THE RESONANT DYNAMICS OF SPEECH PERCEPTION:

INTERWORD INTEGRATION AND DURATION-DEPENDENT BACKWARD EFFECTS

by

Stephen Grossberg and Christopher W. Myers

Department of Cognitive and Neural Systems
and
Center for Adaptive Systems
Boston University
677 Beacon Street
Boston, MA 02215

Suggested Running Head: Neural Dynamics of Intersyllabic Speech Integration

January, 1999
Revised October, 1999
Psychological Review, in press.

Technical Report CAS/CNS-TR-99-001
Boston, MA: Boston University.

ABSTRACT

How do listeners integrate temporally distributed phonemic information into coherent representations of syllables and words? During fluent speech perception, variations in the durations of speech sounds and silent pauses can produce different perceived groupings. For example, increasing the silence interval between the words “gray chip” may result in the percept “great chip”, whereas increasing the duration of fricative noise in “chip” may alter the percept to “great ship” (Repp *et al.*, 1978). The ARTWORD neural model quantitatively simulates such context-sensitive speech data. In ARTWORD, sequential activation and storage of phonemic items in working memory provides bottom-up input to unitized representations, or list chunks, that group together sequences of items of variable length. The list chunks compete with each other as they dynamically integrate this bottom-up information. The winning groupings feed back to provide top-down support to their phonemic items. Feedback establishes a resonance which temporarily boosts the activation levels of selected items and chunks, thereby creating an emergent conscious percept. Because the resonance evolves more slowly than working memory activation, it can be influenced by information presented after relatively long intervening silence intervals. The same phonemic input can hereby yield different groupings depending on its arrival time. Processes of resonant transfer and competitive teaming help determine which groupings win the competition. Habituating levels of neurotransmitter along the pathways that sustain the resonant feedback lead to a resonant collapse that permits the formation of subsequent resonances.

Key words : speech perception, word recognition, consciousness, adaptive resonance, context effects, consonant perception, neural network, silence duration, working memory, categorization, clustering.

1. Introduction

How do listeners integrate individual speech sounds, which arrive at the ear as distributed and overlapping acoustic patterns, into coherent percepts of words? Several decades of quantitative research in psycholinguistics (Cutler, Dahan, & van Donselaar, 1997; Lisker, 1985; Repp, 1982; Repp & Liberman, 1987), cognitive neuroscience (Margolin, 1991; Miller, Delaney, & Tallal, 1995; Rauschecker, 1998), and statistical pattern recognition (Lippmann, 1989; Jelinek, 1976, 1995; Nakatani & Hirschberg, 1994) have yielded important partial answers, but this question continues to provide fertile ground for new investigation. For example, two decades ago Repp, Liberman, Eccardt, and Pesetsky (1978) used a recording of the sentence "Did anyone see the gray ship?" to show that increasing the silence interval between the words "gray ship" can cause listeners to perceive them as "gray chip", or at longer silence intervals as "great chip". Further, increasing the duration of the initial fricative noise of the word "chip" can induce a switch in the perception of "gray chip" to "great ship", thus changing the percept of the first word by altering the beginning of the second word. The processes by which newly arriving phonemic information, such as the initial fricative noise in "chip", can modulate the online perception of earlier occurring speech such as the stop consonant /t/ in "great", even across word boundaries, remain largely unexplained.

In this paper, we develop a dynamical model of neural processes, called ARTWORD, that is capable of integrating temporally distributed phonemic items into unitized syllabic representations of phonemic item sequences, or lists. The model elucidates how information occurring *after* a given speech event can alter the dynamics of competition between previously activated unitized representations and thereby alter the percept of an *earlier* word, as in the data of Repp *et al.* (1978). In order to deal with words of variable length, the model introduces unitized list representations that can selectively respond to words of a particular length, yet also be subliminally primed by shorter words. The model posits an ongoing dynamic competition between unitized list representations biased to favor the longest word interpretation that is consistent with the available bottom-up evidence. Top-down feedback to phonemic item representations creates a slowly developing resonance between item and list levels, which is sustained by the feedback. As new phonemic information arrives, the bottom-up evidence may shift to favor a new, larger list representation as support for the currently most active, smaller representation weakens due to transmitter habituation within the active feedback pathways. This combination of dynamic events can create a *resonant transfer* from one list representation to another, during which the resonance between phonemic item and list levels is sustained, and results in a seamless integration of phonemic information into a single unitized percept. The model is used to quantitatively simulate the data of Repp *et al.* (1978). The model hereby further develops processes that have elsewhere been used to explain other speech and language data (Boardman, Grossberg, Myers, & Cohen, 1999; Cohen & Grossberg, 1986, 1987; Cohen, Grossberg, & Stork, 1988; Grossberg, 1986; Grossberg & Stone, 1986; Grossberg, Boardman, & Cohen, 1997) to explain data about interword integration. The main innovation of the ARTWORD model is to show how list chunks that represent words of variable length can be selectively activated, can compete effectively with related list chunks of different length, can deliver the correct levels of top-down feedback to their working memory items, and can then receive the correct amounts of bottom-up feedback from these items, thereby generating resonances whose properties explain challenging speech data.

2. Neural Dynamics of Phonemic Integration

The brain processes that group sounds into coherent speech units exhibit an exquisite sensitivity to the temporal distribution of spectral energy in the speech stream. For example, the speech literature has revealed a number of context effects whereby later-occurring information influences an earlier perceptual grouping decision. These so-called *backward effects* directly constrain theories of how the perceptual units of language spontaneously form under variable-rate speaking conditions. In particular, they show that the time scale of conscious speech is not equal to the time scale of bottom-up processing.

Striking examples of backwards effects come from *phonemic restoration* experiments (Bashford, Riener, & Warren, 1992; Repp, 1992; Samuel, 1987, 1991; Warren, 1970; Warren & Obusek, 1971; Warren & Sherman, 1974; Warren & Warren, 1970; Warren, Hainsworth, Brubaker, Bashford, & Healy, 1997). When a phoneme, such as /s/ in "legislature" is excised from a word and replaced by silence ("legi-lature"), subjects readily localize the silent gap. But if the silence is replaced with broadband noise, such as a cough, subjects not only fail to localize the missing phoneme, they report hearing all phonemes as present. Moreover, the context of the word and carrier sentence determines the identity of the restored phoneme. If the /s/ in "jump on the sandwagon" is spliced out and replaced by noise, subjects will report hearing "bandwagon", despite the absence of the usual acoustic cues for the voiced stop consonant /b/.

Even more striking is the fact that "the resolving context may be delayed for two or three, or even more words following the ambiguous word fragment" (Warren & Sherman, 1974, p. 156). In the phrase "[noise]eel is on the —–", where the resolving context is given by the last word ("axle", "shoe", "orange" or "table"), listeners "experience the appropriate phonemic restoration ["wheel", "heel", "peel", or "meal"], apparently by storing the incomplete information until the necessary context is supplied so that the required phoneme can be synthesized" (Warren & Warren, 1970, p. 32). Thus, despite the fact that we do not perceive "orange" as occurring before "peel", we appear to delay the formation of the "peel" percept until after the word "orange" arrives. In this example, the later occurring top-down effect of meaning influences the phonemic structure which is consciously perceived as coming earlier in time. These data illustrate that the brain mechanisms that generate speech percepts can integrate contextual information across a relatively broad temporal window and still maintain a natural ordering of the linguistically significant acoustic signals that reach our ears.

Just as the semantic context of a phrase can shape the perception of noise into a particular phonemic segment, the acoustic context of segmental durations in a syllable can shape the perception of that syllable's component phonemes. Broadly speaking, speech is characterized by four types of acoustic segments (Anderson & Port, 1994): sustained energy concentrated in narrow frequency bands called *formants*, the transitions linking formants to other acoustic segments, higher frequency spectrally shaped noise, and silent gaps associated with stop and affricate consonants. *Context effects* occur when the perception of one phoneme is altered by changing the acoustic characteristics of nearby sound segments. *Trading relations*, by contrast, occur when a phonemic percept can remain unchanged by simultaneously changing more than one acoustic features of the signal; these features are said to "trade against each other" (Repp, 1982). The data of Repp *et al.* (1978) illustrate both context effects and trading relations occurring across syllable boundaries. These effects, moreover, are distinctively "backwards", in that much later segmental features, like the duration of "sh" (/ʃ/) in "ship" can alter the perception of earlier phonemes like the "t" (/t/) in "great".

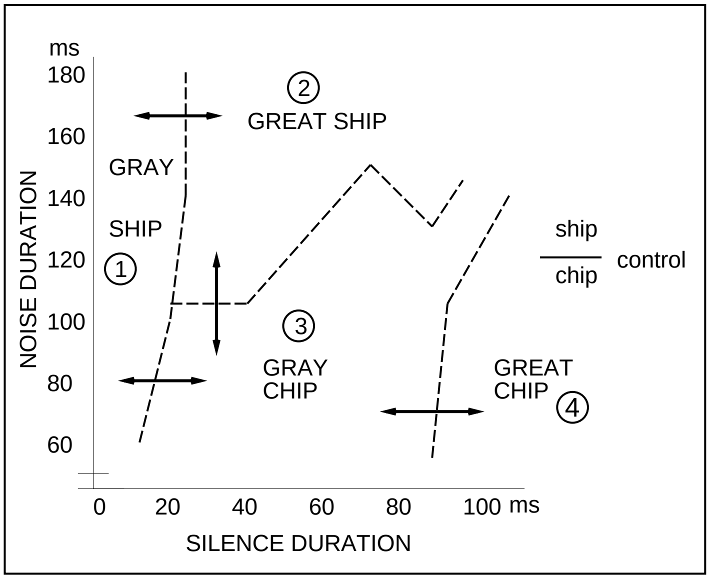


Figure 1: Perceptual boundaries derived from responses (redrawn from Repp *et al.* (1978), Figure 4, p. 630.

The main findings from the Repp *et al.* (1978) experiments are illustrated in Figure 1. This figure shows how the duration of silence between the words "gray ship" (i.e., the abscissa *silence duration*) and the duration of the fricative noise segment /ʃ/ in "ship" (i.e., the ordinate *noise duration*) jointly influence whether listeners perceive "gray ship", "gray chip", "great ship", or "great chip". The original utterance "gray ship" lies in region 1, with no silence between the "ay" and "sh", and a fricative noise of approximately 122 ms. However, when listeners were exposed to the word "gray", followed by a silent interval and then "ship", they would assimilate the silence and the noise in "sh" into cues for the presence of a stop consonant, perceiving "gray" as "great". Given a noise duration of 160 ms, the "t" sound was reliably perceived at the longest silent intervals tested, 100 ms (see regions 2 and 4 in Figure 1). Thus, the assimilation of these cues took place over a relatively long time span and grouped the "t" with the preceding word "gray" without filling the intervening silence with the later occurring "sh" sound. In this range, the perceptual representation of "great" joins the sustained formants of "ay" (/ei/) in "gray" with the later occurring cues for "t" (/t/). Moreover, it does so across the duration of silence instead of linking the "t" sound to the temporally contiguous "chip" signals.

Regions 3 and 4 in Figure 1 illustrate that the second word which listeners perceived can also depend on the silence and noise durations. Simply by shortening the duration of the fricative noise in "ship", Repp *et al.* could induce a switch in the percept from "gray ship" (region 1) or "great ship" (region 2) to "gray chip" (region 3). The transition from region 2 to region 3 is particularly interesting. For a given silence duration, shortening the noise duration caused the perceived stop consonant /t/ to leave the first syllable /grei/, and latch onto the fricative /ʃ/ to form the affricate consonant /tʃ/ ("ch"). Remarkably, without changing the amount of silence separating the words, a variation in the initial segment of the *second* word can alter perception of the *first* word. The boundary between regions 2 and 3 reveals, moreover, a trading relation between silence and noise durations. At longer silence durations, longer noise durations are required in order to cue a switch from “gray chip” to “great ship”. Finally, in region 4, a “stoplike” consonant is perceived in both words – the “t” in “great” as well as the “ch” in “ship”. The transition between regions 3 and 4 (“gray chip” to “great chip”) shows the paradoxical effect that increasing the separation of “chip” from “gray” can change the “gray” percept into “great”.

Several questions about the brain’s underlying perceptual mechanisms need to be answered to develop a unified explanation of these and related data. How and why does the brain generate its perceptual representations in such a way that coherent groupings like “gray” and “chip” can influence each other across such long time spans? How do the representations emerge in such a way that a future sound like “t” can leap over a preceding interval of silence without filling that interval with the “t” sound. Moreover, how does the brain generate these context-sensitive perceptual units without altering the order in which the groupings are perceived?

To answer these questions, Grossberg and colleagues have postulated a hierarchy of processing levels that are linked together by bi-directional pathways, as shown in Figure 2 (Cohen & Grossberg, 1986, 1987; Grossberg, 1978a, 1986). Higher levels in the hierarchy consist of neural populations responsive to successively more compressed representations of activity over the lower levels. These pathways contain adaptive synaptic weights that permit the activations of neurons within each level to differentially influence the activities of neurons in other levels. In other words, the adaptive pathways act as *adaptive filters* that enable each population to selectively respond to particular activity patterns across adjoining levels.

At the lowest levels in the hierarchy, peripheral auditory neurons send signals to higher-level neurons that encode *iconic sensory features*. A pattern of activation across these feature detectors, within a small time interval, activates a compressed *item* representation. For example, He *et al.* (1997) have recently described single-cell tuning to noise bursts of either short or long duration in cat auditory cortex. Such cells could encode, for example, the distinction between “ch”-like sounds with brief fricative bursts and “sh”-like sounds with longer duration fricative noise. In the perception of speech and language, sequences of item representations are temporarily stored in a working memory as a temporal succession of sounds occurs. The working memory transforms a sequence of sounds into an evolving spatial pattern of activation that represents the items and the temporal order in which they occurred (Bradski, Carpenter, & Grossberg, 1994; Grossberg, 1978a, 1978b). Network dynamics within the working memory can store the serial position of items in a sequence using a gradient of activity across the working memory item representations. In the present simulations, parameters were set in the working memory so that a *recency* gradient emerged; that is, the most active item representations correspond to the most recent events. As later network processes alter the activity levels in the working memory, they preserve relative activities across items, and thus serial order information. Other temporal gradients could be generated, depending on network parameters, notably primacy gradients in which the least active item activities correspond to the least recent events, or bowed gradients in which item activities are largest at the beginning and end of a list; see Bradski *et al.* (1994) for examples.

The activity patterns across the item-and-order working memories, in turn, activate *list chunks*, which are unitized, context-sensitive representations of a particular temporal sequence of items. These list chunks may represent, for example, phonemes, syllables, or words. Because each pattern across the working memory represents both items and their order of activation, the list chunks encode particular list sequences. Active list chunks feed back to the item working memories to support the neural activations there via reciprocal connections. At the same time, top-down

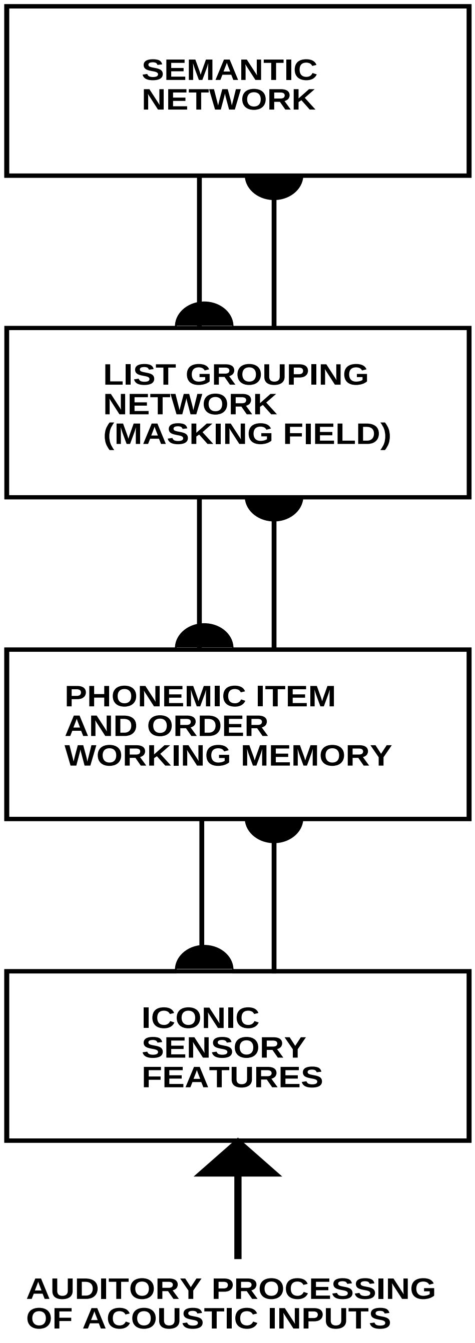

SEMANTIC NETWORK
LIST GROUPING NETWORK (MASKING FIELD)
PHONEMIC ITEM AND ORDER WORKING MEMORY
ICONIC SENSORY FEATURES
AUDITORY PROCESSING OF ACOUSTIC INPUTS

Figure 2: Macrocircuit for neural speech and language perception.

feedback suppresses items in the working memories that are not represented by the active list chunks via a nonspecific inhibitory gain control pathway. These interactions between the chunking network and the working memory — namely, non-specific top-down inhibition combined with specific top-down confirmation of expected items — can naturally begin to explain aspects of some speech perceptual phenomena. For example, in phonemic restoration experiments, broadband noise may be perceived as different phonemes depending on the context. These percepts may be attributed to a process by which active list chunks use their learned top-down expectations to select the noise components that are consistent with the expected formants and suppress those that are not (Grossberg, 1995, 1999d). Future information can influence this selection process because list chunk feedback is delayed in time relative to the bottom-up arrival of signals.

When a phonemic sequence present in the working memory excites, and receives confirmatory top-down feedback from, a list chunk or chunks, the positive feedback loop that is hereby created enhances activity in both fields through a process known as *resonance*. The model proposes that when listeners perceive fluent speech, a wave of resonant activity plays across the working memory, binding the phonemic items into larger language units and raising them into the listener's conscious perception (Grossberg, 1978a, 1986).

The specification of resonant dynamics within a speech perception neural network must solve a key problem: The multiple time scales that are used to activate and group phonemic items need to be coordinated to form a unified speech percept. In particular, the processing of acoustic information prior to its storage in the working memory unfolds on a very rapid time scale – consonants, for example, are typically uttered in tens of milliseconds. As items become rapidly activated by their partially compressed auditory codes, they are stored in a working memory that preserves them on a slower time scale, even as they activate list chunks. The chunks also become active on a slower time scale, since their bottom-up evidence is only completely available once all the items in their list have been presented. Word durations are typically hundreds of milliseconds, and many words cannot be reliably perceived until well after their acoustic offsets (Bard, Shillcock, & Altmann, 1989; Grosjean, 1985). In addition to the response times of list chunks and items in working memory, the *interactions* between the chunks and items create an emergent resonance time scale that reacts quickly enough to keep up with the incoming speech stream, but slowly enough to allow contextual information to affect it, as in phonemic restoration and Repp *et al.* (1978) data. The context-sensitive resonance time scale is proposed to be the primary coordinating factor. According to this hypothesis, speech is perceived only when both phonemic items and their chunks are co-active in a resonant loop, and hence the rate of conscious speech is equal to the time scale of the resonance between multiple processing levels. The variously timed factors that determine the rate of resonance, and hence the rate of conscious speech perception, may themselves not be available to introspection. Only together do these finely timed processes generate a wave of resonant activity corresponding to the conscious stream of speech percepts.

Under the assumption that the conscious speech code is a resonant wave, the dynamics governing the propagation of the wave also delimit the temporal window in which items, activated by bottom-up inputs, can be bound together into a larger conscious percept. A large body of data in the speech literature has examined the temporal constraints on the perception of phonemes and words in specific contexts. One major effect concerns the fusion, doubling, or breaking of a set of consonants. Repp (1980) studied the silence durations that allow different consonants in VC–CV pairs to be perceived as two consonants rather than one. In particular, he investigated when /Ib/-/ga/ and /Ib/-/ba/ are perceived as /Iga/ and /Iba/, respectively. Repp's data revealed that a silent interval approximately 150 ms longer was required to perceive two occurrences of the same consonant (e.g., the *geminate* consonant pair in /Ib/-/ba/) than to perceive two different consonants (e.g., the *cluster* consonant pair in /Ib/-/ga/). Grossberg *et al.* (1997) have modeled how the perceptual distinction between the cluster and geminate stop consonants can be explained by the dynamics of speech resonance. In brief, if the representation of /g/ becomes active while the representation of /b/is active, then /g/ begins to actively inhibit /b/ while initiating its own resonance. In contrast, if the second occurrence of /b/ arrives while the first is already resonating, then it can extend the ongoing resonance and thereby prolongs the fused percept /Iba/. The first /b/ resonance must self-terminate (by a process called habituative collapse that is later explained) before a second /b/ resonance can be initiated and perceived.

These simulations illustrated how resonance between working memory items and chunks can contextually reorganize temporally variable presentations of inputs into perceptually fused or separated percepts, depending on the phonetic context. In addition, while the Grossberg *et al.* (1997) model simulations do not incorporate learning of these interactions, the model developed therein belongs to a broader theory called Adaptive Resonance Theory, or ART, which describes how learning occurs within the pathways that mediate these interactions and thereby builds the list respresentations that are capable of temporally deforming items into larger word groupings (Carpenter & Grossberg, 1991; Grossberg, 1999b; Grossberg & Stone, 1986).

Other speech data suggest that the rate at which resonances develop is sensitive to more global aspects of the incoming speech. For example, Bashford *et al.* (1988) found speech-rate effects in the perceived continuity of fluent speech. When a spoken passage was interrupted by silence or noise, the mean duration of the interruption necessary to be detected varied with the rate at which the passage was presented. For a noise interruption, the detection threshold was very close to the average word duration in the passage. This result held for each of three speech rates tested. Thus, an estimate of the mean rate of the incoming speech appears to modulate the rate at which resonance unfolds.

These considerations converge on two prominent issues in the modeling of phonemic integration. The first issue concerns how to design the working memory so that it stores sequences of items with a representation that is (approximately) independent of speaking rate. Such a working memory representation helps to explain how variations in segmental durations corresponding to different speech rates can determine the perceptual distinction between the stop consonant /b/ and the glide /w/: If the vowel /a/ in the syllable /ba/ is shortened sufficiently, then the syllable may be perceived as /wa/, despite identical spectral energy in the initial formant frequency transitions. The particular backwards effect whereby vowel duration determines whether listeners perceive /ba/ or /wa/ is an example of *durational contrast*. Durational contrasts occur when a segment of given duration seems longer in the context of a short segment than in the context of a long segment. This perceptual effect is consistent with the existence of a rate-based scaling mechanism that maintains *relative* activation levels in the working memory over variable speech rates. Durational contrasts occur in other phonemic contexts as well, as when the perception of the affricate /tʃ/ in /tʃa/ can "switch" to the fricative /ʃ/, when the following vowel /a/ is shortened (Kluender & Walsh, 1988). These durational contrast phenomena illustrate how changing the relative duration of the working memory inputs (for example, how /b/ is processed relative to a short or long /a/) can change the hypotheses selected by the grouping network (/ba/ or /wa/).

Recently, Boardman *et al.* (1999) developed a working memory model, called PHONET, that was used to quantitatively simulate how the /ba/-/wa/ distinction depends on the subsequent vowel duration. The model begins to provide a more sensitive account of how speech preprocessing influences how working memory items are defined and interact. Such preprocessing, can for example, alter the fusion intervals in experiments such as those of Repp (1980).

In particular, PHONET proposes that speech is separated into transient (e.g., formant transitions in consonants) and sustained (e.g., vowel) components, and that separate working memories are activated that are sensitive to these transient and sustained portions of the speech stream. The model also proposes how interactions between these working memories can store rate-invariant representations of phonemic items. In the model, as different formant transitions excite different transient working memory cells, network interactions enable this working memory to estimate the input rate. Output signals from the transient working memory act to modulate, or control the gain of, the processing rate in the sustained working memory. In other words, when the system determines that initial transitions are arriving more rapidly, it sets the vowel processing channel to a correspondingly higher integration rate. The transient-to-sustained gain control tends to preserve the *relative* activities across both working memories as speech rate changes. The stored activities provide a basis for rate-invariant perception. The PHONET model quantitatively describes how phonetic category boundaries can shift as a function of speech rate (Miller & Liberman, 1979; Miller, 1981). The need for rate-invariant representations, however, does not preclude the existence of other working memories that are sensitive to rhythmic information, and other forms of prosodic information in general. In the model developed below, the working memory stores temporal order information in a rate-invariant way, but prosodic interplay needs to be an important component of any larger model (Cutler et al., 1997; Grossberg, 1986; Mannes, 1993; Pitt & Samuel, 1990).

The second issue concerns how to design the list grouping network that resonates with the working memory. This network must be able to pick out the best hypothesis consistent with the available bottom-up data. In some instances, even small list chunks may be selected and may command their own resonances, while at other times these small chunks are supplanted through time by larger chunks as new bottom-up data streams in. For example, consider the perception of the word “great”. The initial formant transitions specifying the /gr/ cluster and the following diphthong /ei/ jointly represent the word “gray”, and so a list chunk GRAY may become active prior to the arrival of the word-final /t/. However, even within the /grei/ sequence, the list chunk RAY has evidence from all its constituent phonemes because both the /r/ and /ei/ codes are active in the working memory. In fact, when the stop consonant /t/ arrives in the working memory, at least five list chunks that are themselves words — ATE, RAY, GRAY, RATE, and GREAT — can be assumed to be in active competition to establish a resonance with the phonemic codes in working memory. The design of the chunking network ensures that the largest chunk receiving activity from all of its phonemic inputs will win this competition. Due to the competition, or masking, between these multiple-scale chunks, such a network has been called a *masking field* (Cohen & Grossberg, 1986; Grossberg, 1978a, 1986). In order for a masking field to work correctly, its list chunks must exhibit *list selectivity*; that is, until all items supporting a given chunk receive bottom-up activation, that chunk can not become active enough to engage in a resonant feedback loop. In the example above, if the /t/ were not to arrive in the working memory within a suitable temporal window, then despite the masking field’s bias towards larger chunks, chunk GRAY would win the competition over chunk GREAT and would resonate with its items in the working memory.

Masking fields were introduced to solve a problem that is called the *temporal chunking problem* (Cohen & Grossberg, 1986; Grossberg, 1978a, 1984, 1986). This is the problem of unitizing an internal representation for an unfamiliar list of familiar speech units; e.g., a novel word composed of familiar phonemes or syllables. In order to even know what the novel list is, all of its individual items must first be presented. Thus, before the entire list is fully presented, all of its sublists will also be presented. What mechanisms prevent the familiarity of these smaller units from forcing the list always to be processed as a sequence of individual units, rather than eventually as a new familiar unitized whole? How does a not-yet-established word representation overcome the salience of well-established phoneme or syllable representations?

A masking field does this by giving the chunks that represent longer lists a prewired competitive advantage over those that represent shorter sublists. The intuitive idea is that, other things being equal, the longest lists are better predictors of subsequent events than are shorter sublists that comprise the longer list, because the longer list embodies a more unique temporal context. As a result, the *a priori* advantage of longer, but unfamiliar, lists enables them to compete effectively for activation with shorter, but familiar, sublists, thereby suggesting a solution of the temporal chunking problem.

It has elsewhere been shown how such a masking field can develop from simple developmental growth laws (Cohen & Grossberg, 1986). It has also been shown how it can naturally explain key data about list coding, such as the Magic Number Seven Plus or Minus Two (Grossberg, 1978a, 1986; Miller, 1956). Properties of the masking field also anticipated data about such properties as the word length effect (Samuel, van Santen, & Johnston, 1982, 1983), which shows that a letter can be progressively better recognized when it is embedded in longer words of lengths from 1 to 4. This property follows from the greater weight given to longer list chunks, together with the effect of these list chunks on their working memory items via top-down feedback; see Grossberg (1986) for further discussion.

Until the present time, all masking field simulations have been done using only bottom-up inputs from a working memory in order to demonstrate how longer list chunks can inhibit shorter list chunks without a loss of selectivity, how longer list chunks can be primed by bottom-up evidence from their sublists, and how the distribution of activity across the masking field can become more focused as more bottom-up evidence becomes available (Cohen & Grossberg, 1986, 1987). The present article takes the major step of showing how a multiple-scale masking field can be incorporated into a feedback loop with a working memory, with both bottom-up and top-down interactions operating continuously through time, and how the ensuing resonant dynamics of this feedback loop can be used to quantitatively simulate challenging data about phonemic grouping data in human speech perception, notably data about context-sensitive backward effects in time.

Thus, in the ARTWORD model developed below, phonemic representations dynamically emerge through working memory and masking field feedback interactions so as to support the perception of different combinations of the words “gray”, “great”, “ship”, and “chip” according to the segmental durations of silence and fricative noise. The serial position information in these representations emerges from several interactive properties. First, there are the different position-sensitive activity levels of items stored in working memory. Second, there are different relative sizes of the bottom-up and top-down weights in the pathways between the working memory items and the list chunks. When the working memory activities are filtered by the bottom-up weights, those list chunks are activated most whose weights best match the activity pattern across the working memory. After competition selects a subset of winning chunks, the order information represented by them determines the percept that arises through resonance.

The degree to which two chunks in the masking field compete with each other depends on how much they share inputs from phonemic items. Chunks like GRAY and CHIP are not in strong competition with each other, because the two chunks have no common input from phonemic item codes in the working memory. Both chunks, however, compete with the GREAT chunk, because of shared item codes. In particular, GREAT and GRAY both receive input from the /g/, /r/, and /ei/ items, while GREAT and CHIP are both sensitive to the initial noise present in the items /t/ and /tʃ/). Likewise, the chunks encoding GREAT and SHIP both inhibit the CHIP chunk, but do not strongly inhibit each other. In general, the greater the overlap of item input between two chunks, the greater the strength of the inhibitory interaction between those chunks. Previous work has shown that the rules governing the competition between masking field chunks can self-organize during development using activity-dependent self-similar cell growth laws (Cohen & Grossberg, 1986, 1987). Although the present model considers how only a single list chunk level works, one can imagine that a hierarchy of such levels exists in which higher levels can code larger language contexts, as well as smaller groupings that can propagate across levels.

In the ARTPHONE model (Grossberg *et al.*, 1997), the PHONET model (Boardman *et al.*, 1999), and the ARTWORD model developed below, quantitative simulations of isolated data sets are provided to illustrate how general principles of network processing can explain particular context effects and trading relations. The speech literature is replete with data on other context effects, in which the temporal properties of specific segment types, play important roles in their perception. Neither previous models nor ARTWORD have been developed to the point where all of these details have been incorporated into the network dynamics. These models have only begun to address the role of contextual temporal factors in speech perception, using simplified inputs in their simulations. While a completely realistic level of quantitative specificity remains a goal for future work, the previous and current ART models all contribute to the gradual elucidation of the dynamical processes that are involved in speech perception. In particular, ARTWORD is perhaps the first real-time model of speech perception that simulates speech context effects using a chunking network which generates retroactive re-segmentations of phonetic inputs that can leap backwards in time over the silent interval that separates two words.

3. ARTWORD: Adaptive resonance in word perception

The processes by which auditory signals activate phonemic item codes in the working memory, excite chunks in the masking field, and close a resonant feedback loop have been described within the framework of *adaptive resonance theory*, or ART (Grossberg, 1976a, 1976b, 1980). ART principles and mechanisms have been used to explain data about visual development, perception, learning, and object recognition (Carpenter & Grossberg, 1991; Chey, Grossberg, & Mingolla, 1997; Grossberg, 1994, 1999b; Grossberg & Merrill, 1996; Grossberg & Williamson, 1998, 1999; Grunewald & Grossberg, 1998). Within the domains of audition, speech perception, and language, ART models have been developed to explain data on auditory streaming (Grossberg, 1999c), word recognition and recall (Grossberg & Stone, 1986), manner distinctions in consonant perception (Boardman *et al.*, 1999), and consonant integration and segregation in VC-CV syllables (Grossberg *et al.*, 1997). These models embody several key ART design principles, including storage of temporal pattern information via the phonemic representation in working memories, automatic gain control to maintain rate invariance, and top-down matching to confirm expected bottom-up activation. In the present article, a model called ARTWORD applies these principles to the integration of multiple phonemic items into larger perceptual units by incorporating a multiple-scale masking field into a word recognition model.

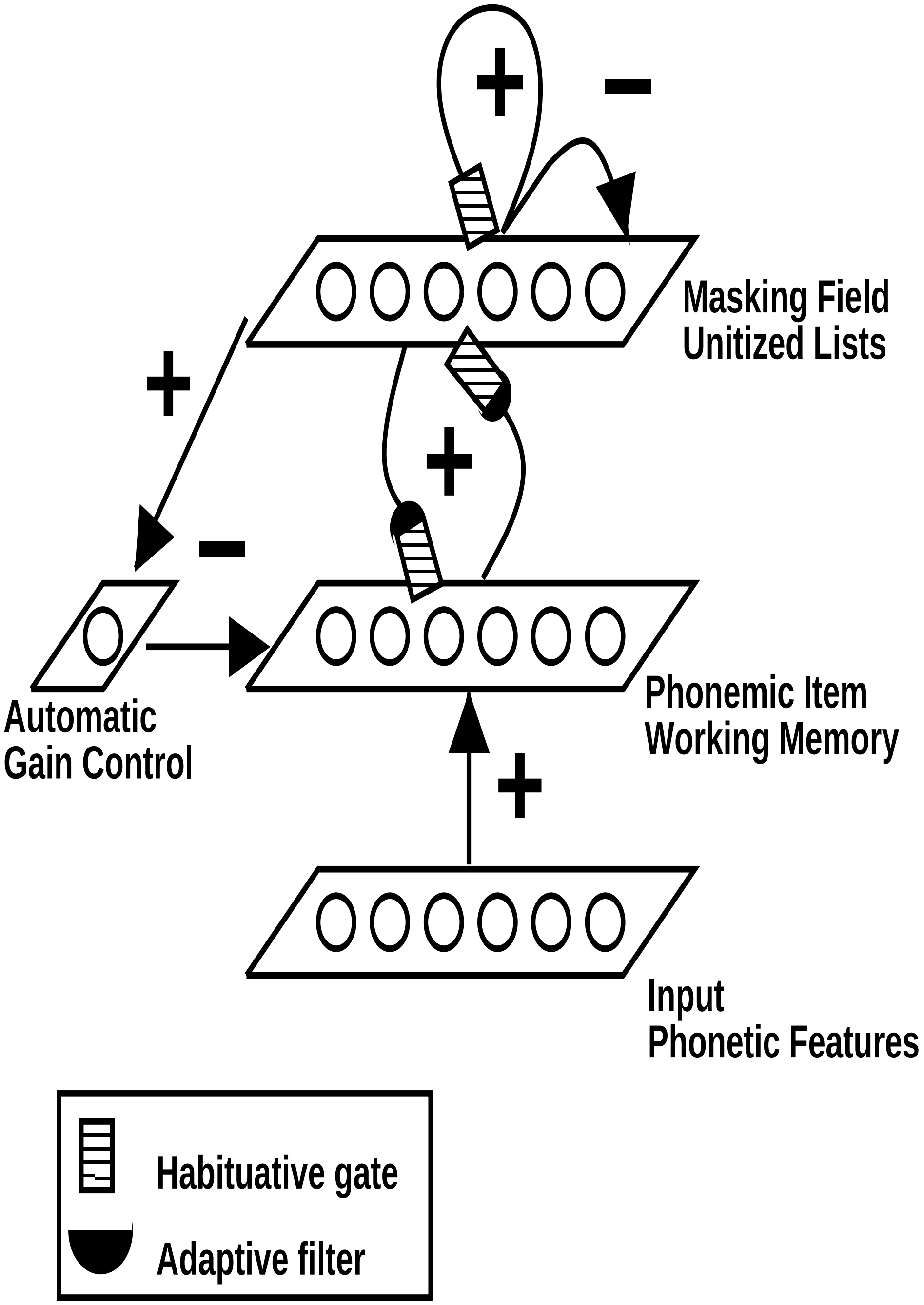


Figure 3: ARTWORD model architecture.

The ARTWORD model is shown schematically in Figure 3. Both the working memory and list chunk levels in Figures 2 and 3 can represent phonetic features, phonemes, syllables, and words, albeit in different ways. The phonetic context helps to determine which type of representation emerges. While it is still an open issue among psycholinguists whether phonemes are extracted prior to word identification, numerous data indicate that the nervous system performs an analysis of incoming speech into relatively primitive neural responses before resynthesizing them into a unitized percept. Exactly what the features, and the corresponding levels, represent remains an area of active research. In ARTWORD, these features correspond to standard units of psycholinguistic analysis of English. In general, the psycholinguistic data relevant to a given language will determine what units are present in each model level.

In ARTWORD, bottom-up processing of the acoustic signal, transduced through a learned acoustic-phonetic mapping, produces activation of item representations in the working memory (Fig. 4A). As each subsequent phonemic item is activated by current bottom-up input, competition within the working memory forces previously activated items to become less active, thereby forming a recency gradient wherein the most recent items are most active (Fig. 4B). Similar conclusions can be drawn if parameters are chosen to yield a primacy gradient in working memory. These short-term memory dynamics within the working memory network have been elaborated in the STORE working memory models; e.g., Bradski *et al.* (1994).

As the items exceed a critical threshold level of activation in the working memory, they excite masking field chunks that are tuned to prescribed activation patterns across the working memory items. Only those list chunks that receive input from all their item codes will reach supraliminal activity (Fig. 4C). As each list chunk receives its full complement of bottom-up activation, it crosses a positive feedback threshold and begins to support the items that excited it. Additionally, it sends inhibitory signals to the other list chunks in the masking field. Other things being equal, the list chunks that receive input from the largest array of items in the working memory (up to some maximal list length) have the strongest masking parameters, so they send the strongest inhibitory signals to the other chunks. In this way, the chunk with the most bottom-up support begins to hold sway within the masking field, and is able to suppress the competing list chunks and establish resonance with its working memory items (Fig. 4C). The resonance between the masking field and working memory is characterized by high activity levels among the items and the chunk(s) they select, and by suppressed activity among the other chunks and items. The chunk-item positive feedback signals are transmitted in both directions via the adaptive filters linking the two neural fields. For the duration of the resonance, both the resonating chunk and its items attain higher levels of activation than would be attained in a non-resonant state. This “resonant boost” of activation is proposed to represent the percept that emerges when the bottom-up input interacts with top-down expectations.

For a sequence of resonant events to occur during fluent speech perception, the positive feedback loop of any one resonance cannot continue indefinitely. Instead, the network is *reset* into a non-resonant state, so that the next resonance can be initiated. Two ART control structures govern reset of network activities. The first, known as *mismatch reset*, occurs when new phonemic information arrives which is sufficiently different from the currently active working memory pattern to warrant an arousal burst that rapidly resets activity in the masking field (Carpenter & Grossberg, 1991; Grossberg et al., 1997; Grossberg & Stone, 1986). The currently active items in the working memory reflect the most active hypothesis in the chunking network that is consistent with the

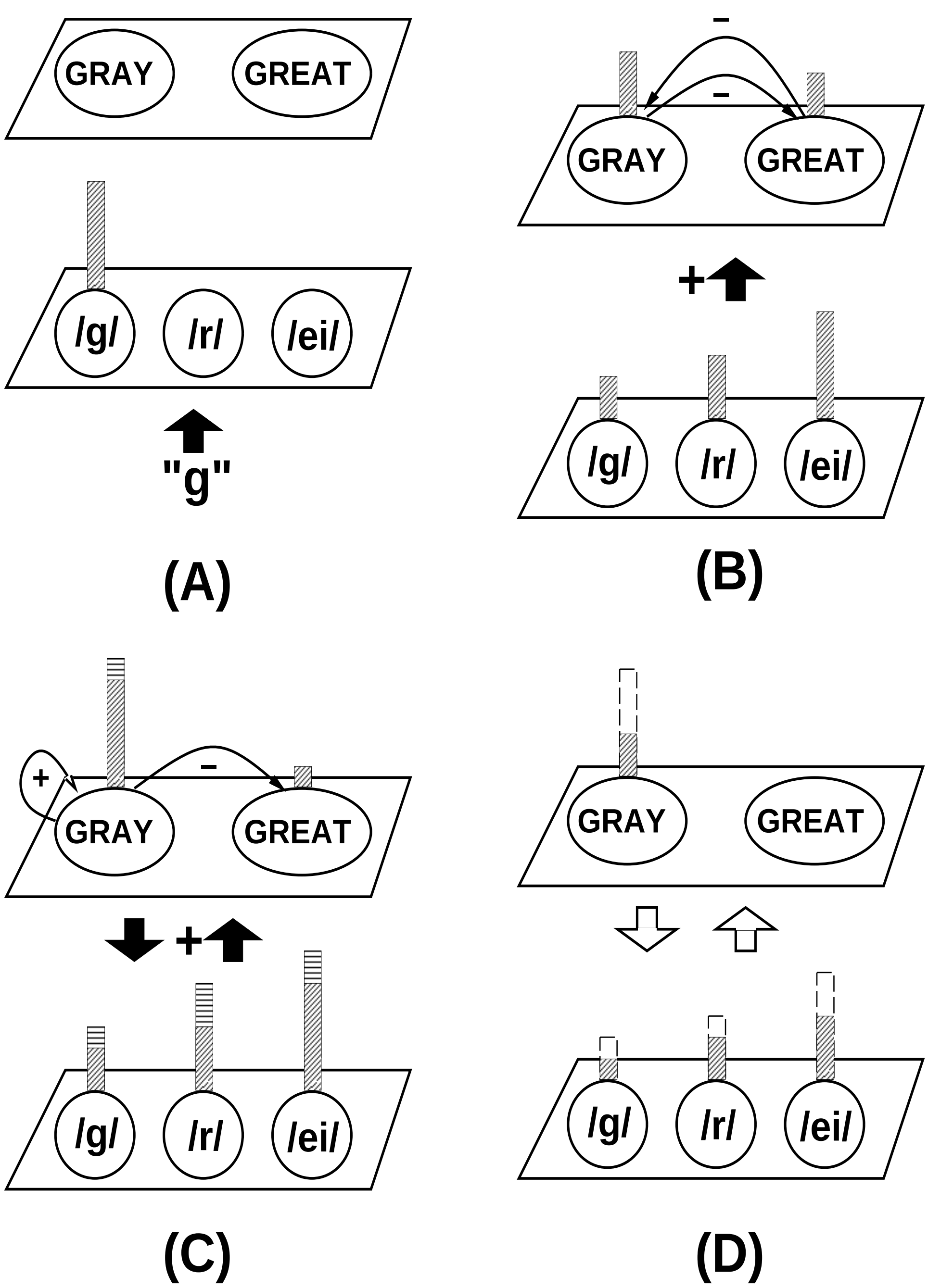


Figure 4: ARTWORD perception cycle: (A) Bottom-up activation. Acoustic inputs are processed and stored as phonetic items in working memory. (B) Chunk competition. A sequence of phonetic items forms a recency gradient in working memory. The list chunks which are activated by these items compete with each other in the masking field. (C) Item–list chunk resonance. The winning chunk crosses the resonance threshold, and enters a positive feedback cycle, exciting itself and its phonetic items in the working memory. (D) Chunk reset due to habituative collapse. As neurotransmitter levels habituate, the signals between levels fall below the resonance threshold, and the positive feedback cycle is broken. The vertical gray bars designate the activation of the corresponding item or list chunk.

Figure 4.

top-down feedback from the resonating chunk. The bottom-up input is compared with these items within the model's *orienting system*, whose cells are sensitive to mismatches between bottom-up and top-down information. If the mismatch is great enough to exceed a *vigilance* threshold, then a nonspecific arousal burst is emitted from the orienting system and quickly drives chunk activity in the masking field to zero and shuts down its top-down feedback. The working memory activity pattern can then select a different chunk with which to establish a new resonance.

The second reset mechanism, called *habituative collapse* (Grossberg *et al.*, 1997), provides a means for resonances to self-terminate in the absence of externally stimulated reset signals (Fig. 4D). This occurs when the synaptic neurotransmitters that convey excitatory activity between the working memory and the masking field habituate. The transmitters replenish at a slower rate than they are inactivated when signaling occurs along their synaptic pathways, so sustained activity between items and chunks results in an eventual depression of available transmitters and a consequent cessation of resonance (Grossberg, 1986). ART models have used properties of habituation, or depression, to explain a variety of perceptual phenomena, ranging from visual persistence and afterimages (Francis & Grossberg, 1996; Francis, Grossberg, & Mingolla, 1994; Grossberg, 1976a) to phonemic integration and segregation (Grossberg *et al.*, 1997).

Complex dynamics can arise within the competitive environment of the masking field before the network settles into a stable resonant state, as illustrated in Figure 5. In particular, variations in the amount of bottom-up evidence for particular items in the working memory can shift the balance within the masking field competition. Consider, for example, a masking field that is tuned to expect the three chunks, WX, XY, and YZ, where the chunks WX and YZ both strongly inhibit the chunk XY because of the shared items X and Y, but WX and YZ do not actively inhibit each other (Fig. 5A). If the bottom-up input supports the activation of the items W and X, followed by Y and Z, then all masking field chunks receive partial evidence from the active items in the working memory. The chunk XY, though, receives combined inhibition from the other two chunks, while the other two chunks are inhibited only by chunk XY. Such a scenario supports the *competitive teaming* of the two chunks WX and YZ against the single chunk XY. The teamed chunks, then, win the competition and establish the sequence WX and YZ of resonances with the working memory. If the inputs to the working memory were, instead, W followed by a sustained or doubled X, followed by Y, then under suitable temporal conditions, the network could generate a sequence of WX and XY resonances. In this example the possibility of a WXY resonance is precluded because no such chunk is assumed to exist in the masking field. Competitive teaming illustrates how differences in such input parameters as duration can result in different perceived groupings.

In addition to competitive teaming, a phenomenon of *resonant transfer* can occur when an additional input is added, after a suitable delay, to an already presented list of items. By this means, a resonance with the initial list can occur during the delay, but can be seamlessly replaced by a larger grouping as the temporal context unfolds. For example, consider a masking field containing the chunks XY and XYZ, and assume that items X and Y are presented sequentially, stored in working memory, and initiate a resonance with chunk XY (Figs. 5B-C). Suppose that an additional item, Z, is then presented as the XY resonance is winding down due to habituative collapse (Fig. 5D). The resonating chunk XY is then temporarily at a disadvantage in any ensuing masking field competition. Since there is a chunk XYZ present in the network, it has already been primed by the previously supported X and Y items and can thus initiate an XYZ resonance shortly after item Z is presented. During resonant transfer from chunk XY to chunk XYZ, the resonance shifts from the

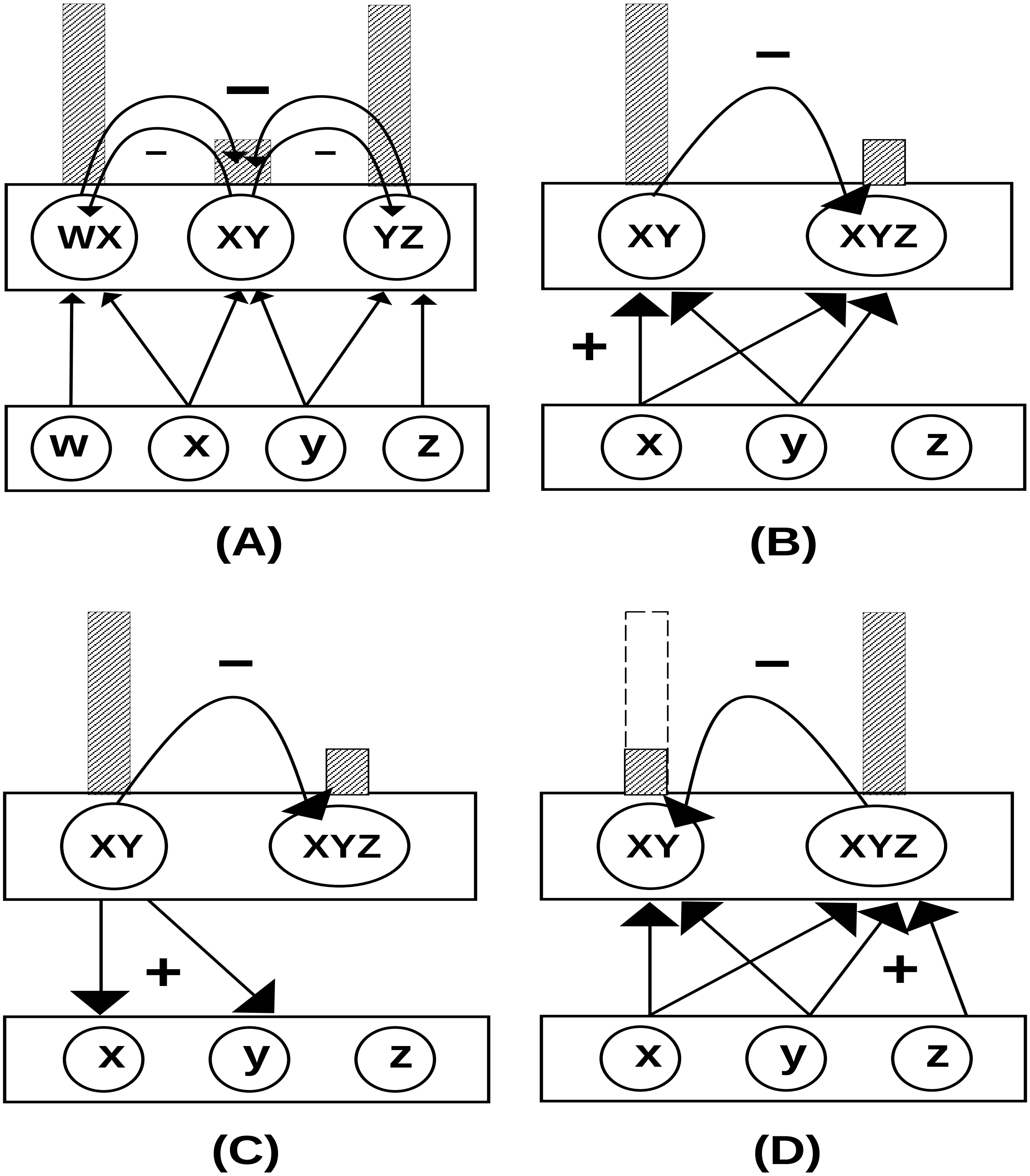


Figure 5: Grouping consequences of competitive teaming (A) and resonant transfer (B)-(D). In (A), each chunk receives complete support from its items, but chunk XY gets twice as much inhibition from competing chunks as do WX and YZ. Thus XY will not resonate, despite its large bottom-up input. In (B)-(D), items x and y initiate resonance with chunk XY (B-C), but when item z arrives as the chunk XY resonance weakens, chunk XYZ builds on its partial activation by x and y to form an XYZ resonance (D).

smaller chunk to the larger chunk. There is only a narrow temporal window under which such a transfer can occur. For example, if the final item occurs too late, the prior items will have fallen to lower activation levels, rendering them incapable of supporting a larger list resonance. The “final” item would then be treated by the system as a single item, or the initial item of a later list.

The two dynamic processes of resonant transfer and competitive teaming show how a masking field can go beyond the single-item grouping simulations in Grossberg *et al.* (1997) to explain multiple-item grouping data, such as the data of Repp *et al.* (1978). The ART processes described above are defined mathematically and illustrated with computer simulations below. Before presenting the model, we first describe in detail the relevant perceptual data of Repp *et al.* (1978) and others.

4. Identification and grouping of stop and affricate consonants into words

To perceive speech, listeners must integrate acoustic information on multiple levels and time scales (Repp, 1988). The coarticulation of consonants and vowels during speech produces an overlapped, interwoven arrangement of sounds that is perceived as a temporal succession of phonemes (e.g., Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967). Which phonemes are perceived depends crucially on the surrounding context, including the duration of silence, or the *lack* of acoustic energy, in ongoing speech. The classical study of Bastian, Eimas, and Liberman (1961) established in tape-splicing experiments that if a short interval of silence is spliced between the /s/ and /lit/ portion of the word “slit”, listeners perceive the result as the word “split”. The silent interval artificially inserted into the signal is sufficient to cue the perception of the voiceless stop consonant /p/. The experiments of Bastian *et al.* (1961) thus showed that the absence of acoustic energy can generate the perceived presence of a speech sound. These *silence cued stop consonants*, and the acoustic parameters that contribute to their perception, have since been the subject of detailed study, in the /s/-/l/, “say”–“stay”, “sa”–“spa”, and other contexts (Bailey & Summerfield, 1980; Dorman, Raphael, & Liberman, 1979; Fitch, Hawles, Erickson, & Liberman, 1980; Repp, 1984, 1985; Summerfield, Bailey, Seton, & Dorman, 1981).

The principal explanation given for listeners’ perception of silence-cued stop consonants stems from a proposed speech-specific mode of perception that makes reference to tacit knowledge of the articulatory gestures which produce stop consonants. Explanations at the level of purely psychoacoustic interactions have also been considered, but several studies seemed to argue against these. For example, with training, listeners can selectively attend to broadband noise in noise-silence-/laet/ stimuli and thereby avoid perceiving a stop (/p/ or /b/) (Repp, 1985). Also, listeners failed to perceive a stop in analogs of /sei/-/stei/ constructed from broadband noise (analogous to /s/) and sine wave tones (analogous to the formants of /ei/) when instructed to perceive them as “non-speech” stimuli (Best, Morrongiello, & Robson, 1981). The explanation in terms of articulatory knowledge relies on the fact that, in natural speech, stop consonants are those which by definition are produced by a temporary closure of the vocal tract and hence give rise to a brief pause in acoustic energy of the speech signal. Affricate consonants, or “stop-initiated fricatives”, such as “ch” (/tʃ/), and “dg” as in judge, likewise begin with a brief closure of the vocal tract (Hardcastle, Gibbon, & Scobbie, 1995; Stevens, 1993). Thus, the formant transitions into and out of vowels surrounding stop and affricate consonants are always present in the context of a brief silence. A speaker will thus be familiar with silence intervals that occur in these speech contexts. As Repp (1988, p. 251) put it, “a listener’s long-term representation of the acoustic pattern corresponding to a stop consonant thus includes the spectro-temporal properties of the signals preceding and following the closure as well as the closure itself...The silence thus is not really ‘actively’ integrated with the surrounding signal portions; rather, the integration has already taken place during past perceptual learning and is embodied in the perceiver’s long-term knowledge of speech patterns to which the input is referred during perception.” The ARTWORD model developed below shows how previously learned differential responses to input stimuli preceded by silence may combine with the temporal displacement effect of the silent interval itself to produce trading relations between silence and the acoustic characteristics (e.g., segment durations) of the following phoneme.

Subsequent experiments have determined a complex relationship between the relative duration of the silence interval and its surrounding context. As noted by Repp (1988, p. 250), relative silence duration is a cue for voicing, manner, and place of stop consonant articulation. For example, Bailey and Summerfield (1980) found that after inserting silent gaps of various duration in /s/-vowel stimuli, listeners perceived /s/-stop-vowel. On average, 20-30 ms of silence were sufficient to induce perception of a stop consonant. Which stop consonant listeners perceived depended crucially on the duration of the silent interval. For example, for a given stimulus series, a 60 ms closure might give a high probability /ska/ percept while an 90 ms closure might give a high probability /spa/ percept). Similarly, Repp (1984) reported that silence closure duration in an /s-l/ context was a primary cue for stop place, with shorter gaps perceived as “t” and longer ones as “p”. The silence durations that can cue stop perception vary according to many acoustic properties of the signal, but, for example, in the /s-l/ context typically range from “60 ms to 300 ms, with the peak occurring at 100–150 ms of silence” (Repp, 1985, p. 802). Relative silence duration interacts with other acoustic cues including spectra and duration of /s/, presence of a release burst and formant transitions after the silence, and duration of the following voiced segment. Together, these spectral features and their temporal arrangement all contribute to perception of the stop in a context-specific manner (Repp, 1985). The ARTWORD model suggests how, even when each item in a sequence receives identical bottom-up input, variations in the duration of the silent interval by itself can play a key role in determining how the competition between chunks is resolved, and how the subsequent resonance – and the perceived grouping it determines – unfolds.

Motivated by knowledge that silence can cue the perception of stop-consonant manner within a syllable, Repp *et al.* (1978) went on to show that the perceived stop or affricate can cross word boundaries. As described earlier, they presented listeners with versions of the sentence “Did anyone see the gray ship?” that varied both the duration of the fricative noise /$\int$/ in the beginning of “ship” and the duration of the silent interval between the words “gray” and “ship”. Depending on the lengths of the two intervals, listeners reported perceiving “gray ship”, “great ship”, “gray chip”, or “great chip”. The introduction of a sufficiently long silent gap brought about the perception of a “stop-like” sound – either the stop /t/, the affricate /t$\int$/, or both. Depending on how the different cues varied, though, that stop-like sound could attach to a different word. Strictly temporal manipulations in the acoustic signal could shift the balance of perceptual evidence one way or another.

To create the test stimuli, Repp *et al.* (1978) inserted silence intervals of duration from 0 to 100 ms in 10 ms steps before the word “ship”. The duration of the fricative noise in the word “ship” (originally 122 ms) was varied by excising or duplicating a 20 ms interval from its center. This procedure left the onset (up to the first 62 ms) and offset of the fricative noise unaltered. Four noise durations (62, 102, 142, and 182 ms) were generated, giving a total of 44 test stimuli (11 silence durations $\times$ 4 noise durations). The stimuli were recorded in 5 different randomizations with 2 sec

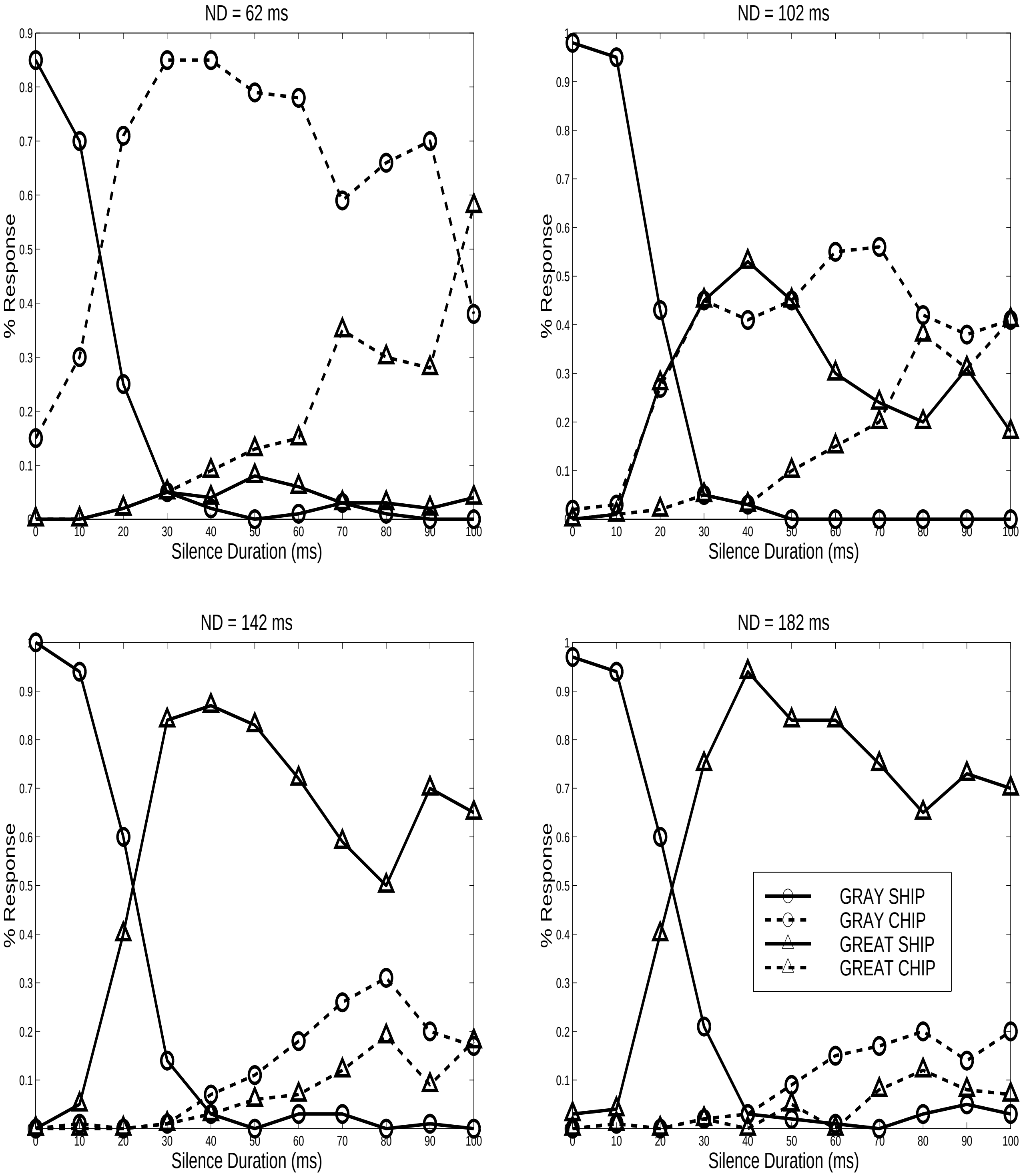


Figure 6: Repp *et al.* (1978) two-word response probabilities (redrawn from Repp *et al.* (1978), Figure 3, p. 629.

intervals between sentences, and presented to each of 10 subjects twice, so that each subject gave 10 responses to each stimulus. Repp *et al.* (1978) reported the averaged responses across the 10 subjects; individual variability for these data were not reported.

Figure 6 shows the results of the Repp *et al.* (1978) experiment. For each of the four noise durations (ND), the four alternative response probabilities are plotted as a function of silence duration. Figure 6 reveals significant patterns in the subjects' responses. First, a minimum silence duration of approximately 20 ms was necessary for any response containing a stoplike percept (/t/ or /tʃ/) to be reported consistently. For silence durations above this, either one or two stops were reported nearly 100% of the time, with the probability of two stops ("great chip") increasing with both increasing silence duration and decreasing noise duration. At the longest silence durations, the dominant response preference is seen to become less probable at all four noise durations, but this is particularly noticeable at the 102 ms noise duration. At this noise duration, the most probable response over the mid-range (60-80 ms) of silence durations, "gray chip", is roughly equiprobable with two different responses at lower and higher silence durations: "great ship" between 20 and 50 ms, and "great chip" between 80 and 100 ms. One of these two secondary alternatives accounted for at least 20% of the responses at every silence duration above 20 ms. The uncertainty, or compatibility of multiple responses, at the 102 ms noise duration suggests the conjoint activation of multiple percepts. (An alternative explanation, which cannot be ruled out from the reported results, is that a single percept was reliably determined by each individual, but variability across individuals created the reported psychometric functions. However, the existence of multiple responses reported with high probability in this region indicates uncertainty, whether due to individual variation, the inherent activation of multiple competing percepts, or both.) Figure 7 parcels out the single word, or marginal, response probabilities for "gray" and "chip" obtained for each word by summing across the two relevant response alternatives (e.g., $P(\text{GRAY}) = P(\text{GRAY SHIP}) + P(\text{GRAY CHIP}))$. The uncertainty at shorter noise durations (62-102 ms) is reflected in Figure 7 at the nearly 50% probability of a "gray" response, indicating the approximately equal likelihoods of grouping the stop consonant percept /t/ with /grei/ to yield "great", with /ʃIp/ to yield "chip", or with both words to yield "great chip" responses. These results reveal trading relations between silence and noise durations, such that for certain ranges, an increase in silence duration that would normally cause a perceptual switch can be offset by a corresponding increase in noise duration.

Dorman *et al.* (1979) further probed the affricate/fricative contrast observed in the Repp *et al.* (1978) data by inserting silent intervals between the words "say" and "shop" in the utterance "please say shop", thereby generating the perception of "please say chop". As in the Repp *et al.* (1978) experiments, silence was a sufficient cue for the manner distinction between the fricative /ʃ/ and the affricate /tʃ/. Dorman *et al.* (1979, p. 1526) found that a silent closure of 70 ms resulted in a 75% "chop" response rate. Notably, this effect disappeared if the "please say" and "shop" portions of the stimuli were uttered by different speakers (a male and a female): no amount of silence between the two utterances caused subjects to perceive "shop" as "chop". This suggests that listeners use their sensitivity to the vocal tract that produced the utterance to determine whether silence is perceived as a closure in an ongoing speech stream – thus providing acoustic evidence for the production of a stop or affricate – or as an ecological change in source which generates a separate perceived auditory stream (e.g., Bregman, 1990; Govindarajan, Grossberg, Wyse, & Cohen, 1994). Dorman *et al.* (1979) also showed that the chop-shop boundary shifts systematically with variations in the

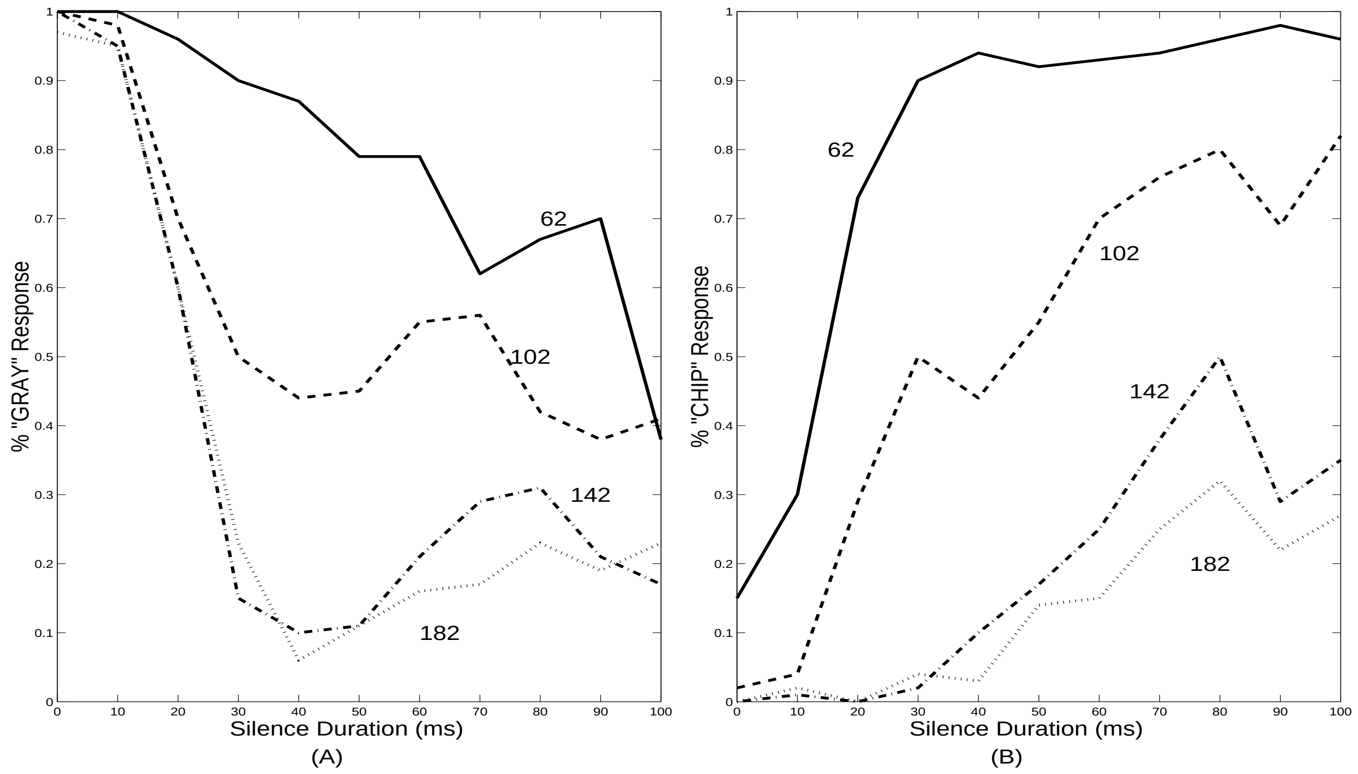


Figure 7: Single word (marginal) probabilities obtained from Repp *et al.* (1978) data. (A): “GRAY”. (B): “CHIP”. Numbers indicate duration of fricative noise.

duration of the fricative noise and the rise-time of its amplitude envelope. By halving the noise duration (from 320 ms to 160 ms), the chop-shop boundary shifted from 75 ms of silence to 55 ms of silence. The shorter noise, more characteristic of an affricate, required less preceding silence to be perceived as an affricate. Similarly, making the noise onset more abrupt by removing 30 ms of the initial /ʃ/ rise time (originally 35 ms long), Dorman *et al.* (1979) were able to shift the chop-shop boundary to silence durations approximately 20 ms shorter. These data indicate the interaction of expected acoustic cues to signal a phonetic contrast (e.g., noise duration and rise time) with local variations in the presentation rate caused by silence. As in the Repp *et al.* (1978) data and in the ARTWORD model presented below, a change in the silence duration differentially alters the percept depending on the acoustic context in which it occurs.

In the Repp *et al.* (1978) experiments, the perceptual system must decide both *what* phonemes have occurred (e.g., /t/, /ʃ/, /tʃ/), and *where* they go; that is, to what larger units they should be bound. This is a special case of the problem of detecting syllable and word boundaries, or junctures. Early studies of juncture perception focused on the local acoustic cues normally available to aid listeners in such decisions (Christie, 1974; Nakatani & Dukes, 1977). Disjunctures often function as a primary cue. For example, in the phrases “lighthouse keeper” and “light housekeeper”, the relative durations of silence between “light” and “house”, and “house” and “keeper” determine the resulting percept (Wickelgren, 1976). Many other acoustic cues associated with the phonemes immediately preceding and following the juncture also, in general, contribute to the percept. For example, aspiration of syllable-initial voiceless stops (“a|sta” vs. “as|ta”), the presence of formant transitions before or after the disjuncture, and allophonic variation can all function as cues to juncture (Christie, 1974; Darwin, 1976; Mattys, 1997).

Nakatani and Dukes (1977) tested perception of juncture by constructing hybrids from phrases like “play taught” and “plate ought”. The transitions to and from the juncture consonant were spliced out and replaced in the different original phrases in various orders, producing four possible percepts for each phrase (e.g., play ought, play taught, plate ought, and plate taught). They found that only the immediate neighborhood of the juncture consonant contained juncture cues, and that “the strongest cues for juncture perception occurred at the beginning of the word” (Nakatani & Dukes, 1977, p. 719).

Samuel *et al.* (1984) used a selective adaptation paradigm to probe whether an intervocalic stop (e.g., /b/ in /aba/) was perceived as belonging to the first or second syllable. Constructing a stimuli series that varied from /aba/ to /ada/, they presented adaptors to shift the /b/-/d/ category boundary. Only CV syllables (“ba” and “da”), and not VC syllables (“ad”, “ab”), were effective adaptors. Further selective adaptation experiments with VCCV stimuli indicated that the perceptual system treats an intervocalic stop “more like a syllable-initial stop than a syllable-final one”, although “it is not really perceptually the same as either kind” (Samuel *et al.*, 1984, p. 1661). The findings of Samuel *et al.* (1984) and Nakatani and Dukes (1977) both point to the importance of the syllable-initial segment in providing juncture cues. The model developed below to explain the Repp *et al.* (1978) data demonstrates how altering a “syllable-initial segment,” or, more properly, the segment immediately following the disjuncture, can shift the competitive balance between units, resulting in a difference of perceived juncture.

More recent studies of juncture perception have analyzed the role of prosodic information, and in particular lexical stress, as a primary cue for juncture perception (see, e.g., reviews by Mattys (1997) and Cutler, Dahan, and van Donselaar (1997)). Analyses of large vocabulary databases by Cutler and colleagues (Cutler & Carter, 1987; Cutler & Norris, 1988; Cutler, 1990; McQueen, Cutler, Briscoe, & Norris, 1995) have shown that the large majority of content words in English (roughly 90% when frequency of occurrence is accounted for) begin with stressed syllables. This suggests a "metrical segmentation strategy", in which listeners attempt to begin a new grouping of speech units with each occurrence of a stressed syllable, backtracking as necessary to correct errors generated by this strategy (Cutler & Norris, 1988; Cutler, 1990). Mattys (1997) reviewed several features of stressed syllables, including physical salience, phonemic stability, and perceptual distinctiveness, which support the idea of syllable stress as a key factor in generating word segmentations. The role of other prosodic factors, and in particular speech rate (see e.g., Pickett, Bunnell, and Revoile, 1995), as a cue to syllabification have recently come to bear on computational models of speech recognition (Price & Ostendorf, 1996; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991).

Together, these speech data support the view that both the perception of phonetic contrasts and the perceived phonemic groupings that result from these contrasts depend critically on the time scale and persistence of item activation in the phonemic working memory. As competition evolves between chunks, the changing neural activity patterns stored across the working memory provide different degrees of evidence to the chunks. The emergent resonant time scales which determine the perceived groupings, then, must be commensurate with how the input to phonemic item codes is traded against silent intervals, changes in speech rate, and lexical stress that modulate the dynamic processing windows within which the chunk-item resonances develop.

5. Sensitivity to Informational and Durational Phonetic Evidence

Variations in the durations of intersyllable silence and syllable-initial noise impact network behavior in two distinct ways: either by directly altering the strength of the input to the working memory, or, indirectly, by arriving at different times during the network processing cycle. These two routes by which segment durations can alter network responses may be considered in terms of what Mattys (1997) has recently described as "informational" and "durational" factors in speech perception. While the influence of coarticulatory smearing of phonetic information in speech is significant, the speech stream is predominantly sequential. But, "despite the intrinsic correlation between *time* and the *speech information* that it brings to the listener, these two variables have an independent impact on lexical processing" (Mattys, 1997, p. 311, italics added). Thus, for example, a silent interval spliced between "gray" and "ship" not only begins to provide evidence to the listener of a stoplike sound between the vocalic /ei/ and the fricative noise, it also allows the listener more time to process the /grei/ input before the next phoneme arrives, and hence the internal representation of the GRAY chunk may reach greater levels of activation by the time the noise does arrive. We describe below the distinction between these two factors in the ARTWORD model: the informational, defined by the local, low-level transduction of the acoustic stimulus into phonemic inputs, and the durational, which affects processing dynamics globally.

The response of phonemic item codes in the working memory is determined through prior learning which has adapted the long-term memory weights along the pathways between lower auditory processing levels and the phonemic item working memory. These pathways encode phonemic item sensitivity to neural activity patterns defining particular external acoustic events, or an *acoustic-phonetic* mapping (Pisoni & Luce, 1987). This learned acoustic-phonetic mapping represents the combined influence of peripheral auditory neural processing, like short-term adaptation within individual nerve fibers (e.g., Delgutte, 1980) and low-level integrative processes across networks of neurons responsive to specific acoustic patterns (e.g., Boardman *et al.*, 1999). Synaptic adaptation along the pathways reflects the statistical distribution of repeated exposure to speech sounds. In the present article, all learned tuning of synaptic pathways between the input and item levels, and between the item and chunk levels, will be assumed to have stabilized during prior developmental stages.

The tuning of synaptic weights on the pathways feeding into the phonemic working memory derives from the long-term average of the spectro-temporal characteristics of the phonemes which listeners hear. Because of the multiplicity of acoustic cues which specify phonetic contrasts, and their intricate dependence on context, it is likely that multiple phonemic codes representing different cue-combinations exist. For example, Hedrick (1997) lists frication duration, formant transitions, frication spectrum, and relative amplitude between frication and vocalic signals as components influencing the perceived place of fricative consonants. Input to the phonemic working memory in ARTWORD was chosen to roughly correspond to the same relative durational trends reported in the literature. Howell and Rosen (1983) measured tokens of /ʃ/ and /tʃ/ and found, for word-initial segments in running speech, mean rise-time durations of 123 and 37 ms respectively; the duration of the noise from end of the rise-time on was the same (48 ms) for both, yielding net durations of 171 and 85 ms for /ʃ/ and /tʃ/, respectively. Crystal and House (1988b) reported the high frequency of stop consonants occurring without a plosive release burst, or “hold only” stops. For example, at the end of a word followed immediately by another word (i.e., in the word-final, nonprepausal position) only 36% of the occurrences of /t/ in their data ($N = 363$) were complete, consisting of both a closure and a burst. The mean duration for all complete voiceless stops in their data was 92 ms, while the hold only voiceless stops, had a mean duration of 56 ms. However, in detailed studies of a 14 speaker corpus of speech, Crystal and House (1988a, 1988b) have highlighted the variability of speech segment durations, noting that even after separating tokens according to several phonetic dimensions, the distributions of segmental durations overlap considerably. In ARTWORD, the compressed item code for the fricative consonant /ʃ/ responds more vigorously to a longer fricative noise interval than the item code for the affricate consonant /tʃ/, all other things being equal. Likewise, the response of the item code for the stop /t/ shows a greater response when a silent interval precedes the noise which activates this item code.

There is some evidence that these distinctions can be encoded in the average discharge rates of auditory neurons, both peripherally and centrally. For example, based on his studies of peripheral responses to speech-like stimuli, Delgutte (1982) proposed a model by which short-term adaptation can account for the trading relation between silence duration and frication rise time in the affricate/fricative contrast in /atʃa/-/aʃa/ stimuli. The model consisted of a bandpass filter, envelope detector, sigmoidal nonlinearity, and short-term adaptation element. The model output in response to synthetic /aʃa/-/atʃa/ stimuli shows that decreases in rise time or increases in silence duration – both cues for “acha” – produced similar increases in the discharge rate of neurons tuned to the approximate frequency of frication. Delgutte and Kiang (1984, p. 896) similarly provided data suggesting that “the central processor should be able to distinguish between various voiceless fricatives even if limited to information carried in the average discharge rates of the most sensitive auditory-nerve fibers.” Thus even simple, peripheral auditory processing can begin to explain trading relations between preceding silence and rise-time duration like those described by Dorman *et al.* (1979).

The case that the responses of single auditory neurons can encode complex information integrated over relatively long temporal intervals was recently strengthened by the discovery of cells selectively tuned to sound duration within cat auditory cortex (He *et al.*, 1997), extending previous reports of duration tuning in the frog and bat at the brainstem level (e.g., Casseday, Erlich, and

Covey, 1994). He *et al.* (1997) described neurons in the dorsal zone of auditory cortex with complex response profiles, including multi-peaked tuning curves and long latency responses ($\geq$ 30 ms, 85% between 30 and 120 ms) to noise bursts. Of special interest with regard to speech-like stimuli were reports of neurons whose discharge rates showed monotonically increasing, decreasing, or unimodally peaked profiles as a function of the duration of noise bursts that vary between 20 and 500 ms. For example, long-duration-selective neurons, many of which required minimal stimulus durations to exhibit any response, either showed increasing discharge rates with stimulus duration (nonduration threshold neurons), or a saturating response which did not increase with further increases in stimulus duration (duration threshold neurons). Short-duration-selective neurons, by contrast, showed a maximal response to brief (e.g., 50 ms) noise bursts, and decreasing responses as stimulus duration was increased. These data raise the possibility that, for example, neurons responsive to /tʃ/ -like stimuli will first increase and then decrease their discharge rates when presented with the long fricative noise in a typical /ʃ/ stimulus. Likewise, neurons responsive to /ʃ/ -like stimuli may show greater latencies and gradually increasing discharge rates over the duration of a fricative stimulus. ARTWORD adopts a similar scheme, assigning complementary input durations to /tʃ/ and /ʃ/ item codes, with /tʃ/ input durations decreasing as fricative noise duration increases.

Apart from the "informational" phonetic evidence transduced to the working memory based on the statistics of prior speech exposure and the lower-level auditory processing, the segmental durations of silence and noise can influence network behavior "durationally", by arriving at different times and altering ongoing dynamic competitions. Because item and chunk activations grow and decay in real time, a pause or lengthening of any input segment, or any intervening silence interval, will alter the relative pattern in working memory which may in turn unbalance a developing competition between chunks in the grouping network. Recent evidence of Faulkner, Rosen, Darling, and Huckvale (1995) points to the possibility of such dynamic interactions in the /tʃ/-/ʃ/ contrast in the /aʃa/ context. Rosen, Darling, Faulkner and Huckvale (1993) and Faulkner *et al.* (1995) constructed factorial combinations of syllable-initial (/tʃa/, /ʃa/) and intervocalic (/atʃa/, /aʃa/) stimuli by varying frication duration (120-220 ms), rise time (0-100 ms), and, for the intervocalic stimuli, silence duration (0-80 ms). The averaged responses of nine subjects were analyzed. Contrary to the previous data reviewed above showing a shorter rise time as a positive cue for affricate perception, Faulkner *et al.* (1995) found that at short silence durations (0 and 20 ms), longer rise times actually produced more affricate responses. Only in the syllable-initial stimuli did the proportion of affricate responses decrease with increasing rise times. These data thus cannot be explained solely on the basis of the Delgutte (1982) peripheral auditory model. Faulkner *et al.* (1995) point out that it is unclear how other models that do not permit the statistical interaction of acoustic features (e.g., the fuzzy logical model of Massaro, 1987)) can satisfactorily account for the observed interactions. While models based on acoustic features and auditory processing go part of the way to explaining these data, Faulkner *et al.* (1995) argue, further explanation by way of a top-down or cognitive interaction is needed. In ARTWORD, durations of segmental excitations in the item field directly shift the competitive balance in the grouping network. When a word chunk does emerge as the winner, it feeds back to the item field, boosting phonemes over a perceptual threshold. By delaying the formation of the perceptual code until the top-down feedback supplies later-occurring information, ARTWORD provides a quantitative realization of the type of hypothesis suggested by Faulkner *et al.* (1995).

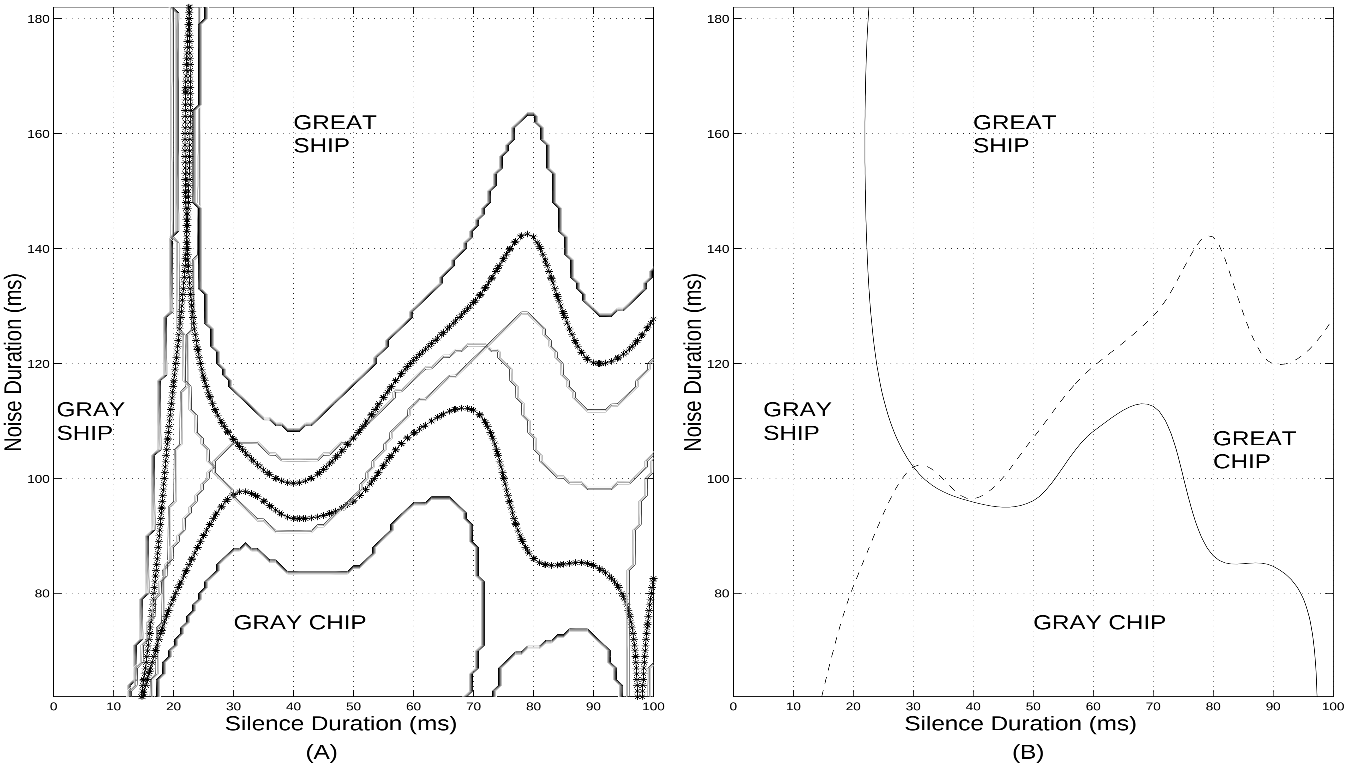


Figure 8: (A): Category boundaries derived from the probabilities in Figure 6 by interpolation. (B): Category boundaries derived from the single-word response probabilities in Figure 7.

Together, the activation of the phonemic item codes and the competitive grouping processes provide explanations of the percepts reported in the Repp *et al.* (1978) data. While Figure 1 provides a good indication of how the perceptual regions depend on silence and noise, the actual response probabilities bely a complexity not apparent in this representation. Figure 8 shows this complexity, and in particular the uncertainty associated with these regions, in greater detail. Because the responses were sampled at only four noise durations, the derivation of any representation of the perceptual space must interpolate to estimate the category boundaries. For example, Repp *et al.* (1978) derived the boundaries of Figure 1 using the probit method (which effectively performs an inverse cumulative Gaussian transform and interpolates by linear regression) to estimate the combination of silence and noise durations at which each of two alternative responses were equally likely. That is, each boundary in Figure 1 was computed between only two alternatives. However, because of the sparsity of noise durations and the fact that “great chip” responses were comparatively rare, this method appears to overestimate the size of the “great chip” region. As Repp *et al.* (1978, p.631) note, “There was no obvious dependency of this boundary on noise duration; the uppermost data point, which may suggest such a dependency, was based on only a few observations, since at this noise duration (142 ms) GREAT SHIP responses predominated.”

In Figure 8, two alternative representations of the perceptual boundaries are presented. To derive the boundary curves in both panels, the response probabilities were interpolated with a cubic polynomial and the contours of 50% probability for each percept were determined. In Figure 8A, the category boundaries are derived from the two-word responses in Figure 6 and are plotted in thick lines, with the corresponding 60% and 40% boundaries in thinner lines. This figure makes it evident that, for silence durations greater than 20 ms, at noise durations between 100 and 120 ms, the large perceptual uncertainty (discussed above) exists. The “great chip” percept is only the most probable response at the longest silence durations and at noise durations below 120 ms. However, either “great” or “chip” is always perceived provided the silence exceeds about 20 ms. This is made evident in Figure 8B, which shows the single word (gray-great and chip-ship) boundaries derived from the data in Figure 7. This representation conveniently partitions the entire perceptual space and shows the dominant first and second word responses at each combination of silence and noise. In order to avoid postulating a higher-level decision mechanism for probabilistically combining single chunk activations, we chose to fit the ARTWORD model to the single word responses of Figure 7. Note, however, that this does not imply, either in the data or the model predictions, that these single word response probabilities are independent of each other. Indeed, a chi-squared test for statistical independence of the first and second word responses (i.e., a test of the hypothesis P(GRAY SHIP) = P(GRAY)P(SHIP), etc.) rejects at high significance levels; likewise, in ARTWORD the generation of all chunk activations are crucially interdependent. The perceptual boundaries are emergent properties of network interactions and, as such, merely reflect one representation of the underlying dynamic generation of resonant events.

Because the ARTWORD model generates the perceptual codes dynamically from the system interactions between bottom-up driven working memory responses and top-down grouping processes, the behavior of these perceptual codes cannot be simply attributed to a single parametric source such as the presence or absence of an acoustic feature. However, considerations of the network responses to inputs presented with different combinations of silence and noise can provide insight into the transitions between perceptual regions in Figure 1 For example, the percept of “gray ship” in region 1 can be primarily attributed in ARTWORD to the strength of the phonemic item responses to the input at brief silence durations. In particular, because silence is an important cue for the perception of stops and affricates, neither the /t/ or /tʃ/ items receive strong excitatory input when the fricative noise immediately follows the vocalic /ei/ segment. With increasing silence, the /t/ and /tʃ/ items are excited for longer durations, and with increasing durations of fricative noise, the /t/ item receives greater excitation. Thus the transitions out of region 1 can be expected on the basis of these phonemic responses: the unitized representations most likely to resonate with working memory will be naturally selected based primarily on the match between the acoustic signal and the learned phonemic representations.

The transition between regions 2 and 3, however, requires an explanation based on the grouping operation involved: the acoustic signal in both cases contains sufficient cues for the perception of a stoplike sound. The only difference is where the stop is grouped. The model explains this transition by describing a competitive grouping operation that dynamically emerges at a slow enough rate to allow the first competition (GRAY vs. GREAT) to be influenced by the later-occurring noise and the second competition which it engenders (GREAT vs. CHIP). When evidence for the /tʃ/ item is strong, at lower noise durations, the GRAY and CHIP chunks can both win their competitions with the GREAT chunk by virtue of their competitive teaming. At longer noise durations, the /ʃ/ item receives proportionally more excitation, so the CHIP vs. SHIP competition weakens the CHIP chunk's activation. This, in turn, permits the GREAT chunk to attain greater levels of activation and win its competition with the GRAY chunk. In this way, the activation level of the SHIP chunk can indirectly help determine whether the GRAY chunk resonates with its items, despite the fact that the SHIP and GRAY chunks do not receive input from any shared phonemic items. ARTWORD also suggests why, at increasing silence durations, the boundary between regions 2 and 3 is slanted upwards, so that more noise is required to perceive "great" than "gray" when the silent interval between /grei/ and the noise is increased. As the GRAY chunk attains greater activations during the longer silent interval, the GREAT chunk is correspondingly inhibited, so greater /t/ activation is required to initiate a resonant transfer from GRAY to GREAT.

The GREAT chunk can also resonate if the /t/ input arrives late enough so that the GRAY chunk has begun to weaken due to the habituation of its transmitters. The transition between region 2 and region 4 (GRAY CHIP to GREAT CHIP) indicates that at sufficiently long silence durations, the resonance between GRAY and its items is susceptible to a transfer. Thus, in region 2, GREAT is inhibited by the proximal future activation of CHIP. In region 4, the stop manner cues associated with /t/ are distal due to the long silence duration. The GRAY chunk initially wins its competition with the GREAT chunk as in region 2. However, the /t/ item then becomes active and, as GRAY completes its natural resonance cycle, all items for GREAT are present, so GREAT enters its own resonant cycle, completing the transfer of /grei/ item information forward in time to adjoin the /t/ information.

6. Simulations of Resonant Transfer and Competitive Teaming

Computer simulations of the ARTWORD model were performed to illustrate aspects of multiple item grouping and resonant dynamics. Appendices A and B describe the network equations and parameters, respectively, that were fixed for all simulations included in the present article. Simulations were performed by second order Runge-Kutta numerical integration with an adaptive step size (MATLAB 5.2).

6.1 Bottom-up activation of list nodes

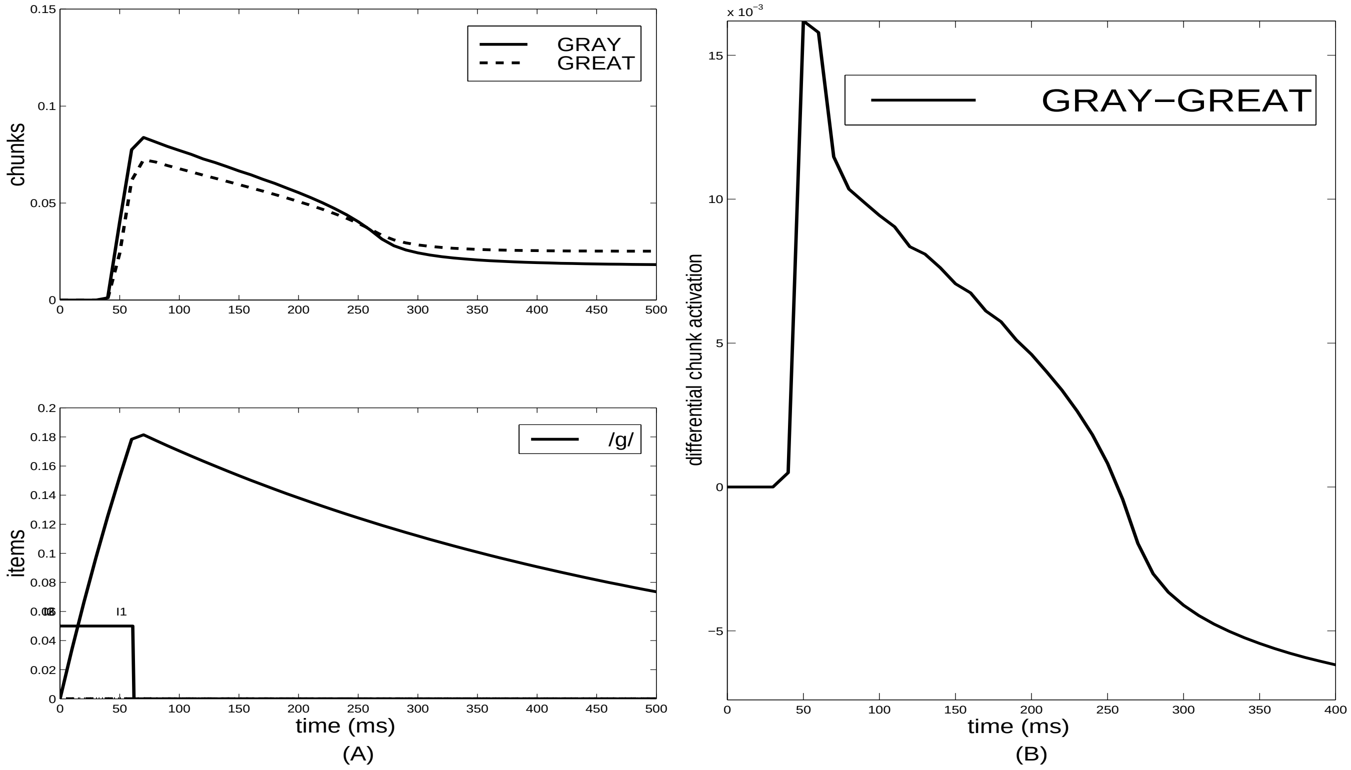


Figure 9: (A): Response of two chunks to a single item. (B): Differential activation of chunks.

The first group of simulations demonstrates the bottom-up effect of item activation on chunk activities in the absence of top-down feedback. Figure 9 shows the response of two chunks in the grouping network, GRAY and GREAT, to the presentation of the single item /g/. Both chunks show brief bursts of activity, but do not receive sufficient input to sustain their climbs. The GRAY chunk responds more strongly than the GREAT chunk to the single item /g/ for two reasons. The first is due to the normalization of input to chunks, via conservation of synaptic sites: larger chunks, like GREAT, receive input from more neurons in the working memory and therefore each input contributes relatively less excitation. A second reason results from synaptic learning as a result of long-term exposure to specific patterns. The GREAT chunk has been tuned through competitive learning to expect a four-item pattern (/g/, /r/, /ei/, and /t/), while the GRAY chunk expects only a three-item pattern (/g/, /r/, and /ei/). Because of the passive decay and lateral inhibition that occurs within working memory, when longer lists are fully stored, the activity of the items that are stored early in the list are smaller than those of shorter lists. Thus, the synaptic weights between the /g/ item and the GREAT chunk have been tuned to expect smaller values than the weights between /g/ and the GRAY chunk. Figure 9B shows the differential activity between the two chunks, which quantifies their competitive balance. GRAY's advantage over GREAT is maximal just as the input to the /g/ item ends. Once the /g/ item begins to decay, both chunks immediately begin to decay. The GRAY chunk decays faster, and thus progressively loses its competitive advantage until its activation falls below that of the GREAT chunk at approximately 260 ms. (The more rapid decay of the GRAY chunk is due to its weaker self-excitatory feedback via term $\phi f(u)z_u$ in Equation (A2) of Appendix A, since for a chunk j coding a list of N items, ϕ_j is proportional to N.)

Figure 10 shows how these effects extend to multiple items, again in the absence of top-down feedback. The inputs /g/, /r/, and /ei/ are presented as a sequence of pulses of constant magnitude and duration of 62 ms, so that the total duration of the sequence is 188 ms, which is the duration of the word "gray" in the Repp *et al.* (1978) experiments. As the working memory integrates the sequence of inputs, the differential activation between the GRAY and GREAT chunks increases, due to the input normalization and synaptic weights described above. As shown in Figure 10B, GRAY is able to maintain a competitive advantage over GREAT for a longer duration, nearly 300 ms, than with the single item input. The plot of transmitter activation (A, middle) shows that with all three items active, the GRAY chunk begins to consume trace amounts of its synaptic transmitter. Because chunks can self-excite more easily than they can send top-down feedback to their items, chunks can begin to consume their neurotransmitters prior to establishing a resonance with the working memory; see Equations (A2)-(A3) and accompanying text in Appendix A. The GRAY chunk shows a much stronger response to the input sequence than to a single input, since its entire complement of supporting items are active. However, without top-down feedback to support the working memory items, neither chunk is able to establish a full-fledged resonance.

6.2 Multiple item grouping and masking sensitivity

When top-down feedback is incorporated into network dynamics (via term ($\sum_{j \leftrightarrow i} \tau_{ij} u_j^+ z_{ju}$) in Equation (A1) of Appendix A), the GRAY chunk selectively enhances its active items in working memory and generates a resonant event. Figure 11A shows that the initial response of the network is identical to that of the open loop simulation in Figure 10. However, once the GRAY chunk exceeds its top-down threshold γ_u (c. 200 ms), both item and chunk trajectories undergo a resonant boost and begin to climb. The resonant event unfolds gradually over the next 100-200 ms. Items and chunks reach their maximal activations approximately 100 ms after the offset of the /ei/ input.

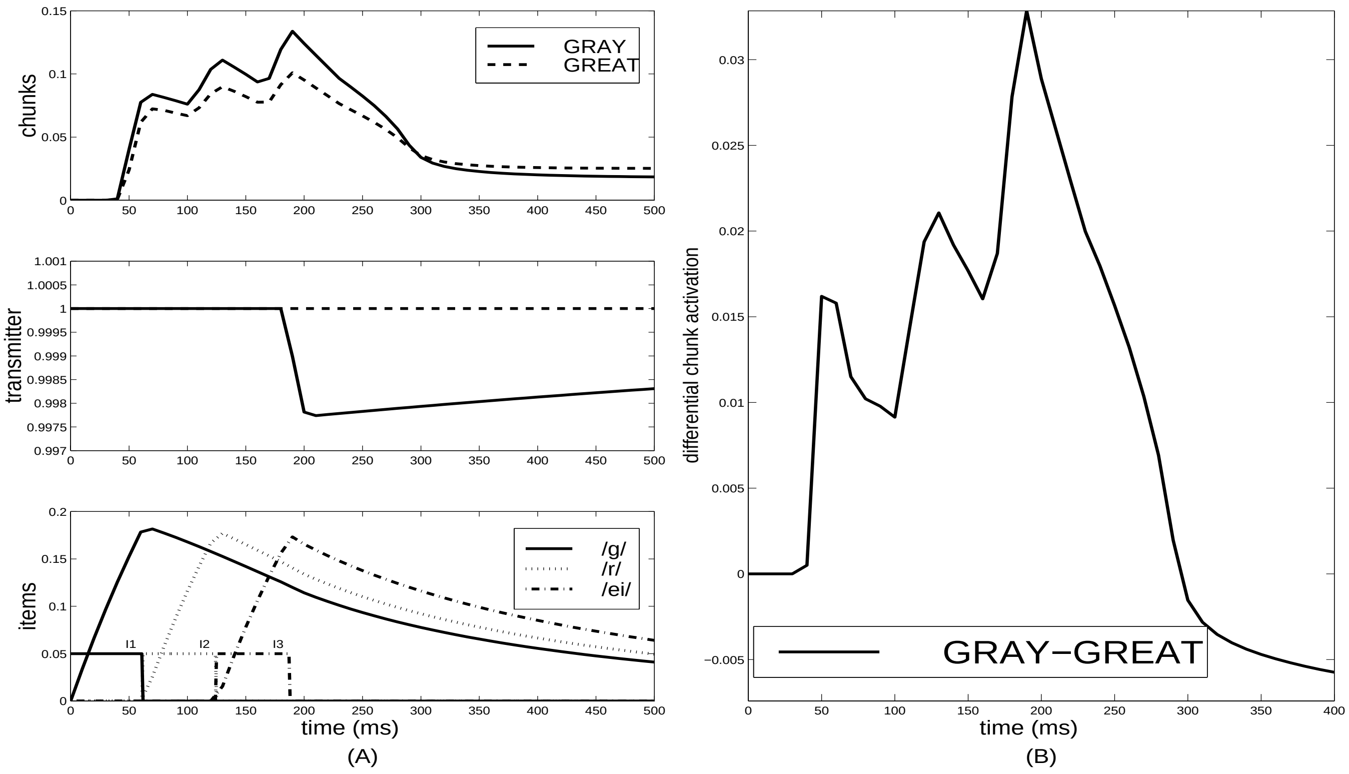


Figure 10: (A): Response of two chunks to a sequence of three item inputs (rectangular bars in lower left figure) in the absence of top-down feedback. (B): Differential activation of chunks.

That the GRAY chunk is fully resonating while the GREAT chunk remains in a subliminal state of activation can be observed from the tracing of transmitter activation in the middle panel. The sharp downwards inflection in the GRAY transmitter, which occurs at approximately 225 ms, indicates the onset of the positive feedback cycle. As the cycle continues, the GRAY chunk consumes transmitter more rapidly than it can be replenished until chunk activity peaks and begins to decay in a habituative collapse. As chunks and items passively decay, GRAY's transmitter slowly begins to replenish.

Figure 11B shows that when a /t/ input of comparable strength follows the /grei/ sequence immediately, it is able to push the GREAT chunk activation over its resonant threshold. The GRAY chunk begins its resonance while the /t/ item is being presented, at the same time as in Figure 11A. But once the /t/ item crosses its bottom-up threshold $\gamma_{/t/}$, it delivers a sustained excitation to the GREAT chunk of sufficient magnitude for the GREAT chunk to overcome GRAY's advantage and dominate the resonance. The resonance of GREAT is reflected in the single peak, at around 260 ms, of the working memory activation trajectories.

Figure 11A also shows that while the GREAT chunk cannot engage in resonance without the bottom-up input /t/, it does benefit from GRAY's top-down support of the /g/, /r/, and /ei/ items. Thus GREAT receives a subliminal boost from GRAY's resonance, priming the network to generate a grouping of the /t/ with the preceding items should it be presented. Such dynamics illustrate a critical aspect of masking sensitivity in the grouping network. Because the grouping network contains a bias towards longer lists by giving their chunks stronger masking parameters, the network design also needs to avoid a cascade of resonances wherein a smaller chunk, by supporting its own items, inadvertently pushes its competitor into a supraliminal state, and so on until the largest list present resonates with all of its items. Thus, the masking field implements larger chunk *potency* without a loss of chunk *selectivity*. In the present simulations, the larger chunk GREAT has a higher top-down feedback threshold ($\gamma_{GREAT} = 0.14 > \gamma_{GRAY} = 0.12$) – that is, needs more evidence to fire – so that even with the greater activation GREAT experiences during GRAY's resonance, GREAT remains below threshold. The subliminal priming of GREAT during GRAY's resonance also prepares the network for a transfer of resonant events between the two chunks in the event that /t/ does occur.

6.3 Resonant transfer

The third group of simulations, illustrated in Figures 12 and 13, shows how the grouping of an additional item with preceding items depends crucially on the temporal window during which it is activated. As a consequence of the competitive dynamics within the working memory, two input pulses with identical magnitude and duration will not be treated identically by the network if they arrive at different times in the processing cycle. Figures 12A and 12B show how a slight delay in the presentation of the /t/ input after the /g/, /r/, /ei/ sequence, relative to its presentation in Figure 11, can actually facilitate the resonance of the GREAT chunk over the GRAY chunk. This behavior mimics that of the Repp *et al.* (1978) data, which shows the apparently paradoxical effect at short noise durations that listeners are more likely to perceive "great" than "gray" when a longer silent interval separates the end of the vocalic segment /ei/ and the word initial fricative noise. In Figure 12A, the /t/ input arrives after a silent interval of 60 ms. During that interval, the GRAY chunk has initiated its resonant cycle with the /g/, /r/, and /ei/ items as evidenced by the depletion of the GRAY transmitter. The activation of the /t/ item in this instance is a case of "too little, too soon": because the /t/ item integrates to its maximal activity just as the activation of the GRAY

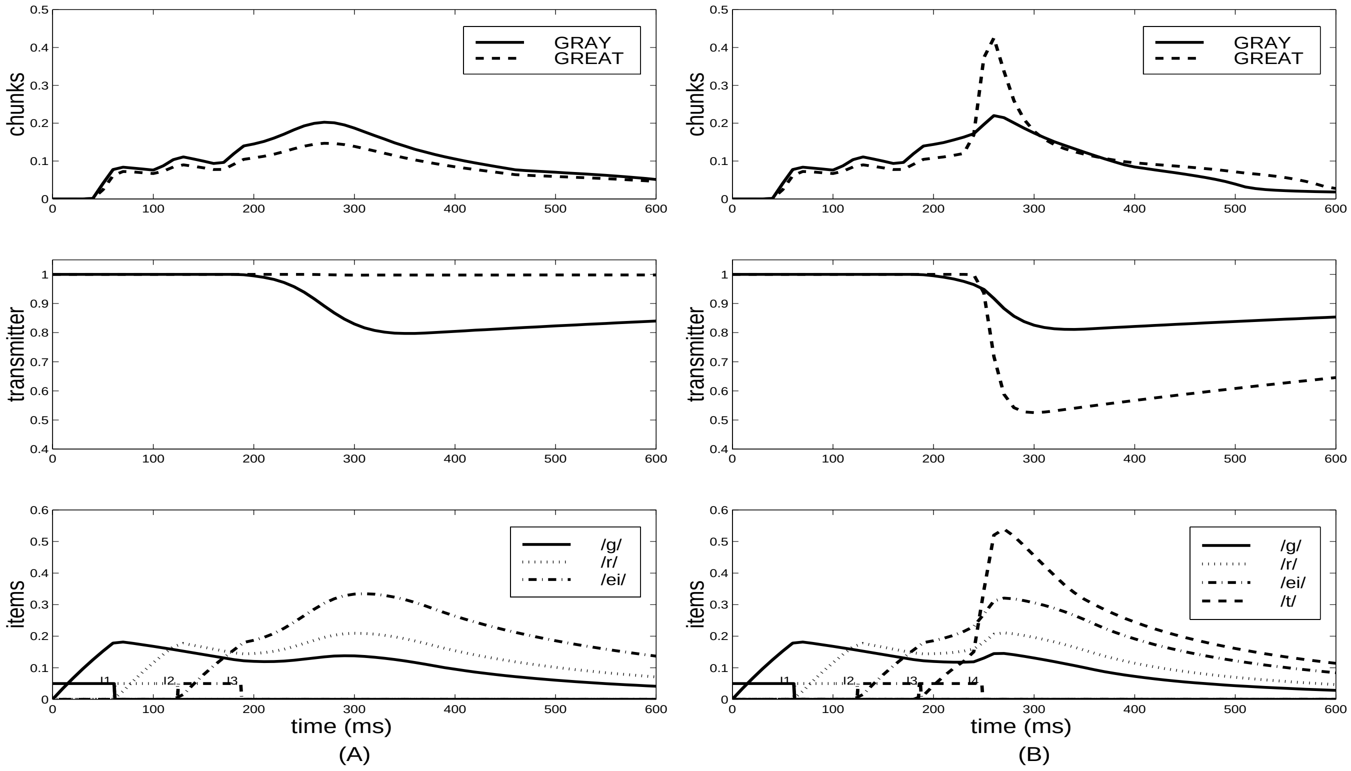


Figure 11: (A): Response of two chunks to a sequence of three (A) or four (B) items with top-down feedback.

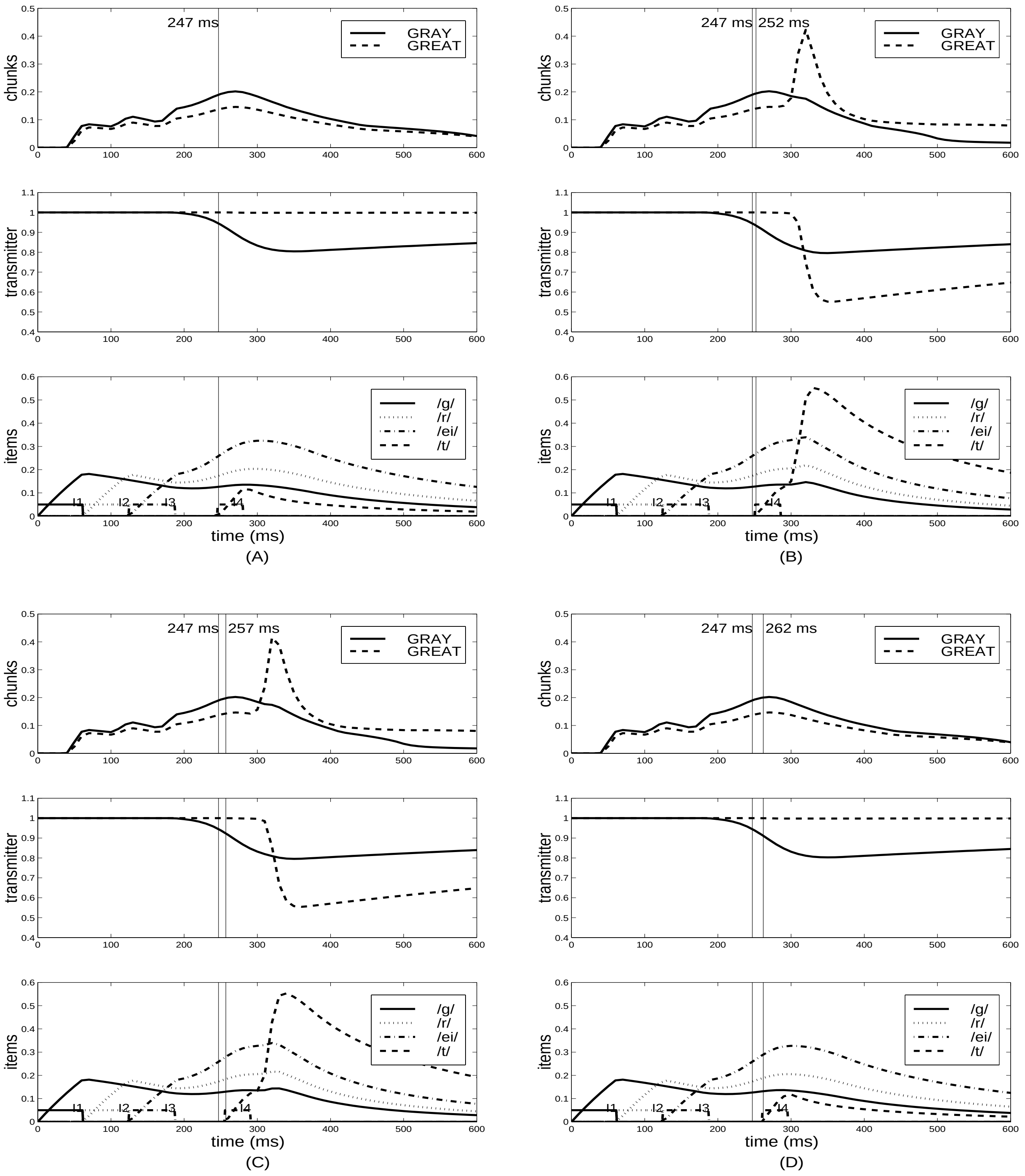


Figure 12: Transfer of resonance from GRAY to GREAT and back to GRAY with successively longer silent intervals between presentation of /ei/ and /t/ inputs. Silence duration = 60 ms (A), 65 ms (B), 70 ms (C), and 75 ms (D). Vertical lines indicate /t/ onset relative to panel (A), where onset occurs at 247 ms.

chunk peaks, GRAY is strongly inhibiting GREAT and, as a consequence of this inhibition, the /t/ item effectively passes undetected by GREAT.

A small additional delay in the presentation of the /t/ item can exert a profound effect on which chunk resonates, as shown in Figure 12B. By providing evidence which arrives to support the GREAT chunk *after* the GRAY chunk's activation has peaked, the /t/ item determines a qualitative change in how the competition in the grouping network unfolds. At this longer silence duration, GREAT can win its competition with GRAY through a resonant transfer. Because the end of the silent interval coincides with GRAY's habituative collapse, the network is primed to integrate the bottom-up activation of the /t/ item with the items that have been supported by GRAY's resonance. Thus, at relatively long silence durations, GREAT may win by piggy-backing on the previously supported /g/, /r/, and /ei/ items, and inhibiting the GRAY chunk whose neurotransmitters have become depressed. The process of resonant transfer thus explains why after being presented with the word "gray", followed by a silent interval of 100 ms in the Repp *et al.* (1978) experiments, the subsequent noise may be perceived as belonging to the word "great": the GRAY chunk has transferred its supported items to the GREAT chunk, by virtue of its habituative collapse. The transfer can be seen in Figure 12B in the trajectories of the chunks and their transmitter activation levels, which indicate that both chunks are able to resonate in a feedback cycle with their working memory items. The trajectories of the working memory items themselves (bottom panel, Fig. 12B) do not, however, reveal that two discrete resonant events have occurred. The network predicts that a listener under these conditions would not perceive the word "gray" followed by the word "great". Instead, from the perspective of the working memory, a single resonant event has developed, with the silence between /ei/ and /t/ enabling the coherent integration of the items into a single list.

The time window over which a subliminally activated chunk can integrate a subsequent item into a resonant event is limited. Thus, while the GREAT chunk can benefit from a delayed presentation of the /t/ input by competing with a weaker GRAY chunk, if the delay is too large, then the GREAT chunk itself will be too weak to achieve resonance. Figures 12C and 12D show that as the silent interval is extended from 70 ms (C) to 75 ms (D), the network undergoes a shift from GREAT's resonance back to GRAY's resonance. As in the simulations of Figures 12A and 12B, the significant determinant of the resonant grouping is the time at which the /t/ item becomes active *relative* to the developing competition between the GRAY and GREAT chunks. In the current simulations, the strength of the /t/ input and the gain Γ on the network integration rate are such that an 80 ms silent interval between activation of the /ei/ and /t/ items exceeds the window over which the GREAT chunk can group its chunks. Changes to many network parameters, either individually or jointly, can affect the precise duration of this integrative window. For example, a slower integration rate Γ will permit GREAT to resonate if longer delay intervenes. In the Repp *et al.* (1978) experiments, the GREAT chunk integrates over silent intervals in excess of 100 ms.

Figure 13 illustrates how resonant transfer depends on the relative timing and strength of the input items, and in particular how the silence duration can trade against the duration of the /t/ input to generate equivalent "great" percepts for different combinations of silence and noise. It shows the integrated GREAT chunk activation as the durations of /t/ input activation varies from 32-52 ms as a function of the duration of the intervening silence interval. Lighter shades represent less GREAT chunk activation, indicating that GRAY resonates with its items and a resonant transfer fails to occur; darker shades reveal that GRAY transfers its resonance to GREAT when the /t/ input is sufficiently strong. The diagonal curves dividing the light and dark regions show that as the

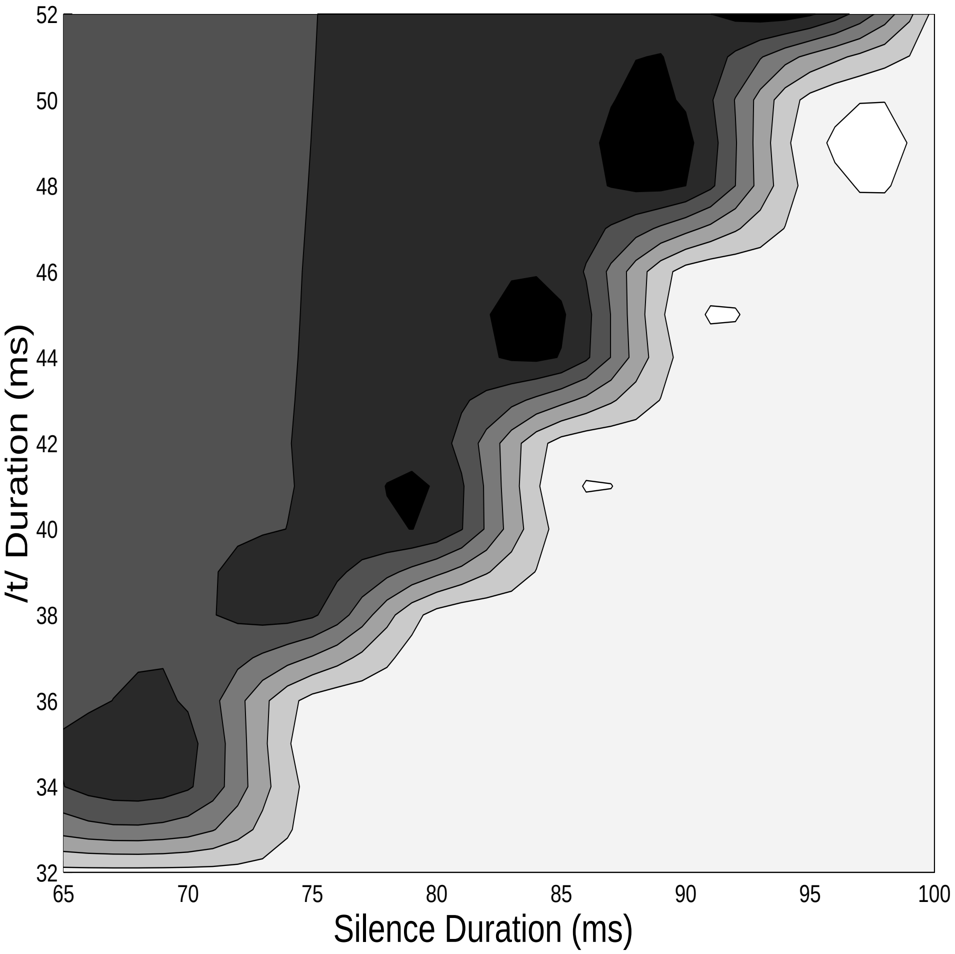


Figure 13: Trading relation between duration of the /t/ input and of the silence interval between /grei/ and /t/. Shading represents total GREAT chunk activation, with darker shades indicating greater activation (GREAT resonance).

silence duration increases, greater /t/ input is needed to excite the GREAT chunk above its feedback threshold and thereby facilitate a resonant transfer. Figure 13 thus illustrates how resonant transfer partially explains the trading relation between “gray chip” and “great ship” (cf. regions 2 and 3 in Figure 1). As noted by Repp *et al.* (1978, p. 631), the boundary function between these regions “shows a clear rise at intermediate silence durations (40-80 ms): GREAT SHIP responses were more frequent at short silence durations and GRAY CHIP responses were more frequent at longer silence durations.” That is, for a fixed duration of fricative noise, a longer silence interval produces a greater likelihood of perceiving “gray” instead of “great”. This occurs in Figure 13 because, through the acoustic-phonetic relations specified in Equation (A6), a longer fricative noise interval will deliver longer excitation to the /t/ phonemic item code, and thus generate a higher probability “great” percept. In a larger network, the competitive roles of the subsequent chunks CHIP and SHIP also function to alter the dynamics and the shape of the boundary between GRAY and GREAT resonances, as shown below.

The total GRAY chunk activation (not shown) behaves as the inverse of Figure 13; that is, when GREAT resonates, GRAY achieves less total activation due to the competitive inhibition from the GREAT chunk. The depression in total activation occurs despite the fact that the GRAY chunk reaches the same maximal activation (cf. Figures 12A and 12B), whether or not GREAT resonates. This suggests that total chunk activation over a specified time interval reflects the relative contrast between grouping patterns more robustly than simply the maximal chunk activation.

Figure 13 also demonstrates a nonlinear interaction between silence interval and input strength such that total chunk activation can actually reach greater values at longer silence intervals. In particular, the darkest shades, or greatest GREAT chunk activations, occur at silent intervals of 80-90 ms when the /t/ duration is just long enough to elicit a resonant transfer. This preference for /t/ inputs which are “strong enough, but not too strong”, provided they are of sufficient duration to drive their items above the bottom-up threshold γ_w, results from lateral inhibition in the working memory. When a given input is presented for a longer stimulus interval, its item inhibits the previously activated items more. The net result is to drive total item activity to a lower state, resulting in weaker support for the resonating chunk and a smaller total chunk activation. Thus a weaker input presented following a longer silence interval can, paradoxically, elicit a greater total chunk activation than a stronger input presented after a shorter silence interval; see, for example, coordinates (80,40) vs. (70,50) in Figure 13.

6.4 Competitive teaming

The preceding simulations illustrate that complex network dynamics can arise with only two chunks in the multiple item grouping network. The next group of simulations, shown in Figures 14A-14D, describe how the inclusion of additional chunks, encoding partially overlapping lists of items, adds a further dimension of complexity to the competition that develops in the grouping network. In these simulations, the grouping network consists of three chunks: GRAY, GREAT, and CHIP. Figure 14 shows that when the onset of the /t$\int$/ input coincides with the /t/ input, following the /g/, /r/, /ei/ sequence, the duration of the /t$\int$/ input relative to the /t/ duration determines whether or not GREAT will resonate. Because of shared sensitivity to high frequency spectral energy contained in the noise of the stop and affricate consonants “t” and “ch”, the GREAT and CHIP chunks compete with each other directly. Thus, if the CHIP chunk becomes sufficiently active, as in Figure 14B, it can prevent the GREAT chunk from resonating. Even though the CHIP chunk receives no input from the /I/ or /p/ items in the simulations of Figures 14A and 14B, the

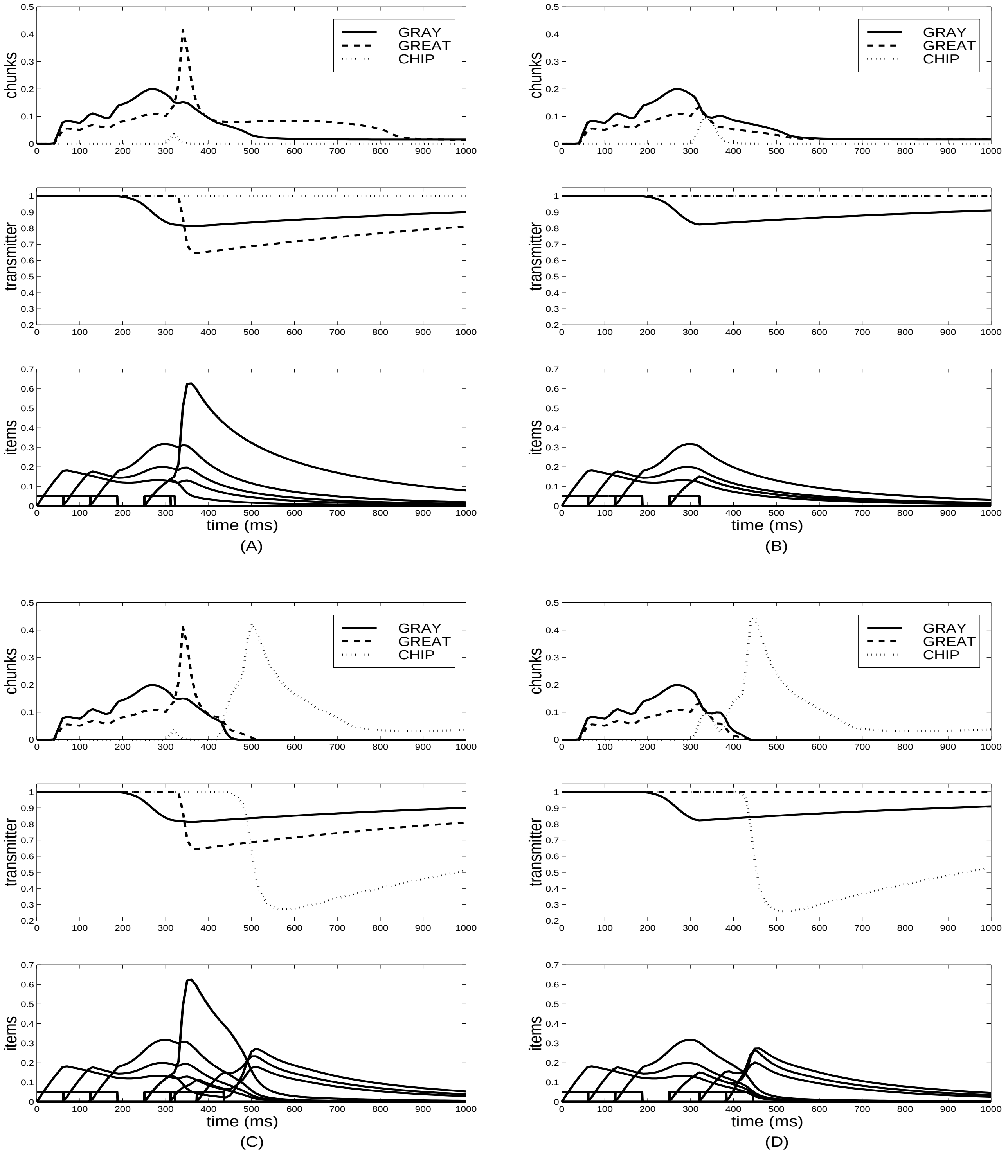


Figure 14: Dynamics of competitive teaming. Presentation of the /t∫/ input may not (A) or may (B) prevent GREAT from resonating via CHIP → GREAT inhibition. GREAT and CHIP resonances can coexist (C), or CHIP can prevent GREAT from resonating (D). A, C: /t∫/ input duration = 60 ms. B, D: /t∫/ input duration = 70 ms.

subliminal activation of the CHIP chunk by a /tʃ/ input 70 ms in duration inhibits the GREAT chunk sufficiently to prevent it from reaching its resonant threshold. A briefer /tʃ/ input of 60 ms duration (A), by contrast, can produce a small activation of the CHIP chunk without interfering in the ability of the GREAT chunk to resonate. Figure 14B thus illustrates the network principle of competitive teaming by which one chunk's resonance is prevented by conjoint activation of multiple competitors.

The consequences of competitive teaming are further illustrated in Figures 14C and 14D, which are identical to the simulations of Figures 14A and 14B except that the /I/ and /p/ items are presented following the /tʃ/ item. In Figure 14C (/tʃ/ duration=60 ms), the network first undergoes a resonant transfer from GRAY to GREAT, as the /t/ and /tʃ/ items become active following the presentation of the /g/, /r/, /ei/ sequence. As in Figure 12, this resonant transfer results in a single grouping event in the working memory indicated by the resonant boost at approximately 350 ms. However, the subsequent presentation of the /I/ and /p/ are able to build on the residual activity of the /tʃ/ item in the working memory and elicit a CHIP resonance. The CHIP resonance defines a second distinct resonant event in the working memory that corresponds to the activation boost at approximately 520 ms. Because the /tʃ/ item remains weakly active during GREAT's resonance, both GREAT and CHIP can resonate in sequence with their working memory items. By creating two distinct resonances under these conditions, the network illustrates how a single noise interval, exciting both /t/ and /tʃ/ item codes in working memory, can be grouped both backwards in time with GREAT and forwards in time with CHIP, as in the "great chip" percepts of the Repp *et al.* (1978) experiments. Figure 14D, by contrast, shows that a relatively stronger /tʃ/ input occurring after an identical preceding silent interval will result in the sequential resonances of GRAY and CHIP, resulting in the "gray chip" percept that occurs in the Repp *et al.* data at intermediate silence durations and brief noise durations. The conditions which favor the formation of the "gray chip" percept, then, include /tʃ/ item activation strong relative to /t/ item activation, and the subsequent competitive teaming of the CHIP and GRAY chunks to inhibit the GREAT chunk.

7. Simulations of the Repp et al. (1988) Data

The simulations above illustrate the key dynamic processes that allow the ARTWORD model to successfully simulate the perceptual data of the Repp *et al.* experiment. Multiple-item grouping with resonant feedback, resonant transfer across silence intervals, and the competitive teaming of overlapping chunks, together define system dynamics that describe the perceived phonemic groupings as a function of inter-word silence and syllable-initial fricative noise.

7.1 Method

To simulate the Repp *et al.* (1978) data, the ARTWORD network described above was constructed with 8 phonemic item codes in the working memory (/g/, /r/, /ei/, /t/, /tʃ/, /ʃ/, /I/, and /p/) and 4 chunks in the grouping network (GRAY, GREAT, CHIP, and SHIP). All network parameters were set to fixed values (see Appendix B). Input pulses of fixed magnitude were presented to the working memory, and item, chunk, and transmitter activities were integrated. All items had fixed durations of 62 ms, except /t/, /tʃ/, and /ʃ/, whose durations depended on the durations of the silence and fricative noise intervals. The durations of these items were determined as described in Equations (A6) to (A8) in Appendix A. As in the Repp *et al.* (1978) experiment, silence duration varied from 0 to 100 ms in 10 ms steps and noise duration varied from 62 to 182 ms in 40 ms steps, producing 44 combinations of silence and noise durations. For each of the 44 combinations, the corresponding input schedule was determined and presented to generate all network trajectories for items (w_i), list chunks (u_j), item-to-list chunk transmitters (z_{iw}), and list chunk-to-item transmitters (z_{ju}). Dynamical equations for all of these variables are given in Appendix A.

7.2 Mapping network activations to response probabilities

Once network activations were determined, chunk activations were integrated and mapped to single word response probabilities, in accord with the four alternative forced choice task of the Repp *et al.* (1978) subjects. Chunk activities were defined as the integrated activity from list onset to 200 ms after list offset, a window which encompassed the resonant responses of all chunks. To determine the probability of a "gray" response, a decision variable D_{GRAY} was formed from the activation of the GRAY chunk relative to the combined activation of the GRAY and GREAT chunks (Luce, 1959), and likewise D_{CHIP} was constructed from the integrated activation of the CHIP chunk relative to the combined activation of the CHIP and SHIP chunks. In the following four equations, we denote the temporal limits of integration by writing "/x/ on" to indicate the onset of the first phoneme of a given chunk and "/x/ off + 200" to indicate the time point 200 ms after the offset of the last phoneme of a given chunk, where /x/ is the first or last phoneme. Letting u_j be the activity of list chunk j (see Appendix A for its equation), we define

$$U_{GRAY} = \int_{\text{/g/ on}}^{\text{/ei/ off+200}} u_{GRAY}(t)dt, \tag{1}$$

$$U_{GREAT} = \int_{\text{/g/ on}}^{\text{/t/ off + 200}} u_{GREAT}(t)dt, \tag{2}$$

$$U_{CHIP} = \int_{/t\int\text{/on}}^{\text{/p/ off + 200}} u_{CHIP}(t)dt, \tag{3}$$

and

$$U_{SHIP} = \int_{/\int\text{/on}}^{\text{/p/ off + 200}} u_{SHIP}(t)dt, \tag{4}$$

from which we further define

$$D_{GRAY} = \frac{U_{GRAY}}{U_{GRAY} + U_{GREAT}} \tag{5}$$

and

$$D_{CHIP} = \frac{U_{CHIP}}{U_{CHIP} + U_{SHIP}}. \tag{6}$$

To map the decision variables to response probabilities, each was linearly rescaled and perturbed by Gaussian noise of fixed mean and unit variance (Green & Swets, 1974). That is, letting Φ represent a cumulative normal distribution with zero mean and unit variance, the final response probabilities were computed as

$$P(GRAY) \quad = \quad \theta_1 + \theta_2 \Phi(\theta_3 D_{GRAY} + \theta_4) \tag{7}$$

and

$$P(CHIP) \quad = \quad \theta_5 + \theta_6 \Phi(\theta_7 D_{CHIP} + \theta_8). \tag{8}$$

By construction, the complementary probabilities are $P(GREAT) = 1 - P(GRAY)$ and $P(SHIP) = 1 - P(CHIP)$. The free parameters θ_i were chosen to maximize the log likelihood of the predicted values with respect to the data. Thus 8 free parameters were chosen to fit the integrated network responses to the 88 data points (44 “gray” response probabilities and 44 “chip” response probabilities). Maximization was performed with the Nelder-Mead simplex search, run for 500 iterations (Press, Flannery, Teukolsky, & Vetterling, 1988).

7.3 Simulation results

The computer simulations summarized in Figure 15 show that ARTWORD closely approximates the perceptual data averaged over 10 subjects in the Repp *et al.* (1978) experiments. All of the major trends shown in the reported psychometric data are replicated by ARTWORD. The ARTWORD model globally accounts for 91% of the variance of the single word response probabilities. The probability of either a “gray” response or a “chip” response decreases with longer noise intervals. Figure 15B shows that “chip” responses increase monotonically with increasing silence intervals. Figure 15A shows, as in the data, that the likelihood of a “gray” response increases with increasing silence, for longer noise intervals (102-182 ms). Under these conditions, the psychometric functions for “gray” are non-monotonic. In ARTWORD, at the longer silence durations, the CHIP chunk can more effectively inhibit the GREAT chunk, and so, via competitive teaming, the GRAY chunk attains a relatively greater proportion of the total activation. Thus, when the decision variable is added to Gaussian noise, it is more likely to yield a “gray” response at longer silence durations.

Figure 16 shows the category boundaries derived from the response probabilities plotted in Figure 15. As described above, to derive the boundaries the probability surface defined by the curves in Figure 16 was interpolated with a cubic polynomial in 1 ms steps on a grid spanning silence durations between 0 and 100 ms and noise durations between 62 and 182 ms. For each word pair (gray/great, and chip/ship), the contour of 50% probability was determined and plotted. Figures 17A-D show the category boundaries derived from the data and the model predictions in more detail. Figures 17A-D also include the 60% and 40% probability contours, which give a measure of the uncertainty associated with the perceptual boundaries between the response regions. The data and model show similar certainty regions, or confidence intervals, for the different parametric combinations of silence and noise.

In Figures 17A-B, the 40%-60% GREAT response region is extremely tight for noise durations greater than 120 ms, indicating a steeply sloped decision function. As silence duration increases above 30 ms, the perceptual contours broaden, indicating psychometric functions with shallower slopes, or greater uncertainty. In both the ARTWORD predictions and the reported data, the contours show a tendency to flare outwards at the greatest silence durations tested, showing that the decision between “gray chip” and “great chip” is uncertain. While the ARTWORD 50% boundary exits to the right (i.e. towards longer silence durations), the boundary interpolated from the Repp

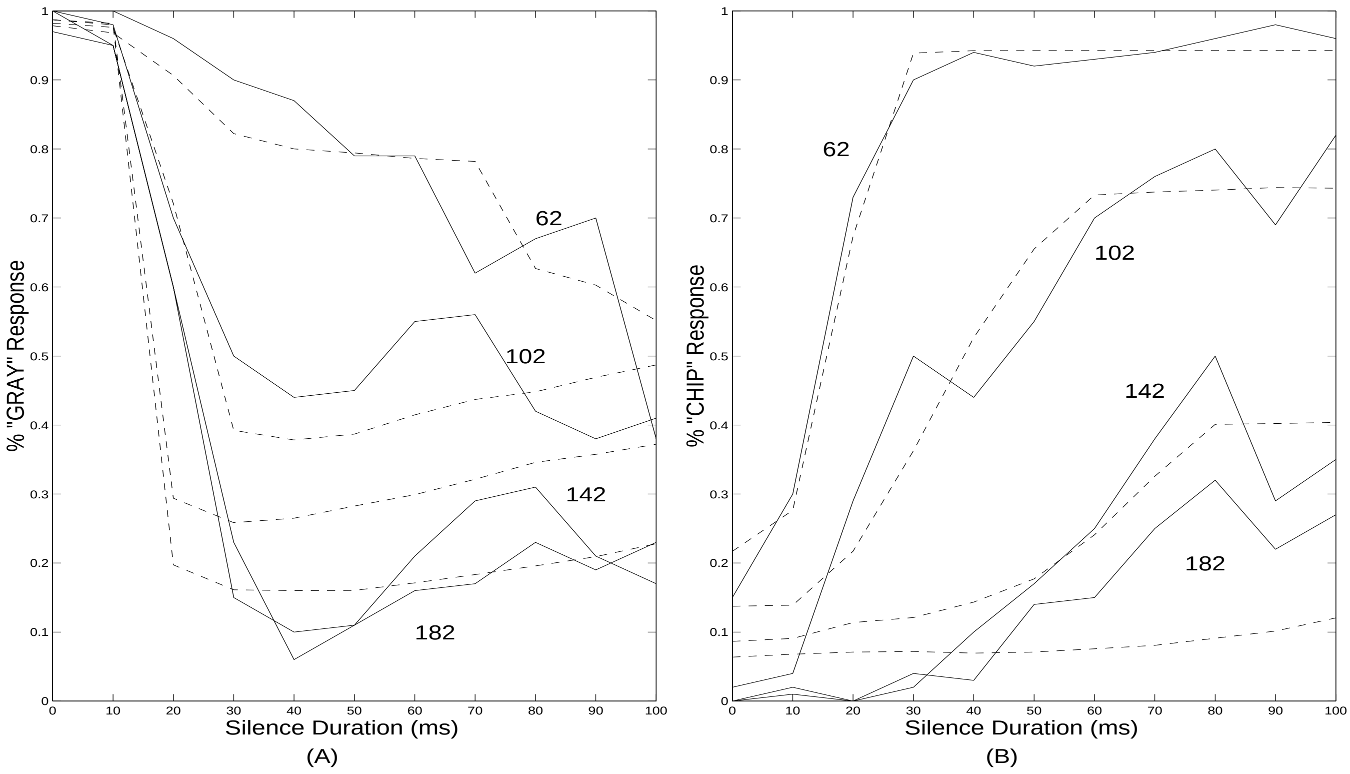


Figure 15: Probabilities of responding GRAY (A) or CHIP (B). Data i nsolid lines, ARTWORD model predictions in dashed lines. Numbers indicate duration of fricative noise interval.

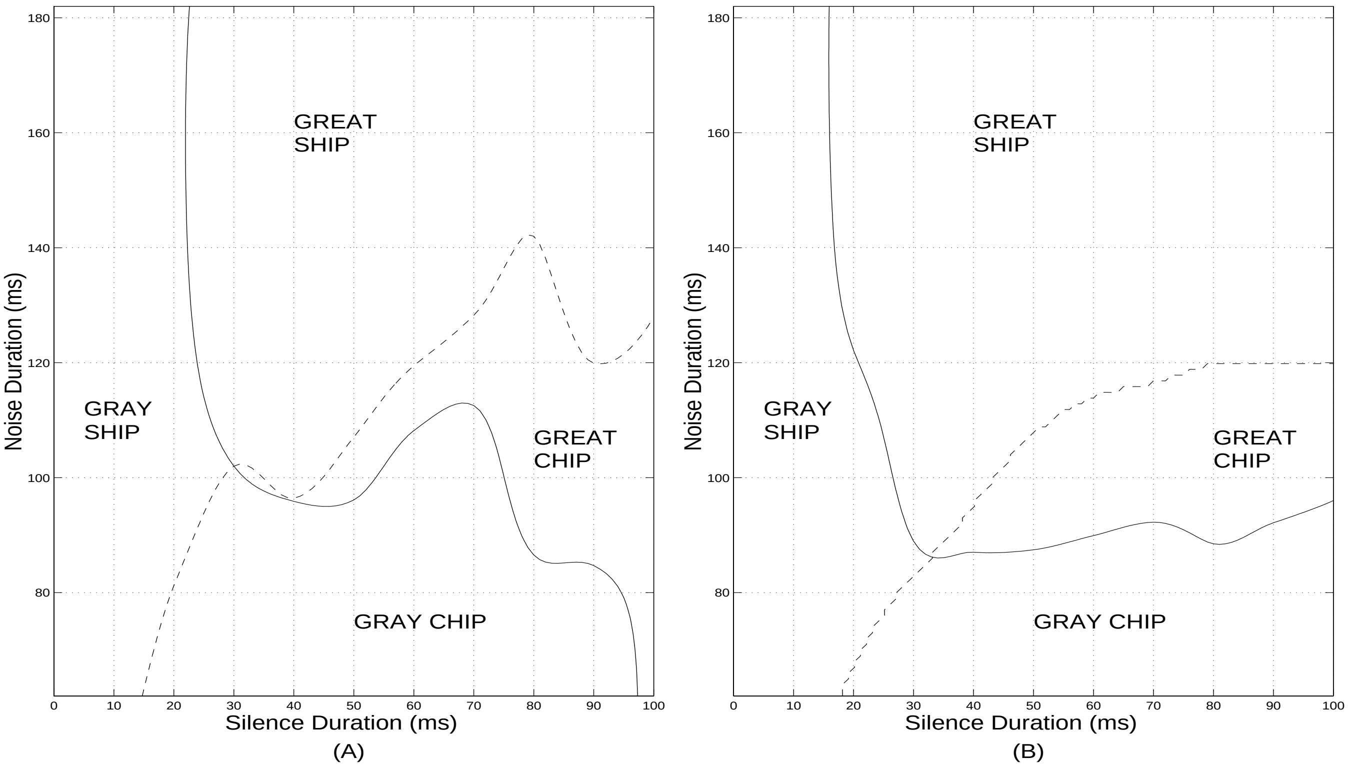
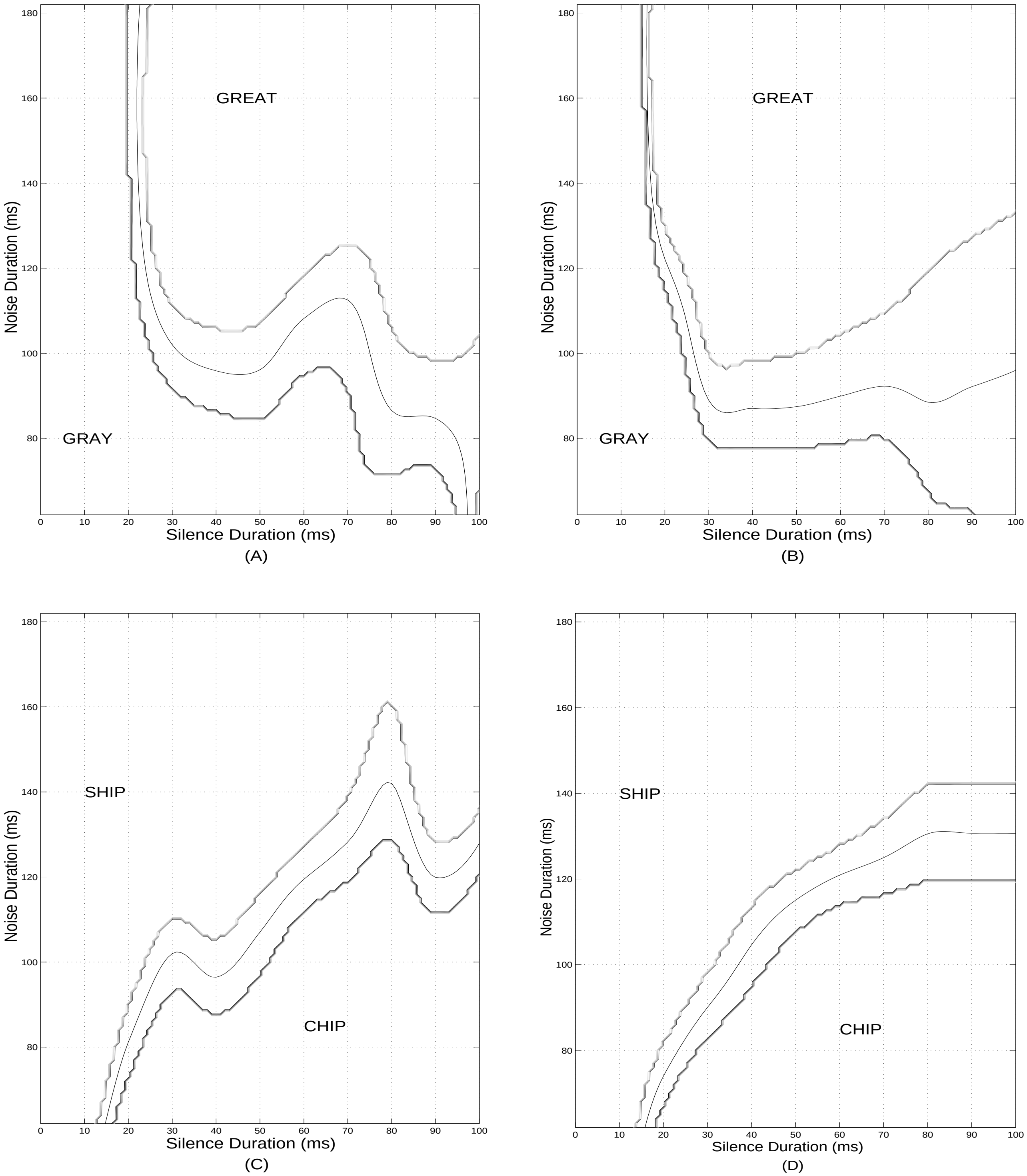


Figure 16: Derived two-word category boundaries. (A): Repp *et al.* (1978) data. (B): ARTWORD predictions.



(A)

(B)

(C)

(D)

Figure 17: Category boundaries derived from the Repp *et al.* (1978) data and from the ARTWORD model predictions. (A): GRAY–GREAT, data. (B): GRAY–GREAT, ARTWORD. (C): CHIP–SHIP, data. (D): CHIP–SHIP, ARTWORD.

et al. (1978) data exits downwards (i.e., towards shorter noise durations). Observing the 40% GREAT decision contours, however, shows that both the model and data show a similar increase in “great” responses at low noise durations at the longest silence durations. The deviation of the ARTWORD model’s predicted 50% boundary at the longest silence durations appears to be due to the shallower slope of the gray-great decision functions at longer noise durations; that is, the ARTWORD model assigns too high a probability to a GRAY resonance in this region.

As noted above, the model boundaries result from systemwide interactions and can be altered by varying parameters. The model boundary in Figure 17B could, for example, be driven downwards in by an input representation that allocated less input to the /t/ item at these longest silence durations. However, without more data to inform the quantitative nature of the acoustic-phonetic mapping between inputs and phonemic item activations, precise determinations of the input representation achieved by the auditory system at the level of the working memory must be deferred. Quantitative exploration of the perceptual space by varying network parameters such as integration rate and chunk thresholds, however, does suggest further perceptual experiments to determine which network processes account for the variations between the ARTWORD boundaries and the boundaries derived from the data.

Figures 17C-D show that, as in the data, the predicted “chip-ship” decision boundary becomes less steep at increasing silence durations. Both the predicted and actual boundaries arc through the same parametric region of silence and noise durations. The reported data generates upwards swerves in both the “gray-great” boundary (60-70 ms silence durations) and the “chip-ship” boundary (30 ms and 80 ms silence durations) which are not apparent in the ARTWORD boundaries. However, without knowledge of the individual responses in the Repp *et al.* (1978) data, it is difficult to assign functionally meaningful interpretation to these swerves. In particular, it is unclear whether these deviations result from systematic competition between co-active lexical representations or merely reflect differing decision thresholds across subjects.

8. Relation of ARTWORD to models of lexical segmentation

ARTWORD was developed primarily to show that the dynamics of resonance can account for the cognitive processes underlying the perceptual integration of phonemic information during conscious speech perception. As a cognitive model of speech perception, ARTWORD bears interesting relationships to several models in the related domain of lexical segmentation. Models of lexical segmentation, driven primarily by psycholinguistic research and by computational analyses of word embeddings in large vocabulary corpora, have converged on strategies that, like ARTWORD, permit the gradual activation of candidate groupings which best match the arriving input stream (see reviews in, e.g., Altmann, 1990; Pisoni and Luce, 1987; and Miller and Eimas, 1995). Three models in particular — Cohort (Marslen-Wilson, 1987), TRACE (Elman & McClelland, 1986), and Shortlist (Norris, 1994) — are interesting in light of the similarity of some of the functional processes they propose. Like TRACE and Cohort, the ARTWORD model explains lexical segmentation on the basis of bottom-up and top-down information flow, and, like all three models, ARTWORD uses some form of competition among candidates. The ARTWORD model shares the quantitative specificity of TRACE and Shortlist while incorporating a number of conceptually attractive features not present in these models, including perceptual resonance, category collapse, and a real-time processing framework that allows it to capture the complex perceptual effects caused by variation of segmental durations in the Repp *et al.* (1978) data.

The Shortlist model has some similarities to processes used in ART networks, although it omits the key ART process of top-down information flow. Shortlist uses “bottom-up mismatch information to penalise mismatching candidate words very strongly” (McQueen *et al.*, 1995, p. 325). This strategy resembles the mismatch reset that occurs in ART networks when bottom-up input to the working memory differs substantially from the expected pattern being read out through long-term memory traces from the active lexical hypotheses in the masking field. Like TRACE, but unlike Cohort, the Shortlist model uses lateral inhibition between active word candidates to decide the competition between them. Cohort, instead, postulates that the activity levels of the candidates do not influence each other, but rather that a higher-level decision mechanism determines the outcome of the competition between them. Despite the difficulty of testing between these hypotheses, McQueen *et al.* (1995) present statistics on lexical embedding and experimental results arguing for the competition between active candidates. In particular, any other decision mechanism must show a number of sensitivities simply accounted for by a lateral inhibitory mechanism; for example, “that the activation of each candidate is sensitive to the impact which that candidate has on the interpretation of both that part and other parts of the utterance”, and an ability “to weigh up each candidate with respect not just to that candidate’s fit to the part of the input with which it is aligned, but also with respect to how that candidate fits with other candidates, spanning other parts of the input” (McQueen *et al.*, 1995, p. 327). Because the competition in ARTWORD is based on the lateral inhibitory connections between unitized representations which continuously integrate the available bottom-up phonemic input, the ARTWORD model shows exactly these sensitivities.

All of these models have provided informative accounts of aspects of lexical segmentation. However, it is difficult to see how these other models would explain grouping data like those of Repp *et al.* (1978). One principle limitation of these models is the absence of a natural reset mechanism which would allow simultaneous competitions to influence sequentially activated and reset word representations. It is also unclear how silence intervals function in the above theories and whether they could contribute evidence for particular groupings of phonemic items by delaying subsequent activations. For example, in TRACE, a hand-coded silence feature inhibits active word representations, and thus silence acts as a fixed, wordlike competitor. In ARTWORD, by contrast, silence is perceived when a temporal break occurs in the rate of resonance. It is an emergent property, not a fixed network feature. The property of resonant transfer can create a fusion event between list chunks only when a delayed item arrives as the first resonance weakens due to reset. Resonant transfer thus requires both ART reset mechanisms and a real-time treatment of silence. In turn, the ARTWORD model can naturally generate trading relations between acoustic cues, including silence, that are problematic for models like TRACE. Other problems faced by the TRACE, MERGE, Shortlist, Interactive Activation, Fuzzy Logical Model of Perception, and related models are discussed in Grossberg (1999a) and Grossberg *et al.* (1997).

9. Discussion: Resonant dynamics and silence in speech perception

The present article has described the ARTWORD neural network model of perceptual integration in speech perception, which quantitatively extends earlier ART-based speech models to allow multiple-item grouping of phonemes into word level representations (Grossberg, 1978a, 1986; Cohen & Grossberg, 1986; Grossberg et al., 1997). The ARTWORD model posits that the grouping process involves bottom-up activation of word chunks which feed back and support their phonemic items. The top-down support of phonemic items, in turn, leads to the dynamic emergence of a resonant event. As inputs stream into the working memory, shifting the evidence for competing chunks, the resonant wave spreads to different phonemes, thereby creating a shifting attentional focus. Categorization and grouping of phonemic inputs is shown to depend both explicitly on phonemic activation strength and implicitly on the durations, or local rates, of input presentation. In particular, silence intervals can play a crucial role in the transfer of perceptual resonance between actively competing candidates.

The ARTWORD model processes of item integration, chunk competition, and resonance also illustrate how later-occurring information can influence the formation of earlier percepts. Again, the duration of silence in the speech stream determines key aspects of these backwards effects. For example, at longer silence durations, ARTWORD remains able to generate "great" groupings because the habituative collapse of the GRAY chunk leads to segmentation of the fricative noise "sh" that supports resonances with both GREAT and CHIP. Analyses of the problems posed by lexical segmentation are beginning to recognize how segmental durations in the speech stream can have profound effects on processing and necessitate the kind of limited temporal integration windows that emerge from resonant dynamics (see, e.g., Newman & Sawusch, 1996). As Mattys noted, "the literature...suggests that segmentation problems dictate how lexical processing unfolds in time. Sequential processing, which for a long time was considered a natural and universal principle, can no longer be viewed as the only mechanism during speech processing...both proactive and retroactive mechanisms seem to be necessary to parse the input successfully" (Mattys, 1997, p. 324).

The resonant dynamics of the ARTWORD system highlight the significance of time itself as a dimension in grouping and generating perceived segmentations. The role of silence and noise *durations*, as distinct from their influence on phonemic item responses, in determining the perceived identity of phonemic and lexical units demonstrates the importance of ongoing temporal integration to the perceptual speech code. While the acoustic cues carried by spectral features are themselves dependent on the temporal aspects of the speech stream, speech research recognizes the importance of the temporal dimension of information. Rosen (1992, pp. 74–75) described how temporal *envelope information*, or "fluctuations in overall amplitude at rates between about 2 and 50 Hz", contribute strongly to perception of manner (e.g., /ʃ/ vs. /tʃ/ rise-times), tempo, rhythm, stress, and syllabification or juncture. Despite the fact that segmental durations have always occupied a prominent role in acoustic and phonetic investigations of vowel and consonant perception (Bastian *et al.*, 1961; Repp *et al.*, 1978; Dorman *et al.*, 1979), psycholinguistic studies of word recognition have only recently begun to take into account time itself as a significant processing dimension (Mattys, 1997).

Many lines of evidence support the view of top-down interactions illustrated in the ARTWORD model whereby higher-level representations (e.g., phrases and words) guide phonemic processing. These interactions are conceptually consistent with both the phonemic restoration data reviewed above and data from the lexical identification shift, which shows that phonemic perception along a stimulus continuum (e.g., /g/-/k/) presented in word-nonword context (e.g., /gift/-/kift/) is biased towards the word (Ganong, 1980; McQueen, 1991; Gordon, Eberhardt, & Rueckl, 1993). Attention has also been shown to modulate the processing of phonemic cues. Gordon *et al.* (1993) showed that distractor tasks differentially affect subjects' perceptions of phonemic distinctions. For example, formant pattern and vowel duration help distinguish the vowels /i/ (in "beat") and /I/ (in "bit"). Attentional demands decrease the relative importance of formant pattern and increase the relative importance of duration. Such top-down grouping effects also occur in the visual processing of lexical items; e.g., in the *word length effect*: letters are perceived more readily when they are embedded in longer words, up to a certain length (Samuel et al., 1982, 1983), which was predicted by ART (Grossberg, 1978a). Available data thus make it clear that top-down interactions in the form of attention, semantic and syntactic context, and lexical and phonemic status, all play a role in shaping acoustic information such as segment durations, formant transitions, and speech-rate estimates into perceived linguistic units.

The ARTWORD model elaborated herein provides a further illustration of how resonant dynamics can explain diverse auditory perceptual events, including auditory stream formation (Govindarajan *et al.*, 1994; Grossberg, 1999c) consonantal geminate-cluster distinctions (Grossberg *et al.*, 1997), and other qualitative aspects of speech perception (Cohen *et al.*, 1988). The further development of these ART models of resonant interactions can proceed along several fronts. Within the domains of audition and speech perception, a prospect for future research concerns the integration of lower-level pitch perception (Cohen, Grossberg, & Wyse, 1995) and phoneme processing networks (Boardman et al., 1999; Cohen & Grossberg, 1997) with higher-level speech and streaming networks. More generally, the ubiquity of resonant events, and in particular their dynamic sensitivity to temporal variations in input, suggests that they reflect universal principles of adaptive sensory processing. As reviewed by Grossberg (1995, 1999d), a growing body of evidence suggests that the ART mechanisms which underly the dynamic formation of resonant events in speech are also common in other attentive brain systems. As specific ART-based networks are developed to explain even more data, the existence of shared dynamic processing mechanisms can further clarify and integrate our understanding of brain function.

AUTHOR NOTE

Stephen Grossberg was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225), the Defense Advanced Research Projects Agency and the Office of Naval Research (ONR N00014-95-1-0409), the National Science Foundation (NSF IRI-97-20333), and the Office of Naval Research (ONR N00014-92-J-1309 and ONR N00014-95-1-0657). Christopher Myers was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225), the Defense Advanced Research Projects Agency and the Office of Naval Research (ONR N00014-95-1-0409), and the Office of Naval Research (ONR N00014-91-J-4100, ONR N00014-92-J-1309, ONR N00014-94-1-0940, ONR N00014-94-1-0597, and ONR N00014-95-1-0657). The authors thank Robin Amos,Cynthia Bradford, and Diana Meyers for their valuable assistance in the preparation of the manuscript. Correspondence concerning this article shold be addressed to Stephen Grossberg,Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, Massachusetts 02215. Electronic mail may be sent to steve@cns.bu.edu.

Reference

Abbott, L. F., Varela, K., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science*, *275*, 220–223.

Altmann, G. T. M. (1990). *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. MIT Press, Cambridge, MA.

Anderson, S., & Port, R. (1994). Evidence for syllable structure, stress, and juncture from segmental durations. *Journal of Phonetics*, *22*, 283–315.

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(3), 536–563.

Bard, E. G. (1990). Competition, lateral inhibition, and frequency: Comments on the chapters of Frauenfelder and Peeters, Marslen-Wilson, and others. In Altmann, G. T. M. (Ed.), *Cognitive Models of Speech Processing*, chap. 9, pp. 185–210. MIT Press, Cambridge, MA.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1989). The recognition of words after their acoustic offset: Effect of subsequent context. *Perception and Psychophysics*, *44*, 395–408.

Bashford, J. A., Meyers, M. D., Brubaker, B. S., & Warren, R. M. (1988). Illusory continuity of interrupted speech: Speech rate determines durational limits. *Journal of the Acoustical Society of America*, *84*(5), 1635–1638.

Bashford, J. A., Riener, K. R., & Warren, R. M. (1992). Increasing the intelligibility of speech through multiple phonemic restorations. *Perception and Psychophysics*, *51*(3), 211–217.

Bastian, J., Eimas, P., & Liberman, A. M. (1961). Identification and discrimination of a phonemic contrast induced by silent interval. *Journal of the Acoustical Society of America*, *33*, 842.

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception and Psychophysics*, *29*, 191–211.

Boardman, I., Grossberg, S., Myers, C. W., & Cohen, M. (1999). Neural dynamics of perceptual order and context effects for variable-rate speech syllables. *Perception and Psychophysics*. In press.

Bradski, G., Carpenter, G. A., & Grossberg, S. (1994). STORE working memory networks for storage and recall of arbitrary temporal sequences. *Biological Cybernetics*, *71*(6), 469–480.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Bradford Books/MIT Press, Cambridge, MA.

Carpenter, G. A., & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks*. MIT Press, Cambridge, MA.

Casseday, J., Ehrlich, D., & Covey, E. (1994). Neural tuning for sound duration: role of inhibitory mechanisms in the inferior colliculus. *Science*, *264*, 847–850.

Chey, J., Grossberg, S., & Mingolla, E. (1997). Neural dynamics of motion grouping: from aperture ambiguity to object speed and direction. *Journal of the Optical Society of America*, *10*, 2570–2594.

Christie, W. M. (1974). Some cues for syllable juncture perception in English. *Journal of the Acoustical Society of America*, *55*(4), 819–821.

Cohen, M. A., & Grossberg, S. (1986). Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, *5*, 1–22.

Cohen, M. A., & Grossberg, S. (1987). Masking fields: A massively parallel neural architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics*, *26*, 1866–1891.

Cohen, M. A., & Grossberg, S. (1997). Parallel auditory filtering by sustained and transient channels separates coarticulated vowels and consonants. *IEEE Transactions on Speech and Audio Processing*, *5*(4), 301–318.

Cohen, M. A., Grossberg, S., & Stork, D. G. (1988). Speech perception and production by a self-organizing neural network. In Lee, Y. C. (Ed.), *Evolution, Learning, Cognition, and Advanced Architectures*, pp. 217–231. World Scientific, Singapore.

Cohen, M. A., Grossberg, S., & Wyse, L. (1995). A spectral network model of pitch perception. *Journal of the Acoustical Society of America*, *98*(2), 862–879.

Crystal, T. H., & House, A. S. (1988a). A note on the durations of fricatives in American English. *Journal of the Acoustical Society of America*, *84*(5), 1931–1935.

Crystal, T. H., & House, A. S. (1988b). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, *83*(4), 1553–1573.

Cutler, A. (1990). Psychology and the segment. In Kingston, J., & Beckman, M. E. (Eds.), *Between the grammar and physics of speech*, Vol. 1 of *Papers in laboratory phonology*, chap. 11, pp. 290–295. Cambridge University Press, Cambridge.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133–142.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 113–121.

Darwin, C. J. (1976). The perception of speech. In Carterette, E. C., & Friedman, M. P. (Eds.), *Handbook of Perception: Language and Speech*, Vol. VII, chap. 6, pp. 175–226. Academic Press, New York.

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. *Journal of the Acoustical Society of America*, *68*(3), 843–857.

Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In Carlson, R., & Ganstrom, B. (Eds.), *The Representation of Speech in the Peripheral Auditory Nerve*, pp. 131–149. Elsevier/North Holland, Amsterdam.

Delgutte, B., & Kiang, N. Y. (1984). Speech coding in the auditory nerve: III. Voiceless fricative consonants. *Journal of the Acoustical Society of America*, *75*(3), 887–896.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, *65*(6), 1518–1532.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In Perkell, J. S., & Klatt, D. H. (Eds.), *Invariance and Variability in Speech Processes*, pp. 360–385. Erlbaum, Hillsdale, NJ.

Faulkner, A., Rosen, S., Darling, A. M., & Huckvale, M. (1995). Modelling cue interaction in the perception of the voiceless fricative/affricate contrast. *Language and Cognitive Processes*, *10*(3/4), 369–375.

Fitch, H. L., Hawles, T. G., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop consonant manner. *Perception and Psychophysics*, *27*, 343–350.

Francis, G., & Grossberg, S. (1996). Cortical dynamics of boundary segmentation and reset: Persistence, afterimages, and residual traces. *Perception*, *25*, 543–567.

Francis, G., Grossberg, S., & Mingolla, E. (1994). Cortical dynamics of feature binding and reset: control of visual persistence. *Vision Research*, *34*(8), 1089–1104.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In Altmann, G. T. M. (Ed.), *Cognitive Models of Speech Processing*, chap. 3, pp. 50–86. MIT Press, Cambridge, MA.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Pyschology: Human Perception and Performance*, *6*, 110–125.

Gaudiano, P., & Grossberg, S. (1991). Vector associative maps: Unsupervised real-time error-based learning and control of movement trajectories. *Neural Networks*, *4*, 493–504.

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology*, *25*, 1–42.

Govindarajan, K. K., Grossberg, S., Wyse, L., & Cohen, M. A. (1994). A neural network model of auditory scene analysis and source segregation. Tech. rep. CAS/CNS-TR-94-039, Boston University, Boston, MA.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and pyschophysics*. Kreiger Press, New York.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, *38*, 299–310.

Grossberg, S. (1969). On the production and release of chemical transmitters and related topics in cellular control. *Journal of Theoretical Biology*, *22*, 325–364.

Grossberg, S. (1973). Contour enhancement, short term meory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, *52*, 217–257. Reprinted in Grossberg, S. (1982), **Studies of Mind and Brain**.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*, 187–202.

Grossberg, S. (1978a). A theory of human memory: Self-organization and performance of sensory-motor codes, maps and plans. In Rosen, R., & Snell, F. (Eds.), *Progress in Theoretical Biology*, Vol. 5, pp. 233–374. Academic Press, New York. Reprinted in Grossberg, S. (1982). **Studies of Mind and Brain.**

Grossberg, S. (1978b). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models?. *Journal of Mathematical Psychology*, *3*, 199–219.

Grossberg, S. (1980). How does a brain build a cognitive code?. *Psychological Review*, *87*, 1–51.

Grossberg, S. (1984). Unitization, automaticity, temporal order, and word recognition. *Cognition and Brain Theory*, *7*, 263–283.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In Schwab, E. C., & Nusbaum, H. C. (Eds.), *Pattern Recognition by Humans and Machines, vol. 1: Speech Perception*. Academic Press, New York.

Grossberg, S. (1994). 3-D vision and figure ground separation by visual cortex. *Perception and Psychophysics*, *55*(1), 48–120.

Grossberg, S. (1995). The attentive brain. *American Scientist*, *83*, 438–449.

Grossberg, S. (1999a). Brain feedback and adaptive resonance in speech perception. *Behavioral and Brain Sciences*. In Press. Also available as Boston University Technical Report CAS/CNS-99-022.

Grossberg, S. (1999b). How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision*, *12*, 163–186.

Grossberg, S. (1999c). Pitch-based streaming in auditory perception. In Griffith, N., & Todd, P. (Eds.), *Musical networks: Parallel distributed perception and performance*, 117–140. MIT Press, Cambridge, MA.

Grossberg, S. (1999d). The link between attention, brain learning, and consciousness. *Consciousness and Cognition*, *8*, 1–44.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 481–503.

Grossberg, S., & Merrill, J. (1996). The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience*, *8*(3), 257–277.

Grossberg, S., & Stone, G. (1986). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review*, *93*, 46–74.

Grossberg, S., & Williamson, J. (1998a). A neural model that links cortical development to visual perception. Tech. rep. CAS/CNS-TR-98-022, Boston University, Boston, MA.

Grossberg, S., & Williamson, J. (1999). A self-organizing neural system for learning to recognize textured scenes. *Vision Research*, *39*, 1385–1406.

Grunewald, A., & Grossberg, S. (1998). Self-organization of binocular disparity tuning by reciprocal corticogeniculate interactions. *Journal of Cognitive Neuroscience*, *10*(2), 199–215.

Hardcastle, W. J., Gibbon, F., & Scobbie, J. M. (1995). Phonetic and phonological aspects of English affricate production in children with speech disorders. *Phonetica*, *52*, 242–250.

He, J., Hashikawa, T., Ojima, H., & Kinouchi, Y. (1997). Temporal integration and duration tuning in the dorsal zone of cat auditory cortex. *Journal of Neuroscience*, *17*(7), 2615–2625.

Hedrick, M. (1997). Effect of acoustic cues on labeling fricatives and affricates. *Journal of Speech, Language, and Hearing Research*, *40*, 925–938.

Howell, P., & Rosen, S. (1983). Production and perception of rise time in the voiceless affricate/fricative distinction. *Journal of the Acoustical Society of America*, *73*(3), 976–984.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556.

Jelinek, F. (1995). Training and search methods for speech recognition. *Proceedings of the National Academy of Sciences USA*, *92*(22), 9964–9969.

Kluender, K. R., & Walsh, M. A. (1988). Effect of vowel duration on the perception of syllable-initial /ʃ/ and /tʃ/. *Journal of the Acoustical Society of America*, *84*(Suppl. 1), S159.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Reivew*, *74*, 431–461.

Lippmann, R. P. (1989). Review of neural networks for speech recognition. *Neural Computation*, *1*(1), 1–38.

Lisker, L. (1985). The pursuit of invariance in speech signals. *Journal of the Acoustical Society of America*, *77*(3), 1199–1202.

Luce, R. D. (1959). *Individual Choice Behavior*. Wiley, New York.

Mannes, C. (1993). *Neural network models of serial order and handwriting movement generation*. Ph.D. thesis, Boston University.

Margolin, D. I. (1991). Cognitive neuropsychology. Resolving enigmas about Wernicke's aphasia and other higher cortical disorders. *Archives of Neurology*, *48*(7), 751–65.

Markram, H., & Tsodyks, M. (1996). Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature*, *382*(6594), 807–810.

Marslen-Wilson, W. (1987). Functional parallelism in spoken-word recognition. *Cognition*, *25*, 71–102.

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In Altmann, G. T. M. (Ed.), *Cognitive Models of Speech Processing*, chap. 7, pp. 148–172. MIT Press, Cambridge, MA.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Mattys, S. L. (1997). The use of time during lexical processing and segmentation: a review. *Psychonomic Bulletin and Review*, *4*(3), 310–329.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *23*, 1–44.

McQueen, J. M. (1991). The influence of the lexicon on phoentic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 433–443.

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes*, *10*(3), 309–331.

Miller, J. L., & Eimas, P. D. (1995). Speech perception: From signal to word. *Annual Review of Psychology*, *46*, 467–492.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later occurring information on the perception of stop consonant and semi-vowel. *Perception and Psychophysics*, *25*(3), 457–465.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In Eimas, P. D., & Miller, J. L. (Eds.), *Perspectives on the Study of Speech*, chap. 2, pp. 39–74. Lawrence Erlbaum Associates, Hillsdale, NJ.

Miller, S. L., Delaney, T. V., & Tallal, P. (1995). Speech and other central auditory processes: insights from cognitive neuroscience. *Current Opinion in Neurobiology*, *5*(2), 198–204.

Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, *95*(3), 1603–1616.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, *62*(3), 714–719.

Newman, R., & Sawusch, J. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception and Psychophysics*, *58*(4), 540–560.

Norris, D. G. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

Pickett, J. M., Bunnell, H. T., & Revoile, S. G. (1995). Phonetics of intervocalic consonant perception: retrospect and prospect. *Phonetica*, *52*, 1–40.

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representation in word recognition. *Cognition*, *25*, 21–52.

Pitt, M. A., & Samuel, A. G. (1990). The use of rhythm in attending to speech. *Journal of Experimental Psychology Human Perception and Performance*, *16*(3), 564–573.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1988). *Numerical recipes in C: The art of scientific computing*. Cambridge University Press, Cambridge, England.

Price, P., & Ostendorf, M. (1996). Combining linguistic with statistical methods in modeling prosody. In Morgan, J. L., & Demuth, K. (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, pp. 67–83. Lawrence Erlbaum Associates, Mahwah, NJ.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, *90*, 2956–2970.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, *8*(4), 516–521.

Repp, B. H. (1980). A range–frequency effect on perception of silence in speech. *Haskins Lab Status Report*, *SR–61*, 151–165.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110.

Repp, B. H. (1984). Effects of temporal stimulus properties on perception of the [sl]-[spl] distinction. *Phonetica*, *41*, 117–124.

Repp, B. H. (1985). Perceptual coherence of speech: Stability of silence-cued stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(6), 799–813.

Repp, B. H. (1988). Integration and segregation in speech perception. *Language and Speech*, *31*(3), 239–271.

Repp, B. H. (1992). Perceptual restoration of a “missing” speech sound: Auditory induction or illusion?. *Perception and Psychophysics*, *51*(1), 14–32.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In Harnad, S. N. (Ed.), *Categorical Perception: The Groundwork of Cognition*, chap. 3, pp. 89–112. Cambridge University Press, New York.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 621–637.

Rosen, S., Darling, A. M., Faulkner, A., & Huckvale, M. (1993). Cue interaction in an intervocalic voiceless affricate/fricative contrast. *Speech, hearing and language: Work in progress*, *7*, 183–197. Department of Phonetics and Linguistics, University College London.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory, and linguistic aspects. *Philosophical Transactions of the Royal Society of London, Series B*, *336*, 367—373. Also published in Carlyon, R.P., Darwin, C.J., and Russell, I.J. (Eds.), *Processing of Complex Sounds by the Auditory System*, pp. 73–79. Clarendon Press, Oxford.

Samuel, A. G., van Santen, J. P. H., & Johnston, J. C. (1982). Length effects in word perception: We is better than I but worse than you or them. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 91–105.

Samuel, A. G., van Santen, J. P. H., & Johnston, J. C. (1983). Reply to Matthei: We really is worse than you or them, and so are ma and pa. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 321–322.

Samuel, A. G. (1987). Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language*, *26*(1), 36–55.

Samuel, A. G. (1991). A further examination of attentional effects in the phonemic restoration illusion. *Quarterly Journal of Experimental Psychology: Human experimental psychology*, *43A*(3), 679–699.

Samuel, A. G., Kat, D., & Tartter, V. C. (1984). Which syllable does an intervocalic stop belong to? A selective adaptation study. *Journal of the Acoustical Society of America*, *76*(6), 1652–1663.

Stevens, K. N. (1993). Modelling affricate consonants. *Speech Communication*, *13*, 33–43.

Summerfield, Q., Bailey, P. J., Seton, J., & Dorman, M. (1981). Fricative envelope parameters and silent intervals in distinguishing ’slit’ and ’split’. *Phonetica*, *38*, 181–192.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393.

Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. A., & Healy, E. W. (1997). Spectral restoration of speech: intelligibility is increased by inserting noise in spectral gaps. *Perception and Psychophysics*, *59*(2), 275–283.

Warren, R. M., & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception and Psychophysics*, *9*(3B), 358–363.

Warren, R. M., & Sherman, G. L. (1974). Phonemic restorations based on subsequent context. *Perception and Psychophysics*, *16*(1), 150–156.

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American*, *223*, 30–36.

Wickelgren, W. A. (1976). Phonetic coding and serial order. In Carterette, E. C., & Friedman, M. P. (Eds.), *Handbook of Perception: Language and Speech*, Vol. VII, chap. 7, pp. 227–264. Academic Press, New York.

APPENDIX

A. ARTWORD Model Equations

ARTWORD is defined mathematically by differential equations which indicate how item and chunk activities change in time. These network equations extend those developed in Grossberg *et al.* (1997) to include chunks sensitive to multiple items, using the principles of the masking field architecture of Cohen and Grossberg (1986, 1987). Below, Greek letters denote fixed parameters, and I_i, w_i, and u_j denote the activation levels of the ith input and working memory item, and jth list chunk, respectively. Likewise, z_{iw} and z_{ju} denote the quantity of transmitter activated by the ith item in the working memory and the jth chunk, respectively.

Item Working Memory

Working memory activation is described by a membrane, or shunting, network equation (Grossberg, 1973). The activity w_i of the ith item coded in working memory changes according to the equation

$$\frac{dw_i}{dt} = \Gamma\left[(1-w_i)\left(I_i + \eta H(w_i)\sum_{j\leftrightarrow i}\tau_{ij}u_j^+ z_{ju}\right) - w_i\left(\alpha + \beta\sum_k u_k + \kappa\sum_{k\neq i} w_k\right)\right], \qquad (A1)$$

where the symbol $j \leftrightarrow i$ denotes the relation "j is connected to i"; i.e., the existence of an excitatory synaptic pathway between item i in the working memory and chunk j in the grouping network. Parameter Γ, common to both the working memory and the list chunking network, defines the global processing rate at which neurons in the network integrate their inputs. In general, Γ tracks the mean rate of incoming speech so that processing can adjust for variations in segmental durations that occur at different speaking rates (Grossberg *et al.*, 1997). In the present article, ARTWORD simulates the Repp *et al.* (1978) data presented at a single speaking rate, so it suffices to fix Γ to a constant value for all simulations.

By Equation (A1), working memory activity increases to a maximum of 1 via excitatory inputs that are shunted by term (1-w_i). Shunting the excitatory inputs ensures that network activity remains bounded below 1. For a given item i, two sources of excitation exist: the bottom-up input I_i and the summed activity of all chunks j connected to item i by positive weights τ_{ij}. The activity u_j of the jth chunk must exceed a positive threshold γ_{ju} before it can begin to send excitatory top-down signals $u_j^+ = \max(u_j - \gamma_{ju}, 0)$ to working memory items. The signals emitted by each chunk are then multiplied, or gated, by the supply of neurotransmitter z_{ju} currently available to that chunk. The net top-down signal is scaled by the global parameter η which indicates the influence of top-down feedback on all working memory item activations relative to bottom-up input I. Top-down input is also gated by the Heaviside function of item activity, $H(w_i)$ defined to be zero when $w_i = 0$ and to be 1 when $w_i > 0$. This gating ensures that top-down feedback does not activate a particular item i until after that item is first excited by bottom-up input I_i. Thus, it performs a matching process that prevents the top-down expectations themselves from activating their own items in the absence of external input. Some partial activation from bottom-up input, however weak, is necessary for top-down feedback to begin to support phonemic item codes.

Three sources of inhibitory input act to counter the excitation of each working memory item. Each item decays passively at rate α, and actively due to both non-specific top-down inhibition and on-center, off-surround competition within the working memory. The top-down inhibition via the term $\beta \sum_k u_k$ serves as an automatic gain control to attenuate or suppress unexpected features in the working memory as chunks in the grouping network become active, and to balance the excitatory support which expected items receive from their active chunks. The on-center, off-surround competition via the term $\kappa \sum_{k \neq i} w_k$ keeps the total activity in the working memory normalized by attenuating old items as new ones become active. This competition also produces a natural recency gradient of temporal order information, so that, other things being equal, a more recently presented input will command a higher item activation – and thus a greater proportion of the total activity forming the pattern across working memory – than would a less recently presented input. The three inhibitory inputs to each item are shunted by the term $-w_i$, keeping item activity bounded to be greater than or equal to zero.

List Chunking Network

Each list node, or chunk, in the grouping network is connected via top-down synaptic pathways to the same items in the working memory that excite it in a bottom-up fashion. The top-down weights τ_{ji} are identical to the corresponding bottom-up weights τ_{ij}. Like items in the working memory, chunks in the grouping network obey shunting membrane equations whose integration rate is modulated by parameter Γ. For list node j in the grouping network, activity u_j changes according to the equation

$$\frac{du_j}{dt} = \Gamma \left[(1 - u_j) \left(\frac{\rho}{\phi_j} \sum_{i \leftrightarrow j} \tau_{ji} w_i^+ z_{iw} + \phi_j f(u_j) z_{ju} \right) - u_j \left(\delta + \sum_{k \neq j} \psi_{kj} g(u_k) \right) \right], \qquad (A2)$$

where the sigmoidal signal functions f and g act to contrast-enhance the excitatory and inhibitory chunk interactions, respectively, and are defined by

$$f(x) = \frac{x^2}{0.75^2 + x^2} \quad \text{and} \quad g(x) = \frac{x^2}{0.15^2 + x^2}. \qquad (A3)$$

As in the working memory, both excitatory and inhibitory inputs to list nodes are shunted via the terms $(1 - u_j)$ and $-u_j$, respectively, thereby keeping list node activity bounded between zero and one. List nodes are excited by their working memory items $w_i, i \leftrightarrow j$, when item activity exceeds a threshold γ_{iw}. Thresholded activity $w_i^+ = \max(w_i - \gamma_{iw}, 0)$ is multiplied by synaptic weights τ_{ji} and further gated by the neurotransmitter available to each item, z_{iw} before exciting list nodes via term $\frac{\rho}{\phi_j} \sum_{i \leftrightarrow j} \tau_{ji} w_i^+ z_{iw}$. The item activations that excite a list node are normalized by the number of items that form synaptic connections with that node; that is, by the number of items encoded by the list. Term ρ/ϕ_j, where ϕ_j is proportional to the number of items encoded by list chunk j, affects this normalization, or conservation of synaptic sites. Normalizing by list length helps to prevent chunks from becoming active above their positive feedback thresholds γ_{ju} before all of their constituent items have been activated in the working memory.

Each list node also sends self-excitatory input $\phi_j f(u_j) z_{ju}$ via the sigmoidal signal function, $f(u)$. This positive feedback is scaled to be larger for chunks encoding longer lists, via the term ϕ_j. Scaling self-excitatory feedback in proportion to list length gives larger chunks a competitive advantage when all of their items are active in the working memory by allowing them to overcome the greater activations of chunks coding for sublists of their items. Self-excitatory feedback is also gated by a chunk's available transmitter supply z_{ju}. Such gating ensures that resetting of chunk activation by habituative collapse is possible.

Inhibitory input to each chunk comes from only two sources: passive decay and competition from other chunks. Chunk passive decay is determined by parameter δ, chosen to be smaller than item passive decay α so that chunk activity lags the item activity it is integrating (Grossberg *et al.*, 1997). The inhibition $\sum_{k \neq j} \psi_{kj} g(u_k)$ from other chunks is scaled by the feedback function g defined in Equation (A3), and by the inhibitory synaptic coefficients ψ_{kj} defining the competitive strength between two chunks k and j. The ψ coefficients are set to zero for two chunks that code for mutually exclusive lists, and grow with increasing overlap between chunks. Cohen and Grossberg (1986) defined the strength of the inhibitory interaction from chunk k to chunk j in proportion to the product $|K|(|K \bigcap J| + 1)$, where $|K|$ and $|J|$ denote the lengths of the lists coded by chunks k and j. Such a rule specifies that chunks coding for longer lists have stronger masking parameters (via $|K|$), that inhibition grows proportionally to list overlap (via $|K \bigcap J|$), and that all chunks maintain weak long-range inhibitory interactions (via the term 1). In the present article, the ψ_{kj} were selected based on chunk size and overlap, but varied as necessary because, unlike in the network developed by Cohen and Grossberg (1986), the masking field did not contain all possible chunks coding the items in the working memory.

By Equation (A3), inhibitory feedback $\sum_{k \neq j} \psi_{kj} g(u_k)$ between chunks becomes active earlier than chunk self-excitatory feedback $\frac{\rho}{\phi_j} \sum_{i \leftrightarrow j} \tau_{ji} w_i^+ z_{iw}$ because of the smaller term 0.15^2 in the definition of $g(x)$ than term 0.75^2 in $f(x)$. This predominantly inhibitory interaction within the grouping network helps to prevent chunks from entering self-sustaining positive feedback loops when they are presented with an insufficient input.

Transmitter Dynamics

Equations (A1) and (A2) show that all specific signals between the working memory and grouping network, as well as excitatory feedback within the grouping network, are gated by their respective levels of available transmitter. Transmitters for items and lists nodes obey laws of an identical form, introduced by Grossberg (1969):

$$\frac{d}{dt} z_{wi}(t) = (1 - z_{wi})\,\epsilon - z_{wi}\left[\lambda w_i^+ + \mu (w_i^+)^2\right], \qquad (A4)$$

$$\frac{d}{dt} z_{uj}(t) = (1 - z_{uj})\,\epsilon - z_{uj}\left[\lambda u_j^+ + \mu (u_j^+)^2\right]. \qquad (A5)$$

Similar laws have recently been reported in visual and somatosensory cortex (Abbott, Varela, Sen, & Nelson, 1997; Markram & Tsodyks, 1996). In Equations (A4) and (A5), transmitters accumulate at constant rate ϵ until they attain a maximal level of 1 via the shunting term $(1 - z)$. When no signals are consuming the transmitter supply, so that $w^+ = 0$, or $u^+ = 0$, then the transmitter accumulates until it equilibrates at a value of 1. When suprathreshold signals w^+ or u^+ are sent along the pathways, then transmitter habituates to lower equilibrium values as determined by the strength of the signals and by the parameters λ and μ, which specify linear and quadratic rates of activity-dependent transmitter inactivation, respectively (Gaudiano & Grossberg, 1991; Grossberg et al., 1997).

Input to working memory

ARTWORD incorporates fixed acoustic-phonetic pathways, assumed to have been learned during a prior stage of self-organization, into the activation of working memory items. Thus phonemic item responses will be stronger to sounds that better match the bottom-up pattern extracted by lower levels of transient and sustained auditory signal processing, as in Boardman *et al.* (1999). For example, a shorter fricative noise interval will provide greater input to the /t$\int$/ item in working memory than to the /$\int$/ item, because in natural speech the voiceless affricate consonant has a shorter duration than the voiceless fricative (Howell & Rosen, 1983). Greater input can, in general, take the form of activation at a *greater amplitude* or activation for a *longer duration*. In the simulations, input to the /g/, /r/, /ei/, /I/, and /p/ items were fixed as pulses of equal amplitude and duration for all combinations of silence and noise durations. Input to the stop /t/, affricate /t$\int$/, and fricative /$\int$/ items consisted of fixed amplitude pulses whose durations depended on the segmental durations of silence and noise. In particular, the duration of the /t/ item was chosen to increase monotonically with the interval of preceding silence and the duration of noise, such that its duration ranged from 10 ms (at silence durations $\leq$ 10 ms) to 41 ms (maximal /t/ input, at silence duration=100 ms, noise duration=182 ms). The /t$\int$/ input duration increased monotonically with silence duration but decreased exponentially with increasing fricative noise duration, while the /$\int$/ input behaved in a complementary fashion. Unless otherwise noted, the durations of the /t/, /t$\int$/, and /$\int$/ input pulses reflecting the acoustic-phonetic map are given by the following equations:

$$\text{Duration of /t/} = 10 + 0.025(ND - 52)\sqrt{SD^+}, \qquad (A6)$$

$$\text{Duration of /t}\int\text{/} = (4.5)2^{-0.025ND}\min(SD, \frac{ND + 18}{2}), \qquad (A7)$$

and

$$\text{Duration of /}\int\text{/} = ND - \text{Duration of /t}\int\text{/}, \qquad (A8)$$

where ND = duration of the fricative noise, SD = duration of the preceding silence interval, and $SD^+ = \max(SD - 10, 0)$.

B. Parameters Used in ARTWORD Simulations

This section describes the parameters used to generate the simulations depicted in Sections 6 and 7. Greek letters refer to parameters in Eqns. (A1)–(A5). In the following, phonemic item codes are indexed by $i = 1, ..., 8$, denoting respectively items /g/, /r/, /ei/, /t/, /t$\int$/, /$\int$/, /I/, and /p/. Chunk codes are indexed by $j = 1, ..., 4$, denoting respectively chunks GRAY, GREAT, CHIP, and SHIP. All weights given below are item-to-chunk weights τ_{ij}, since the reciprocal weights are equal; i.e., $\tau_{ij} = \tau_{ji}$. Thus, for example, in describing the weights τ_{ij}, the value τ_{12} denotes the weight between the first item (/g/) and the second chunk (GREAT). In describing the inhibitory coefficients between chunks, ψ_{kj}, the first subscript k denotes the source of the inhibitory signal and the second subscript j denotes the target. Thus, ψ_{32} denotes the inhibitory influence of chunk 3 (CHIP) on chunk 2 (GREAT). Unless otherwise noted, $\psi_{kj} = |K|(|K \bigcap J|)$, where $|K|$ and $|J|$ denote the lengths of the lists coded by chunks k and j. Each normalization coefficient ϕ_j was set equal to the number of inputs encoded by chunk j; i.e., $\phi_j = |\{\tau_{ij} : \tau_{ij} > 0\}|$. For all simulations, the following parameters were fixed: Γ = 0.7; α = 0.03; κ = 0.1; β = 1; ρ = 70; δ = 0.02; λ = 0.10; μ = 3.0; $\gamma_{iw} = 0.12, \forall i$. In Section 6, ϵ = 0.01 and in Section 7, ϵ = 0.05.

Section 6.1, Figures 9 and 10: The masking field contains two chunks, u_1 = GRAY, u_2 = CHIP.

Top-down thresholds were set to $\gamma_{1u} = 0.12$, $\gamma_{2u} = 0.14$. Top-down feedback scale parameter η in Equation (A1) was set to 0. Weights between items and chunks were as follows: τ_{11} = 0.1333; τ_{21} = 0.2333; τ_{31} = 0.4334; τ_{12} = 0.1000; τ_{22} = 0.1500; τ_{32} = 0.2500; τ_{42} = 1.2000. All other weights $\tau_{ij} = 0$. Inhibitory coefficients were $\psi_{1,2} = \psi_{2,1} = 12$. For Figure 9, the /g/ item was activated from $t = 0$ to $t = 62$. For Figure 10, the items /g/, /r/, and /ei/ were activated sequentially for 62 ms each, beginning at $t = 0$.

Section 6.2, Figure 11: All parameters were chosen as above, except the top-down feedback parameter η in Equation (A1) was set to 3. In Figure 11A, item activation was as in Figure 10. In Figure 11B, item activation was as in Figure 11A with the additional activation of the /t/ item following the offset of the /ei/ item, for a duration of 62 ms.

Section 6.3, Figures 12 and 13: All parameters were chosen as above, except for the input presentation. For Figure 12, the duration of the /t/ item activation was 34 ms. The items /g/, /r/, and /ei/, were activated sequentially for 62 ms each, beginning at $t = 0$ ms. The /t/ item was activated after a silence duration of 60, 65, 70, and 75 ms, in A, B, C, and D, respectively.

For Figure 13, the silence duration between /ei/ offset and /t/ onset varied from 50 ms to 100 ms in steps of 5 ms. The duration of the /t/ item activation varied from 32 ms to 52 ms in steps of 2 ms. For each combination of silence duration and /t/ duration, the entire network was integrated and total activation of the GREAT chunk was computed. To produce Figure 13, a two-dimensional grid with 1 ms steps in each dimension between 50-100 and 32-52 ms was created and the GREAT chunk activation was interpolated over this grid using a cubic polynomial. Figure 13 is a contour map of the resulting values, with darker shades representing greater GREAT chunk activation.

Section 6.4, Figure 14: Chunk inhibitory coefficients were $\psi_{21} = 12, \psi_{23} = 3, \psi_{32} = 4$. Top-down threshold $\gamma_{3u} = 0.12$. Weights between items and chunks 1 and 2 were as above, with the exception that $\tau_{42} = 1.6$. The weights between items and chunk 3 (CHIP) are as follows: τ_{53} = 0.35; τ_{63} = 0.01; τ_{73} = 0.20; τ_{83} = 0.25. All other parameters were chosen as above.

Section 7, Figures 15–17: Top-down thresholds were set to $\gamma_{ju} = 0.18$, all j. Weights between items and chunks 1-3 were as above, except that $\tau_{43} = 0.02$ and $\tau_{63} = 0.02$. The weights between items and chunk 4 (SHIP) are as follows: τ_{54} = 0.02; τ_{64} = 0.35; τ_{74} = 0.20; τ_{84} = 0.25. Chunk inhibitory coefficients were $\psi_{32} = 1, \psi_{23} = 15, \psi_{34} = 9, \psi_{43} = 9$. Silence durations and item activation durations for items 5, 6, and 7 were as specified in Equations (A6)–(A8). Input amplitudes were 0.18 (items 5, 6, 7, and 8), 0.36 (item 4), and 0.12 (items 1, 2, and 3). All other parameters were chosen as above.