

**Measuring the flexibility of localist connectionist models of speech perception**

Mark A. Pitt, Jay I. Myung, Maximiliano Montenegro, and James Pooley

Ohio State University

running head: Flexibility

word count: 7121

Correspondence information:

Mark A. Pitt  
Department of Psychology  
1835 Neil Avenue  
Ohio State University  
Columbus, OH 43210-1222  
office (614) 292- 4193  
fax (614) 688-3984  
pitt.2@osu.edu  
<http://lpl.psy.ohio-state.edu>

## Abstract

This study examines how network connectivity affects the flexibility of localist models of speech perception. TRACE (McClelland & Elman, 1986) has serially ordered layers with bidirectional excitatory connections between adjacent layers. ARTphone (Grossberg, Boardman, & Cohen, 1997) differs by having the initial input layer connect directly to all subsequent layers (e.g., biphone, word), with selective inhibitory links between layers. Flexibility was evaluated using parameter space partitioning (Pitt, Kim, Navarro, & Myung, 2006), which measures the number and representativeness of the data patterns a model can generate in an experimental design, in this case that of Vitevitch and Luce (1998). Results showed that network differences are greatest when the lexicon is small (<17 words). With larger lexicons (901 words), the models perform much more similarly, in part because the lexicon itself heavily constrains model behavior.

## Issues in Model Comparison and Evaluation

A goal of modeling in cognitive science is to infer the properties of cognition from the regularities present in experimental data. Computational models are a means by which we evaluate the accuracy of these inferences because they specify the mathematical form of the regularity. For example, we might entertain the hypothesis that memory retention curves decay according to a power function ( $y = ax^{-b}$ ) or an exponential function ( $y = a \exp(-bx)$ ). The purpose of such comparisons is to identify the most accurate model.

Selection of the best model is an inductive inference problem in which we generalize from specific instances (i.e., data) to all instances (i.e., model), largely by evaluating how well models can fit empirical data. This problem is ill-posed, however, because information in finite data samples (e.g., one or two experiments) is rarely sufficient to identify uniquely the true function (model) that generated the data. Random noise inherent in behavioral data further obscures the inferential process. Given these obstacles, an achievable goal of modeling is to select the model, among a set of candidate models, that best approximates the cognitive process, in some defined sense (Smelser & Baltes, 2001)

A key to identifying the best approximating model is understanding model flexibility. Model flexibility, or equivalently model complexity, refers to the ability of a model to fit a diverse range of data patterns. A model with many parameters is more flexible than a model with few parameters. Another (often neglected) dimension of model flexibility is the model's functional form, that is, the way the model's parameters are combined to yield the model equation (Myung & Pitt, 1997). For example, the power and exponential models of retention described above have the same number of parameters (2) but they differ in functional form.

Model flexibility is a double-edged sword. We want a model to be flexible enough to capture the underlying cognitive process, which can be complicated, but at the same time, the model should not be so flexible that it can fit idiosyncratic, random noise in the data. The reason for this is that by virtue of its flexibility alone, a highly flexible model can fit a data set better than a less flexible model, even if the simpler model generated the data (Myung & Pitt, 1997; Pitt, Myung & Zhang, 2002). An implication of this fact is that choosing among a set of models based solely on a goodness of fit criterion, such as mean squared error, percent variance accounted for, or maximum likelihood (Myung, 2003), can result in selecting an overly complex model that generalizes poorly, being unable to predict with reasonable accuracy data samples collected in the future.

The relationship among flexibility, goodness of fit and generalizability is illustrated in the top panel of Figure 1. Note that goodness of fit can always be improved by increasing model flexibility (point B versus point A), but at some point it will be at the cost of generalizability. It is at this point (C) along the flexibility axis that overfitting begins to take its toll. These trade-offs are made more tangible in the break-out boxes below the graph, which depict increasingly complex models (lines) fitting the same data set. Of the three models shown, the middle one generalizes best; it captures the trend in the data (which the one on the left fails to do) without fitting noise (which the one on the right does perfectly).

Various measures of generalizability have been proposed in statistics. They include the Akaike Information Criterion (AIC: Akaike, 1973), the Bayesian Information Criterion (BIC: Schwartz, 1978), the Bayes factor (BF: Kass & Raftery, 1995) and minimum description length (MDL: Grünwald, Myung & Pitt, 2005; Grünwald, 2007). In each of these model selection

methods, generalizability is measured by trading off goodness of fit for model flexibility. For reviews and application examples of these and other model selection methods, the reader is directed to two special issues of the *Journal of Mathematical Psychology* (Myung, Forster & Browne, 2000; Wagenmakers & Waldorp, 2006).

Model selection methods are designed to identify among a set of competing models the one that best approximates the underlying cognitive process, in that the model chosen accurately predicts future observations from the “true” model. As such, model selection methods are valuable statistical tools in the cognitive modeling enterprise.

### **Parameter Space Partitioning**

One limitation of *statistical* model selection methods is that they summarize the potentially intricate relationship between model and data into a single real number, a generalizability measure. Often times we want to know more than this. What other data patterns can a model produce at different parameter settings, other than the empirical pattern, which the model may have been originally designed to fit? Does the model produce non-human-like patterns as well as human-like ones? How representative is the empirical pattern of the model’s behavior? Although these are still very much questions about flexibility, they go beyond the realm of statistical model selection, and answering them requires developing other methods.

Pitt, Kim, Navarro and Myung (2006) introduced a model analysis method with these questions in mind. Dubbed *parameter space partitioning* (PSP), it explores fully the parameter space of a model by partitioning it into disjoint regions, each of which corresponds to a qualitatively different data pattern (appropriately defined). PSP enables one to find all the data patterns that a model can produce in an experimental setting by varying the full range of its

parameter values. To illustrate how PSP works, consider an experiment in which mean response times are measured and compared across three conditions, A, B, and C. Suppose that we are interested in evaluating orderings (including inequalities) among the three means, such as  $A > B > C$ ,  $A > B = C$ , and  $B = C > A$ , and further, that out of a total of 13 such qualitative data patterns, the pattern  $B > C > A$  was observed in the experiment. Now consider two models,  $M_1$  and  $M_2$ , each with two parameters. The bottom panel of Figure 1 shows hypothetical PSP results for the two models. Model  $M_1$  simulates three of the 13 possible patterns. One of the three is the empirical pattern, taking up most of the parameter space, and the other two patterns are similar to the empirical pattern. The PSP analysis shows that model  $M_1$  is not only well constrained but also it predicts the empirical pattern as its central feature. Model  $M_2$  produces nine patterns, one of which is also the empirical pattern, but given the model's ability to simulate almost any data pattern, its account of human performance is less impressive. This impression is reinforced by the fact that the empirical pattern occupies a small region of the parameter space. In this paper, we applied PSP to examine the consequences of design differences in connectionist models of speech perception.

### **Connectionist Models of Speech Perception**

Connectionist models are strong contenders in many areas of cognitive science. In some of these fields (e.g., language, memory), multiple network architectures have been proposed to account for a set of empirical findings. For example, in the field of spoken word recognition, there are TRACE (McClelland & Elman, 1986), Merge (Norris, McQueen, & Cutler, 2000), and ARTphone (Grosberg, Boardman, & Cohen, 1997). The task of deciding among them is challenging because the behavior of even simple networks can be complex.

A primary criterion for the evaluation of connectionist models is simulation performance: Are simulation data qualitatively similar to empirical data collected in an experiment? Although simulation accuracy is a necessary requirement of any model, as discussed above, much more can be learned about the model's adequacy and the meaningfulness of its performance by examining the wider performance of the model in an experimental setting using PSP.

The specific question addressed in the present study was whether multi-layer, hierarchical networks, like TRACE, are less flexible than those with a less constrained architecture, like ARTphone. This possibility was prompted by empirical evidence (Vitevitch & Luce, 1999) and by a recent PSP analysis of TRACE and Merge (Pitt et al, 2006). Before discussing this work, the two models are introduced.

A diagram of ARTphone is on the left side of Figure 2. Memory is thought to contain representations of words and smaller sub-word units, such as biphones and phonemes. Collectively, all utterances are referred to as list chunks. List chunks of the same size can inhibit each other, and larger chunks can inhibit smaller ones, which is referred to as masking. Speech is initially encoded as phonemes at the phoneme input layer, which is connected via bidirectional excitatory links to list chunks of all sizes. As speech is fed to the model, phonemes at the input layer establish a mutual excitatory feedback loop with all list chunks to which they match. This interaction is termed *resonance*, and waxes and wanes depending on the degree of match. The chunk to which the strongest resonance is established is deemed the recognized utterance.

A schematic diagram of TRACE is shown on the right side of Figure 2. It is an interactive activation model with three layers: feature, phoneme, and word. There are excitatory connections between nodes in adjacent layers, and inhibitory connections between nodes within

the same layer. Input is encoded in the model as features over time slices. Phonemes, and then words, are activated according to their degree of match to the input. Two salient differences between the models are the connectivity and sublexical representations. Connectivity is limited to immediately adjacent layers in TRACE but not ARTphone, in which the phoneme layer connects to list chunks of all sizes. Also, sublexical chunks of various sizes coexist in ARTphone, whereas there are only phonemes in TRACE.

Vitevitch and Luce (1999) were attracted to ARTphone because its connectivity seemed to provide the flexibility necessary to account for a challenging result they reported (Vitevitch & Luce, 1998). In that study, they explored the impact of phonotactic probability (frequency of phoneme co-occurrence) on word recognition. They orthogonally manipulated phonotactic probability (low vs high) and lexical status of the utterance (word vs. nonword). Listeners had to repeat items, presented over headphones, into a microphone as quickly as possible. When the stimuli were nonwords, listeners were fastest naming the strings high in phonotactic probability. The opposite result was found with words, where listeners were fastest naming the low-probability items. The result with words was argued to be lexical in origin and due to inhibition from dense neighborhoods of words: High-probability words have many neighbors that impede recognition, whereas low-probability words have few neighbors, resulting in much less inhibition. Because nonwords do not have lexical representations, the result obtained with nonwords was thought to occur sublexically. Free from the inhibitory effects of lexical neighbors, facilitory effects of phonotactic probability could emerge, thereby causing high-probability nonwords to yield faster responses than low-probability nonwords.

Pitt, Myung, and Altieri (2007) evaluated the accuracy of this theoretical account by

implementing ARTphone and performing PSP analyses across multiple testing conditions. As Vitevitch and Luce predicted, neighborhood density was the dominant force in producing the lexical effect, and substantial variation phonotactic probability was necessary to produce the sublexical effect.

Vitevitch and Luce (1999) were concerned that TRACE's connectivity was too constraining to produce the reversal in naming speed across the two probability conditions when the stimuli changed from words to nonwords. Bidirectional excitatory connections between lexical and sublexical (e.g., phoneme) layers creates a processing dependency that would seem to restrict model performance, making it especially difficult for the model to generate one data pattern lexically with words (high density words < low density words) and the opposite data pattern sublexically with nonwords (high probability nonwords > low probability nonwords).<sup>1</sup> Rather, connectivity between adjacent layers would seem to reinforce, and possibly amplify, at an earlier layer (phoneme) what occurs at a later layer (lexical). The reversal in naming speed requires a dissociation of processing between layers. In contrast, the greater independence of sublexical and lexical nodes in ARTphone would seem to provide the necessary flexibility. The PSP analyses of Pitt et al (2007) not only confirmed this intuition, but also showed ARTphone to be very flexible with particular parameter settings.

Preliminary evidence to suggest that TRACE is less flexible than models with more independent layers was found by Pitt et al (2006), who performed PSP analyses on Merge (Norris, et al, 2000) and a bare-bones version of TRACE, created by rewiring Merge. Analyses of model performance in two experimental settings showed that Merge generated more data patterns than TRACE, suggesting that the bidirectional connectivity of TRACE does in fact

constrain model performance.

The purpose of the present investigation was to explore more thoroughly the relationship between model design and model flexibility. PSP analyses were performed on ARTphone and TRACE (full-scale version) simulations in the context of the Vitevitch and Luce (1998) experimental design. Comparisons across models should answer the questions of whether TRACE is indeed more constrained in its behavior, and whether this reduced flexibility prevents it from producing the reversal in naming speed, as Vitevitch and Luce wondered.

Because the reversal in naming speed is partly attributed to differences in the size of lexical neighborhoods, PSP analyses of the models were compared across lexicons of three sizes: 4, 16, and 901 words. The size and composition of the neighborhoods differs in each lexicon. A models' sensitivity and adaptability to these differences provides another means of assessing flexibility. A more constrained model should lack the ability to produce the reversal in naming speed across lexicons because it cannot adapt to the impact of the changing contents of the lexicon. In addition, with the 901-word lexicon, we can learn how the models perform with a lexicon whose size and complexity begins to approximate that of adults.

### **Simulation Details**

#### **Lexicons**

The three lexicons were created so that the smaller ones were embedded in the larger ones. Our starting point was the 901-word lexicon, Biglex, which was chosen because it is the largest lexicon that jTRACE (Strauss, Harris & Magnuson, 2007), a user-friendly version of TRACE, can use. From analyses of its neighborhood and biphone characteristics, we selected one word from a dense neighborhood (*rub*). We replaced another word in the lexicon with the

string *shuuk* (rhymes with *shoot*) to create a word from a sparse neighborhood, and one that overlaps minimally with the *rub* neighborhood. The 4-word lexicon was created by combining *rub* and *shuut* with two neighbors of *rub* (*pub*, *cub*). Twelve additional neighbors of *rub* were added to create the 16-word lexicon. The 16-word lexicon was created to exaggerate the difference in neighborhood size between *rub* and *shuut* found in the 4-word lexicon.

Implementations of ARTphone and TRACE are described in detail for the 4-word lexicon. With the larger lexicons, the only difference in implementation is the number of words.

### **ARTphone**

The version of ARTphone is similar to that described in Pitt et al (2007), which is an expanded version of the model described in Grossberg et al (1997). Like in Figure 2, there are four word nodes, which are interconnected with inhibitory links to represent two levels of lexical density, low (zero neighbors; the utterance *shuuk*) and high (two neighbors, the words *rub*, *pub*, *cub*). At the sublexical layer, there are six biphones that make up the words. The total number of inhibitory and masking links impinging on each node is indicated by the numeral in each node. Chunks of the same size inhibit each other if they overlap. For words, overlap was defined in terms of biphones. For biphones, it was defined in terms of phonemes. Words also mask their corresponding biphones.

Variation in phonotactic probability across stimuli is assumed to be critical for generating the reversal in naming speed. Phonotactic probabilities were encoded in the bottom-up activation functions for the biphones and words. The functions were multiplied by an amount that corresponds to the list chunk's probability in a corpus of English. For biphones, the Phonotactic Probability Calculator was used (Vitevitch & Luce, 2004). For words, phonotactic probabilities

were computed by averaging the two biphone probabilities that make up a word (Vitevitch & Luce, 1999). These values were rescaled (increasing or decreasing their range of variation) using a logarithmic function when integrated into the model so that differences in phonotactic probability influenced model performance, but did not cause erratic behavior. To perform this rescaling, two additional parameters were introduced into the model, one to rescale the probabilities for biphones (*a*) and another to rescale the probabilities for words (*b*; see Pitt et al for additional details). Optimal values of these parameters were identified by searching the parameter space of the model. This was done separately for each lexicon.<sup>2</sup>

There were two word inputs to the model, *rub* and *shuuk*. The three phonemes of each word were fed to the network one at a time, each over three time units. Coarticulation was simulated by overlapping the last time unit of one phoneme with the first time-unit of the next. The resonance established by these inputs at the lexical level was used as a measure of word processing, and the resonance established by the second biphone of each word (*ub* and *uuk*) was used as a measure of sublexical processing. This simulation differs from that of Pitt et al (2007), who used as stimuli two nonwords in addition to two words. The current design was adapted from Vitevitch (2003), which is an improvement over Vitevitch and Luce (1998), in that it has the attractive feature of showing that words alone can generate the reversal; the same word can have one effect lexically and an opposite effect sublexically.

As noted above, PSP analyses involve repeating model simulations across the ranges of a model's parameters and then studying the data patterns that were produced. Which parameters should be varied? For ARTphone and TRACE, we varied those parameters that affected interactivity between layers, because these are likely to be the most central to producing the

reversal in naming speed. For ARTphone, three parameters were varied, inhibition, masking and kappa, the excitation parameter responsible for resonance. For inhibition, the parameter range was 0-.15. For masking it was 0-.3, and for kappa it was 0-8.

PSP requires that model performance be defined quantitatively. Peak resonance served as the measure of strength of evidence in favor of a biphone or word. This measure correlates negatively with reaction time, the measure of human performance in the Vitevitch and Luce experiment, where faster naming (smaller values) is indicative of more efficient processing. Therefore, to translate the empirical predictions into resonances, they must be reversed, with faster naming corresponding to greater resonance and slower naming to weaker resonance.

In the 2x2 experiment of Vitevitch and Luce, there are nine possible data patterns if one includes ties. These are shown in Table 1, with the sublexical outcomes defined across columns and the lexical outcomes over rows. The empirical pattern is number 3. Simulation data patterns were categorized as one of these nine by comparing peak resonances of *rub* with *shuuk* and *ub* with *uuk*. To qualify as one of the patterns, all resonances had to exceed a minimum value of .2 for words and .1 for biphones. Differences less than .02 were classified as equal. Patterns that failed to meet these criteria were considered invalid and together are designated as pattern 10. ARTphone and the PSP analyses were implemented in Matlab.

## **TRACE**

The simulations and PSP analyses had to be modified from those performed on ARTphone to accommodate differences between the models. As shown in Figure 2, biphones do not exist in TRACE. The only intermediate representation between the feature layer and the lexicon is the phoneme. Because of this, only the final phoneme (*b* or *k*) was used to measure

sublexical activation. In addition, the lack of biphones made it impossible to encode biphone probabilities in TRACE. As a result, the two extra parameters,  $a$  and  $b$ , that were used to encode phonotactic probabilities in ARTphone, were not added to TRACE. The inability to encode phonotactic probabilities makes the PSP analyses particularly interesting, because they provide the opportunity to learn whether the model can compensate in some way and thereby produce the empirical pattern (3).

Network dynamics in TRACE are different from ARTphone, and this necessitated using a different quantitative definition of the Vitevitch and Luce (1998) data pattern. The effects of lexical neighborhood and other variables do not always reveal themselves as differences in peak excitation. The reason for this is that long after speech input has been fed to TRACE, activation continues to circulate through the model, causing excitation to asymptote with little if any subsequent decay. Word and phoneme nodes can eventually reach high and comparable levels of activation when they fully match speech input, often making this measure insensitive to neighborhood and other differences.

A more appropriate measure was needed. Pilot simulations showed that lexical influences are most salient during the rise of the node activation functions, in particular from cycles 18 through 48. Word and final phoneme activation are underway by cycle 18, and after cycle 48 the activation functions are asymptoting. Differences between activation functions during this 31-cycle window were therefore used to define the data patterns.

The criteria used to define a data pattern as one of the nine possible were similar to that of ARTphone. The activation of *rub* was compared with *shuuk*, and  $b$  was compared with  $k$ . Word and phoneme activations had to exceed .15 from their resting levels to be counted as a

valid pattern. Functions whose values differed by less than .02 were considered equal for that cycle. During the 31-cycle window, the data pattern was defined as the one that dominated across the majority of these cycles. A pattern was considered invalid (10) when a function failed to exceed threshold and when oscillating functions were generated. The original version of TRACE, written in C, was called from the PSP algorithm, implemented in Matlab.

The PSP algorithm was used to map the parameter space of TRACE. As with ARTphone, parameters were varied that were considered central to affecting the lexical and sublexical processing necessary to produce the Vitevitch and Luce (1998) data pattern: Phoneme-Phoneme inhibition, Word-Word inhibition, and Word-Phoneme excitation. The three parameters were varied across their ranges (0-1).

### **Results and Discussion**

The results of the PSP analyses are organized in Figure 3 as a function of lexicon size. The bars in each graph represent the entire volume of the model's parameter space, with each slice containing the proportion of the volume occupied by the designated data pattern. Volume estimates are averages over 30 PSP runs. The patterns are stacked in ascending order. Refer to Table 1 for definitions of each pattern. Slices labeled 10 represent patterns that failed to meet the criteria to be considered a valid data pattern, usually because resonance or activation was too weak to reach threshold.

Comparison of the models when the lexicon has 4 words confirms the suspicions of Vitevitch and Luce: ARTphone is more flexible than TRACE. ARTphone can generate four of the nine data patterns while TRACE can generate only two. Interestingly, TRACE's two patterns are a subset of those produced by ARTphone, suggesting that TRACE's behavior is nested

within ARTphone's when the lexicon is very small.

Of the patterns that ARTphone produces, the empirical pattern is dominant, occupying just over half of the parameter space (.56). A graph of ARTphone producing the reversal in naming speed is shown in the upper left portion of Figure 4. That ARTphone can generate this data pattern over such a large range of the parameter space demonstrates its suitability in accounting for the finding. Pitt et al (2007) obtained comparable results, although in that study ARTphone produced more patterns (8 of the 9) and the volume of pattern 3 was smaller (.15). The difference between studies is due to the fact that three parameters were varied in the PSP analysis of the current study instead of two, which can change the volume of a region, and the parameter settings for  $a$  and  $b$  were different.

TRACE, in contrast, cannot produce pattern 3. Instead, its volume is dominated by pattern 5, occupying .96 of the parameter space. Pattern 5 is the null-effect pattern. It is obtained when the activation functions are comparable for the two words and for the two phonemes; the top right graph in Figure 4 shows the model's output. Because TRACE has been evaluated historically using its default parameter settings, we used them in the three simulations graphed in Figure 4 to maintain continuity with the literature. We also examined performance with values of these three parameters optimized to produce pattern 3. With the 4-word lexicon, the activation functions are very similar to those in Figure 4, although the phoneme functions are separated a bit more.

TRACE cannot produce the empirical pattern because of the size of the lexicon. Inhibitory effects from lexical neighborhoods are the primary means by which differences in activation are generated and cycle through the model. A neighborhood of three words is clearly

too small to generate the inhibition necessary to produce the High < Low pattern lexically, let alone generate enough top-down excitation to influence sublexical processing to yield the High > Low pattern. Only at the most extreme values of the parameter responsible for top-down influences, WP, does the sublexical effect show signs of emerging.

ARTphone succeeds in producing the reversal in naming speed where TRACE fails because differences in phonotactic probabilities across words and biphones (parameters *a* and *b*) compensate for weak lexical inhibition. These extra degrees of freedom in ARTphone are the source of its greater flexibility. When phonotactic probabilities are equated across words and biphones, ARTphone no longer produces pattern 3.

Model flexibility reverses when the neighborhood of *rub* is increased from 3 to 15 members. ARTphone again generates four patterns but TRACE now produces 7. Why does a larger lexicon change the behaviors of the model so significantly? For ARTphone, the 5-fold increase in neighborhood size causes so much masking (and to a lesser extent inhibition) at even low values of this parameters that biphone activation functions only reach threshold at their lowest values (<.01). When they do, the empirical pattern is generated over a significant region of the parameter space (.29). An example of ARTphone's performance in this region of the parameter space is shown in the middle graph on the left side of Figure 4. Compared with the top graph, note how much weaker peak activation of all functions is except that for *shuuk* which as with in 4-word lexicon, has no neighbors.

Whereas the larger neighborhood reduces the flexibility of ARTphone, it empowers TRACE with flexibility by causing *rub* to be inhibited far more than *shuuk*. The bidirectional excitatory connections between the lexical and phoneme layers causes these processing

differences to spread not only to the phoneme layer, but also back up to the lexical layer. By independently varying the three parameters that control the flow of activation through the model, TRACE can produce almost any of the patterns in the experimental design. Although this includes pattern 3 (.05 volume), it is not nearly as representative as many of the other patterns. That said, pattern 3 can be approximated with the model's default parameter settings (middle right graph in Figure 4), showing that it does not need to take advantage of this flexibility to mimic human behavior. Like with ARTphone, the High<Low pattern found lexically is more robust than the High>Low pattern found with the phonemes, which emerges only early in the evolution of activation (e.g., cycles 18-30). With optimal parameter settings, the High < Low effect for words shrinks, and the High>Low effect for phonemes increases.

With Biglex, the PSP analyses of the two models resemble each other. Pattern 10 dominates the majority of the parameter space (> .85), and the region occupied by pattern 3 is small and similar in size in both models (ARTphone=.02, TRACE=.01). Although TRACE generates two more patterns than ARTphone (6 vs. 4), the fact that the extra patterns occupy such a small amount of the total volume tempers the impact of this additional flexibility. These results shows that ARTphone's architecture is not more flexible than TRACE's, at least within the context of the Vitevitch and Luce data. TRACE generates just as many data patterns as ARTphone, including the empirical one.

Despite their structural and representational differences, the two models perform much more similarly when a more realistic lexicon is used. When the lexicon is large, it, more than differences in model architecture, seems to govern model performance. Words act en mass to constrain model behavior significantly. With so many words in the lexicon, "valid" data patterns

(i.e., 1-9) are found in both models only when parameter values are within a narrow range and together occupy less than .15 of the volume, which is far different from what was found with the smaller lexicons. In addition, the parameters for inhibition (and masking for ARTphone) must be very, very low. Otherwise activation functions will not exceed threshold.

Other ways in which the large lexicon affected model performance are visible in the bottom two graphs of Figure 4. For ARTphone, the activation functions become more peaked with a larger lexicon. In addition, the functions for *shuuk* and *uuk* are shifted later in time, probably another consequence of *shuuk* being in a sparse neighborhood. Composition of the lexicon also has temporal processing consequences in TRACE, although they are less obvious to the eye. The High < Low pattern for words emerges early (cycle 11) and the functions diverge quickly. On the other hand, the sublexical High > Low pattern does not emerge permanently until cycle 24, as though there are forces preventing the high probability phoneme, *b*, from rising faster than *k*. Even then, the functions diverge slowly from cycle 24 onward. That the time course of these two effects is due to properties of the broader lexicon can be seen by comparing these data with those in the graph above it, where the lexicon contains only two neighborhoods, one with 15 words and the other with one. Activations functions for the phonemes start to diverge much earlier, at the same time as the words. They then converge at trial 30 and remain overlapping for the remaining cycles. In contrast to the High < Low pattern for words, which is temporally similar across Biglex and the 16-word lexicon, the emergence of the High > Low pattern for phonemes is more complex. Because the 16-word lexicon is contained within Biglex, the presence of other words is at least partially responsible for this change in activation time course.

In the experimental literature, the sublexical High > Low pattern has also proven to be complex. The High > Low pattern is frequently obtained when participants respond quickly, but when responses are slow or the pace of the experiment is slow, the opposite outcome (High < Low) or a null effect is found (Lipinski & Gupta, 2005; Vitevitch & Luce, 2005). Vitevitch and Luce raise the possibility that the High > Low data pattern is a time-dependent process, occurring only early in word processing. The present simulations illustrate one context in which this can be the case, but they more strongly suggest that a broader consideration of overlap in the lexicon could assist in understanding the sublexical effect.

### **Conclusions**

The goal of the present study was to understand the consequences of design differences between two localist models of speech perception. PSP analyses of ARTphone and TRACE in the context of the Vitevitch and Luce (1998) experimental design showed that they are comparable in flexibility when the simulations include a more realistically-sized lexicon, generating a similar number of data patterns and with the empirical pattern occupying a similarly-sized region in the parameter space. When small lexicons were used, design differences were evident, one of which was ARTphone's superior ability to generate pattern 3 across lexicons, an indication of its greater flexibility. In some circumstances (16-word lexicon), it is clear that TRACE can exhibit a high degree of flexibility. The suspicions of Vitevitch and Luce (1999) were partly correct. ARTphone can be more flexible than TRACE, but TRACE itself is sufficiently flexible to generate the reversal in naming speed.

How are these two very different models able to produce the reversal? The challenge lies not so much in producing the lexical effect (High < Low) as it is in producing the sublexical

effect (High > Low). The lexical pattern comes about in both models from greater lexical inhibition in dense than in sparse neighborhoods, resulting in the low-density word achieving a higher level of activation. The origin of the sublexical pattern is quite different in the two models because of how the lexicon interacts with sublexical layers. As shown in Figure 2, in ARTphone, biphones are masked by words with which they overlap, which means that high-probability biphones will naturally receive greater masking than low-probability biphones. To offset this effect, phonotactic probability is explicitly encoded in biphone excitation level so that high-probability biphones achieve a higher level of activation. Phonotactic probability differences, motivated by the data of Vitevitch and Luce are key ingredients to producing the effect in ARTphone.

In TRACE, top-down effects are facilitatory, so dense neighborhoods of words boost the sublexical activation of phonemes in these same words, which means the phonemes of words from dense neighborhoods will achieve a higher level of activation than those from sparse neighborhoods, yielding the High>Low pattern. Phonotactic (or phoneme) probability does not need to be encoded directly in TRACE because it emerges from the combined influence of words in the lexicon. The sublexical pattern is another example of a “gang” effect in TRACE (McClelland & Elman, 1986). The same principle is present in ARTphone, but because top-down effects are inhibitory, the model’s natural behavior is to produce the High<Low pattern sublexically when parameters for phonotactic probability (*a* and *b*) are not present. This connectivity difference between the models is likely to be important in distinguishing between them in future work.

The PSP analyses across lexicons make clear the central role of the lexicon in

determining the behavior of the two models. The vastly different results across the three lexicons show that it's composition can greatly affect model behavior. Compared to small lexicons, large ones squash model flexibility by limiting the effectiveness of parameter variation in generating sensible (i.e., valid) data patterns. In this respect, the lexicon functions almost as a meta-parameter.

Model flexibility is a complex issue that modelers must confront to understand the meaning and implications of their simulations. As mentioned in the introduction, a model must have enough flexibility to generate the empirical pattern of interest. Otherwise the model is clearly inadequate. Because a successful model will eventually be applied to many testing situations, a high degree of flexibility may be necessary for the model to generate very different data patterns. There is a catch-22 here, in that the high degree of flexibility needed for the model to generalize to new testing situations can lead to the model generating most or all data patterns that might be observed in a single experiment. The broad perspective that PSP provides about model behavior allows one to identify such situations, and thus understand what it means for a model to simulate a data pattern. In this regard, it is impressive that with the larger lexicons, TRACE did a reasonably good job of simulating the empirical pattern with its default parameter settings.

Another way to think about the problem of flexibility, and the implications it has for model evaluation, is as a mismatch between a model and the ability of an experimental design to test it or to discriminate between a set of models. PSP provides a means of assessing whether a model's flexibility mismatches the power of the experimental design. With such knowledge in hand, the modeler has some idea of the quality of the data and its potential to advance the field.

PSP extracts much richer information about a model than statistical model selection methods, such as AIC and MDL. Information gleaned from applying PSP can help us deepen our understanding of how and why the model behaves the way it does, thereby improving modeling in the cognitive sciences.

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in B. N. Petrox and F. Caski, *Second International Symposium on Information Theory* (pp. 267-281). Akademiai Kiado, Budapest.
- Grossberg, S., Boardman, I., Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 483-503.
- Grünwald, P. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press.
- Grünwald, P., Myung, I. J., & Pitt, M. A., eds., (2005). *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Lipinski, J. & Gupta, P. (2005). Does neighborhood density influence repetition latency for nonwords? Separating the effects of density and duration. *Journal of Memory & Language*, 52, 171-192.
- McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- Norris, D., McQueen, J.M., & Cutler, A. (2002). Perceptual learning in speech. *Cognitive Psychology*, 47, 204-238.
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation . *Journal of Mathematical Psychology* , 47, 90-100.

Myung, I. J., Forster, M., & Browne, M. W. (2000). Special issue on model selection. *Journal of Mathematical Psychology, 44*, 1-2.

Myung, I. J. & Pitt, M. A. (1987). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95.

Pitt, M.A., Kim, W., Navarro, D J., & Myung, J.I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113*, 57-83.

Pitt, M. A. & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6(10)*, 421-425.

Pitt, M.A., Myung, J. I., Altieri, N. (2007). Modeling the word recognition data of Vitevitch and Luce (1998): Is it ARTful? *Psychonomic Bulletin & Review, 14*, 442-448.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109(3)*, 472-491.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Smelser, N. J., & Baltes, P.B., eds., (2001). Computational approaches to model evaluation. *The International Encyclopedia of the Social and Behavioral Sciences* (pp. 2453-2457). Oxford, UK: Elsevier.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods, 39*, 19-30.

Vitevitch, M.S., & Luce, P.A. (1998). When words compete: Levels of processing in spoken word perception. *Psychological Science, 9*, 325-329.

Vitevitch, M.S., & Luce, P.A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language*, 40, 374-408.

Vitevitch, M.S., & Luce, P.A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory & Language*, 52, 193-204.

Wagenmakers, E.-J. & Waldorp, L. (2006). Model selection: Theoretical developments and applications. *Journal of Mathematical Psychology*, 50(2).

## Footnotes

<sup>1</sup> *Word density* is used to refer to the manipulation of word probability because the two correlate highly and *word density* is most descriptive of the source of the High < Low data pattern, neighborhood size.

<sup>2</sup> Other analyses showed that ARTphone cannot produce the empirical pattern across the three lexicons with *a* and *b*, the parameters that encode phontactic probability, held constant. Such values can be found when only the 4-word and 16-word lexicons are considered. Because PSP analyses with these settings are similar to those reported, we decided to use the optimal settings for these parameters for each lexicon. A similar comparison was performed on TRACE, and is described in the main text.

Author Notes

Mark A. Pitt, Jay I. Myung, Maximiliano Montenegro, James Pooley, Department of Psychology, Ohio State University.

This work was supported by research grant R01-MH57472 from the National Institute of Mental Health, National Institute of Health. Portions of this work were presented at the 2006 Annual Meeting of the Society for Mathematical Psychology.

Correspondence concerning this article should be addressed to Mark Pitt or Jay Myung, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH, 43210. Electronic mail should be sent to [Pitt.2@osu.edu](mailto:Pitt.2@osu.edu) or [Myung.1@osu.edu](mailto:Myung.1@osu.edu)

Table 1. The Nine possible data patterns in the 2x2 experiment of Vitevitch and Luce (1998).

		Sublexical probability relationship		
		High > Low	High = Low	High < Low
Word density relationship	High < Low	3 Empirical	2	1
	High = Low	6	5	4
	High > Low	9	8	7

## Figure Captions

Figure 1. Top panel: Relationship between flexibility, goodness of fit, and generalizability. The lower graphs depict the fits of three models to the same data set, increasing in flexibility from left to right (adapted from Pitt & Myung, 2002). Bottom panel: An illustration of the parameter space partitioning analysis for two hypothetical models in an experiment with three conditions, A, B and C (adapted from Pitt et al, 2006).

Figure 2. Schematic illustrations of ARTphone and TRACE. Only the most relevant links are shown. The number of these links impinging on each node is designated by the numeral inside the node.

Figure 3. The proportion of parameter space in ARTphone and TRACE occupied by each of the data patterns in the Vitevitch and Luce (1998) design.

Figure 4. ARTphone (left side) and TRACE (right side) simulations. Output using the 4-word lexicon is in the top graph, with the 16-word and Biglex lexicons in the lower graphs. For ARTphone, the simulation is of the empirical pattern across all three lexicons. For TRACE, the simulations are of model performance with its default parameter values.

Figure 1







