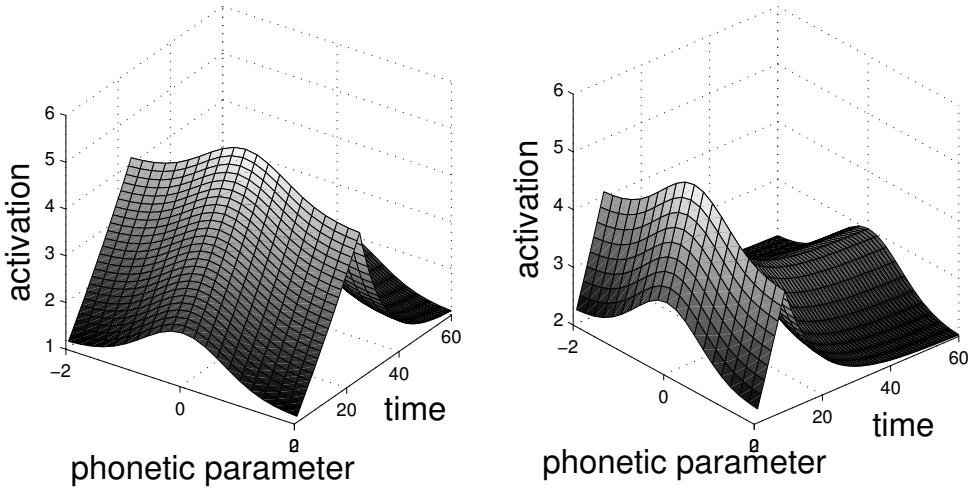

Figure 17: Sampling was simulated with a decision threshold $d = 5$, $\tau = 10$, ω=0, and a positive δ value of 1.1. The first field (left) reached the decision threshold at $t = 25$, and the second field (right) reached the decision threshold at $t = 9$ (where t is an arbitrary unit of simulation time). The field with higher initial resting activation reached the decision threshold faster. Both fields return to their pre-production resting activation after decision threshold is reached.

We now discuss the ways in which representing gestural parameters by fields relates to other proposals.

The field equation used in our model parallels the exemplar model in many ways, but encapsulates much of the functionality of that model in a single dynamical law which does not require the storage of exemplars. Memory wanes over time as the field decays, much as older exemplars are less activated in the exemplar model. Input causes increased activation at a particular area of the field, much as an exemplar's activation is increased with repeated perception. This activation decays with time, as memory does.

Perhaps the most crucial difference between our model and the exemplar model described earlier is the time-course dimension. In the exemplar model discussed, the assignment of a value to a parameter does not have any time-course. The process is instantaneous. The same is true for the relation between our model and those of Saltzman & Munhall (1989), Browman & Goldstein (1990).

Using fields is a generalization of a similar idea put forth in Byrd & Saltzman (2003), where gestural parameters are stored as ranges or windows of possible values. In our model, each window is approximated by an activation field in memory. Finally, representing targets by activation fields is also a generalization of the notion of target in Guenther's model of speech production (Guenther, 1995). In that model, speech targets take the form of convex regions over orosensory dimensions. Unlike other properties of targets in Guenther's model, the convexity property does not fall out from the learning dynamics of the model. Rather, it is an enforced assumption. No such assumption about the nature of the distributions underlying target specification need be made in our model.

3.2.1 Lenition in the Dynamical Model

When a lexical item is a token of exchange in a communicative context, phonetic details of the item's produced instance may be picked up by perception. This will have some impact on the stored instance of the lexical item. Over longer time spans, as such effects accumulate, they trace out a path of a diachronic change. Our model provides a formal basis for capturing change at both the synchronic and the diachronic dimensions.

We focus here on how a single field in a lexical entry is affected in a production-perception loop. The crucial term in the field equation is the input term $s(x,t)$. This input term $input(x,t)$ represents sensory input. More specifically, input is a peak of activation registered by the speech perception module. This peak is located at some detected x-axis value along the field. This value is assumed to be sub-phonemic in character. For example, we assume that speakers can perceive gradient differences in Voice Onset Time values, constriction location, and constriction degree within the same phonemic categories. In the current model, the input term is formulated as $e^{-(x-\mathit{off})^2}$, where *off* is the detected value or offset along the x-axis of the field.

The spike corresponding to the input term $s(x,t)$ is directly added to the appropriate field, resulting in increased activation at some point along the field's x-axis. A concrete example is presented in Figure 18. Once input is presented, a system can evolve to a stable attractor state, that is, a localized peak at a value corresponding to the input. The state is stable in the sense that it can persist even after the input has been removed. In effect, the field for the lexical item has retained a memory of the sub-phonemic detail in the recently perceived input.

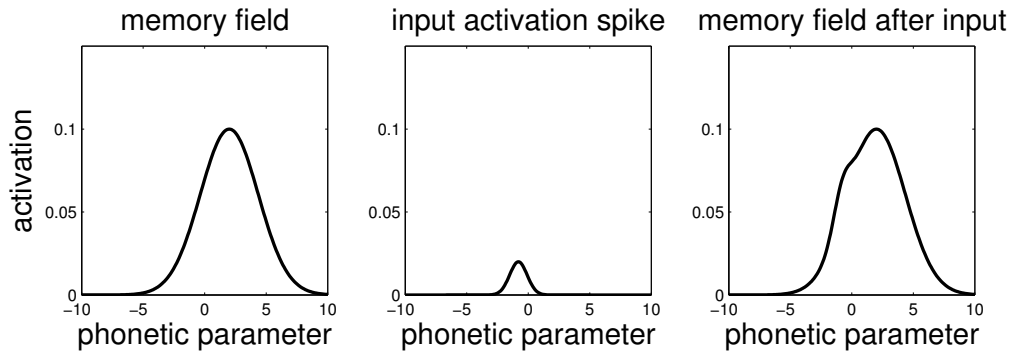

Figure 18: (Left) Field of lexical item in memory. (Middle) Input function (output of perception corresponding to $s(x,t)$ in Equation 5). Represents a localized spike in activation along the field, corresponding in location to, for example, the constriction degree of the input. (Right) Field of lexical item in memory after input is added to it. Field shows increased activation around area of input.

Since activation fades slowly over time, only areas of the field that receive reinforcement are likely to remain activated. Thus, a peak in activation may shift over time depending on which region of the field is reinforced by input. In terms of the lenition model this means that regions of the field representing a less lenited parameter fade while regions representing a more lenited parameter are kept activated by reinforcement from input.

The interaction between localized increase in activation based on input and the slow fading of the field due to memory decay is the basic mechanism for gradual phonetic change. Since activation fades slowly over time, only areas of the field that receive reinforcement are likely to remain activated. So, a peak in activation may shift over time depending on which region of the field is reinforced by input. Regions of the field representing a less lenited parameter fade while regions representing a more lenited parameter are kept activated by reinforcement from input.

Given an initial field (a preshape) representing the current memory state of a lexical item, we can simulate lenition using the model described above. Figure 19 shows the results of one set of simulations. Shown are the state of the simulation at the starting state, after 50 samples of a token, and after 100 samples (in the simulations, the number of samples is small but each sample produces a large effect on the field). Each time step of the simulation corresponds to a production/perception loop. Production was performed as described above by picking a value from the field and adding noise and a bias to it. This produced value, encoded by an activation spike of the form $e^{-(x-off)^2}$, where $off = sample(p) + noise + bias$, was fed back into the system as input.

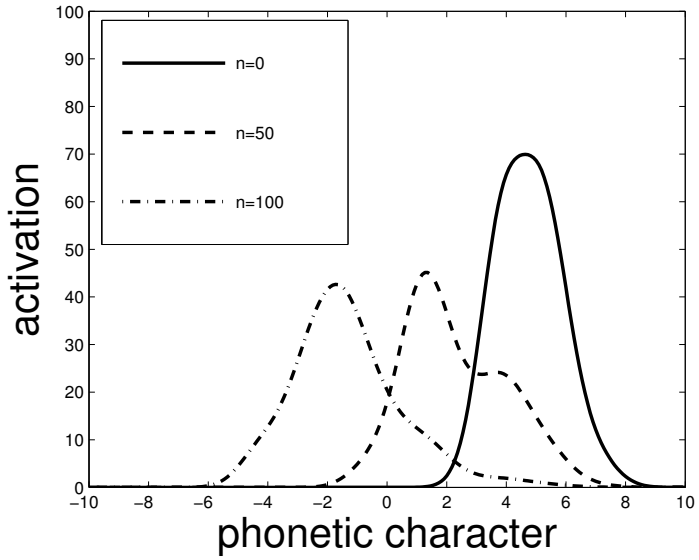

Figure 19: Output of lenition simulation. The x-axis represents a phonetic dimension (e.g. constriction degree during t/d production). Each curve represents a distribution of a particular category over the x-axis at a point in time. As time progresses, the distribution shifts to the left (i.e. there is more undershoot/lenition) and becomes broader.

As can be seen in Figure 19, at the point when lenition begins, the field represents a narrow distribution of activation and there is little variability when sampling the field during production. As lenition progresses, the distribution of activation shifts to the left. During this time the distribution becomes asymmetrical, with a tail on the right corresponding to residual traces of old values for the parameter. It also grows wider, corresponding to an increase in parameter variation while the change occurs.

With small changes in parameterization, our model can more closely represent the entrenchment behavior seen in Pierehumbert (2001). In Figure 20, lowering the strength of memory decay by adding a constant $\epsilon < 1$ factor in the $-p(x,t)$ term in Equation 5, results in less flattening of the parameter field as lenition proceeds. However, the distribution retains a wide tail of residual activation around its base.

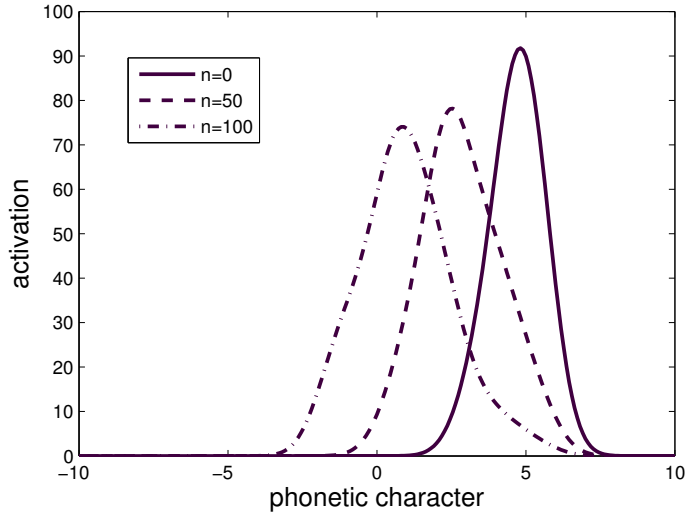

Figure 20: Lowering memory decay results in less flattening of the field as the lenition simulation proceeds.

To keep the field narrow as time proceeds, we can alternate between production/perception cycles with a production bias and without. In effect, this Figure 21 was created by biasing only every other simulated production. This was done in addition to lowering the strength of memory decay as discussed above.

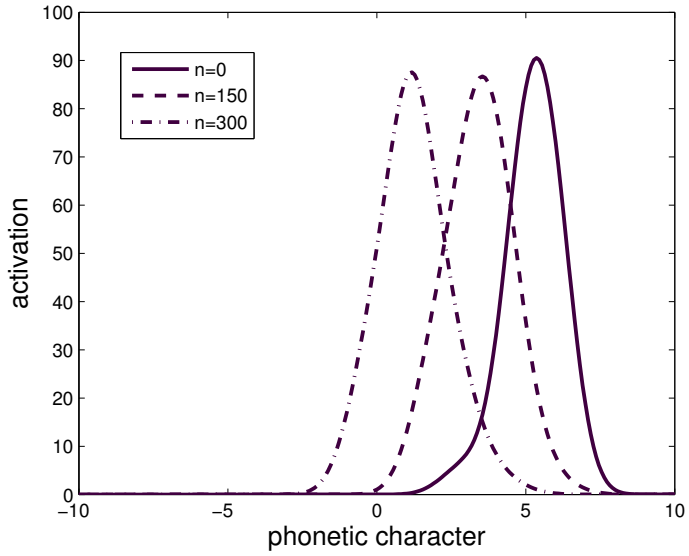

Figure 21: Interleaving biased and non-biased productions leads to consistently narrow field.

Like the exemplar model above, the model described in this section can derive the basic properties of lenition. This is true mainly because both models share a set of assumptions about how the global properties if lenition arise from local production/perception loops. The dynamical model differs from the exemplar model in terms of the representations involved and how they are manipulated. Variability in production is accounted for by noise during the gestural specification process. Each lexical item has its own fields and each production/perception loop causes a shift in the appropriate field towards lenition due to biases in production (see Figure 18 for an example of a field starting to skew to the left). In a given period of time, the number of production/perception loops an item goes through is proportional to its frequency. Thus, the amount of lenition associated with a given item shows gradient variation according to the item's frequency. All the processes described here occur within a single individual, so lenition is clearly observable within the speech of individuals. Diachronically, lenition will proceed at a faster rate for more frequent items, again because they go through more production/perception loops in a given time frame. This same mechanism is evident synchronically as well, since at any single point in time, more frequent items will be more lenited than less frequent items.

In sum, the broad proposal of this section is that diachronic change can be seen as the evolution of lexical representations at slow time scales. The specific focus has been to demonstrate that certain lenition effects, described in a previous exemplar model, can also be captured in our model of evolving activation fields.

4 Conclusion

We have presented a dynamical model of speech planning. The model consists of several formally identical, simultaneously evolving dynamical systems coupled in a hierarchical fashion. The model has been shown to fit psycholinguistic results regarding the effects of lexical factors on speech production. In addition, the model makes testable predictions about the time-course of the planning process, namely, that it is nonlinear in nature. There are rapid jumps in accuracy, rather than a gradual improvement.

We extended the dynamics used in the planning model to the featural or vocal tract variable level. This allowed us to provide an alternative account for lenition in lieu of an exemplar-based model. The dynamical and exemplar-based models cover the same ground as far as their broad agreement with the assumed properties of an evolving lenition process are concerned. However, there are fundamental high level differences between the two. Tables 2 and 3 contrast properties of the exemplar and dynamical models.

Exemplar Model

- Every token of a category (where category could mean any item capable of being recognized - word, phoneme, animal cry, etc.) is explicitly stored as an exemplar in memory. A new experience never alters an old exemplar (Hintzman, 1986).
- Complete set of exemplars forms an extensional definition of a probability distribution capturing variability of a category.
- Distributions are altered by storing more exemplars.

Table 2: Properties of the Exemplar Model

Dynamical Model

- Every token of a category is used to dynamically alter a single representation in memory associated with that category, and is then discarded. No exemplars are stored.
- Variability is directly encoded by the singular representation of a category. The parameters of a category exist as field approximations to probability distributions which are defined intensionally. That is, they are represented by functions, rather than a set of exemplars.
- Distributions are altered by dynamical rules defining the impact of a token on a distribution, and changes to the distribution related to the passage of time.

Table 3: Properties of the Dynamical Model

Two key differences are highlighted. First, the dynamical model remains consistent with one key aspect of generative theories of representation. Instead of representing categories extensionally as arbitrarily large exemplar sets, linguistic units and their parameters can have singular representations.[1] These are the fields in our specific proposal. It is these unitary representations, rather than a token by token expansion of the exemplar sets, that drifts in sound change.

Second, the dynamical model is inherently temporal. Since both the exemplar and the dynamical model are at least programmatically designed to include production and perception, which unfold in time, this seems to be a key property.

Future extensions of the dynamical model will focus on linking perceptual to motor representations and on providing a formally explicit account of other well-known lexical effects such as those involving neighborhood frequency and onset density. Moreover, the proposed model can be used as framework for testing different linguistic theories of representation and seeing how they play

[1]It is useful to distinguish the exemplar approach from a version of the dynamical one where multiple different instances of a category are stored, corresponding to different registers, different speakers, etc. For our purposes, each of these subcategories is considered unique and has a singular representation.

out within a psycholinguistic context. In particular, the simple featural phonetic specification and similarity measure used in the current implementation of the model can be swapped out for another theoretical alternative, such as specification and similarity computation based on natural classes or directly on articulatory gestures (Frisch *et al.*, 2004; Browman & Goldstein, 1986). The ability of the model to fit experimental data using these alternate representational schemes can then be evaluated. In effect, the model can be seen as a tool for strengthening the link between the substantial literature on the time-course of speech planning and linguistic theories of representation.

References

Bailey, Todd M., & Ulrike, Hahn. 2005. Phoneme Similarity and Confusability. *Journal of Memory and Language*, **22**, 339–362.

Beckman, Mary E., de Jong, Ken, Jun, Sun-Ah, & Lee, Sook-hyang. 1992. The Interaction of Coarticulation and Prosody in Sound Change. *Language and Speech*, **35**, 45–58.

Browman, Catherine P., & Goldstein, Louis. 1986. Towards and Articulatory Phonology. *Phonology Yearbook*, **3**, 219–252.

Browman, Catherine P., & Goldstein, Louis. 1990. Gestural Specification Using Dynamically Defined Articulatory Structures. *Journal of Phonetics*, **18**, 299–320.

Bybee, J. 2003. Lexical Diffusion in Regular Sound Change. *Pages 58–74 of:* Restle, D., & Zaefferer, D. (eds), *Sounds and Systems: Studies in Structure and Change*. Mouton de Gruyter, Berlin.

Byrd, D., & Saltzman, E. 2003. The Elastic Phrase: Modeling the Dynamics of Boundary-Adjacent Lengthening. *Journal of Phonetics*, **31**(2), 149–180.

Cser, András. 2003. *The Typology and Modelling of Obstruent Lenition and Fortition Processes*. Akadémiai Kiadó.

Dell, Gary S. 2000. Lexical Representation, Counting, and Connectionism. *Pages 335–XXX of:* Broe, Michael B., & Pierrehumbert, Janet (eds), *Papers in Laboratory Phonology V*. Cambridge University Press, Cambridge UK.

Dell, Gary S., & Gordon, Jean K. 2003. Neighbors in the Lexicon: Friends or Foes? *In:* Schiller, Niels O., & Meyer, Antje S. (eds), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Mouton, New York.

Erlhagen, W., & Schöner, G. 2002. Dynamic Field Theory of Movement Preparation. *Psychological Review*, **109**, 545–572.

Frisch, Stefan A., Pierrehumbert, Janet B., & Broe, Michael B. 2004. Similarity Avoidance and the OCP. *Natural Language and Linguistic Theory*, **22**(1), 179–228.

Guenther, Frank H. 1995. Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production. *Psychological Review*, **102**, 594–621.

Hintzman, Douglas H. 1986. "Schema" Abstraction in a Multiple-Trace Memory Memory Model. *Psychological Review*, **93**(4), 411–428.

Kiparsky, Paul. 1968. Linguistic Universals and Linguistic Change. *Pages 170–202 of:* Emmon, Bach, & Robert, Harms (eds), *Universals in Linguistic Theory*. New York: Rinehart and Winston.

Lund, K., & Burgess, C. 1996. Producing High-dimensional Semantic Spaces FromLexical Co-occurrence. *Behavior Research Methods, Instruments & Computers*, **28**, 203–208.

Pierrehumbert, J. 2001. Exemplar Dynamics: Word Frequency, Lenition, and Contrast. *Pages 137–157 of:* Bybee, J., & Hopper, P. (eds), *Frequency Effects and the Emergence of Linguistic Structure*. John Benjamins, Amsterdam.

Saltzman, E.L., & Munhall, K.G. 1989. A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, **1**, 333–382.

Vitevitch, Michael S. 2002a. Influence of Onset Density on Spoken-Word Recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **28**(2), 270–278.

Vitevitch, Michael S. 2002b. The Influence of Phonological Similarity Neighborhoods on Speech Production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**(4), 735–747.

Vitevitch, Michael S., & Sommers, Mitchell S. 2003. The Facilitative Influence of Phonological Similarity and Neighborhood Frequency in Speech Production in Younger and Older Adults. *Memory and Cognition*, **31**, 491–504.

Wright, Richard. 1994. Coda lenition in American English consonants: An EPG study. *The Journal of the Acoustical Society of America*, **95**, 2819–2836.

A Simulated Words

High Frequency				Low Frequency			
Dense Neighborhood		Sparse Neighborhood		Dense Neighborhood		Sparse Neighborhood	
bail	buck	balm	bob	chore	chap	beige	gash
bill	cache	calf	fig	comb	chop	char	jab
code	chip	chute	gap	dune	knack	cuff	lull
core	dam	kiss	gum	kin	lag	gauze	muff
debt	deal	myth	hub	knoll	lash	hedge	nudge
dot	dome	pool	joke	reel	loom	jade	pub
fate	dull	ridge	theme	rut	mug	jot	rib
gait	hull	shame	tub	soar	poke	shun	shag
mare	lap	shed	vote	tack	putt	soot	thug

Table 4: Simulated Words