Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Phonetics 31 (2003) 305–320
Journal of Phonetics
www.elsevier.com/locate/phonetics

Puzzle-solving science: the quixotic quest for units in speech perception

Stephen D. Goldinger[a,*], Tamiko Azuma[b]

[a] *Department of Psychology, Arizona State University, Box 871104, Tempe, AZ 85287-1104, USA*
[b] *Department of Speech & Hearing Science, Arizona State University, USA*

Received 15 August 2002; received in revised form 18 March 2003; accepted 19 March 2003

Abstract

Although speech signals are continuous and variable, listeners experience segmentation and linguistic structure in perception. For years, researchers have tried to identify the basic building-block of speech perception. In that time, experimental methods have evolved, constraints on stimulus materials have evolved, sources of variance have been identified, and computational models have been advanced. As a result, the slate of candidate units has *increased*, each with its own empirical support. In this article, we endorse Grossberg's *adaptive resonance theory* (ART), proposing that speech units are emergent properties of perceptual dynamics. By this view, units only "exist" when disparate features achieve *resonance*, a level of perceptual coherence that allows conscious encoding. We outline basic principles of ART, then summarize five experiments. Three experiments assessed the power of social influence to affect phoneme-syllable competitions. Two other experiments assessed repetition effects in monitoring data. Together the data suggest that "primary" speech units are strongly and symmetrically affected by bottom-up and top-down knowledge sources.

1. Introduction

In describing the practice of "normal science", Kuhn (1962) noted that scientific inquiry often amounts to *puzzle-solving*. In speech perception, few puzzles can rival the long-standing quest for the *fundamental unit* of perceptual analysis. Given the hierarchical structure of language, spoken communication logically requires unitization, at some level(s). That is, small sets of segmental contrasts are recombined to create endless messages. If listeners can organize features into

*Corresponding author. Tel.: +1-480-965-0127; fax: +1-480-965-8544.
E-mail address: goldinger@asu.edu (S.D. Goldinger).

0095-4470/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0095-4470(03)00030-5

phonemes, and phonemes into higher units, information-transfer rates are greatly reduced (Pisoni & Luce, 1987). Thus, researchers have tried to determine which units are retrievable from the continuous speech signal, and how they build larger percepts.

Research on speech units has an interesting history, beginning with a surprising result: Savin and Bever (1970) played non-words to listeners for two speeded identification tasks. They found slower identification of syllable-initial phonemes than of syllables in their entirety. Savin and Bever proposed that syllables are identified first in speech; phonemes can be retrieved afterward. This counterintuitive hypothesis spawned many monitoring studies, wherein listeners are given targets (e.g., phonemes) to detect in lists of spoken materials.[1] The results were quite inconsistent (see Decoene, 1993). Warren (1971) embedded phoneme targets in contexts of increasing linguistic complexity, verifying that phoneme-detection took longer than syllable-detection. However, RTs were also affected by sentence contexts, suggesting that phonemes may be recovered from higher units of linguistic analysis. Similarly, Foss and Swinney (1973) found that people could detect two-syllable words faster than initial syllables. Reasoning that bisyllables are unlikely as fundamental units, they cautioned that monitoring may not reveal units of *perception*; they may only reflect conscious *identification* of targets.

This caution was reiterated by McNeill and Lindig (1973), who found that monitoring times were remarkably sensitive to changes in selective attention. Listeners were fastest whenever their assigned targets matched the stimulus, even with targets as large as sentences. McNeill and Lindig concluded,

> What is ''perceptually real'' is what one pays attention to. In normal language use, the focus of attention is the meaning of an utterance. Subordinate levels become the focus of attention only under special circumstances. […] there is no clear sense in which one can ask what the ''unit'' of speech perception is. There is rather a series (or a network) of processing stages and each can in principle be the focus of attention (p. 430).

In general, McNeill and Lindig suggested that monitoring data reflect top-down matching processes, which seemingly assumed priority over bottom-up, stimulus factors. However, later studies showed that many variables (including stimulus factors) affect unit-detection times. These variables included the perceptual ambiguity of different phonemes (Healy & Cutting, 1976; Swinney & Prather, 1980), the quality of suprasegmental cues (Cutler, 1976), lexicality (Rubin, Turvey, & van Gelder, 1976), the presence of foils (Norris & Cutler, 1988), and general changes in method (Mehler, Segui, & Frauenfelder, 1981). Moreover, every potential conclusion was advanced. Different authors argued for the primacy of phonemes (Norris & Cutler, 1988) and syllables (Mehler et al., 1981). Healy and Cutting (1976) concluded that ''phonemes and syllables are equally basic to speech perception''. Others reiterated the view that monitoring procedures cannot resolve the issue (Mills, 1980; Shand, 1976).

Given such variable monitoring data, many researchers have applied indirect methods, hoping to remove influences of selective attention. For example, Massaro (1972) used a recognition-masking procedure, implicating syllables as primary units. Based on selective adaptation data, Samuel (1989) argued for demisyllables. Decoene (1993) used a primed-matching task, concluding

[1] Due to space limitations, we can only provide a cursory review, focusing mainly on monitoring studies. See referenced articles (e.g., Decoene, 1993; Plomp, 2002) for more detailed and comprehensive treatment.

that phonemes are more perceptually basic. Citing gating data, Marslen-Wilson and Warren (1994) suggested that listeners use phonetic features directly, without organization into phonemes or syllables. Kolinsky, Morais, and Cluytens (1995) assessed illusory conjunctions in dichotic listening, proposing that syllables are primary. Further complicating matters, numerous studies suggest that *different* units assume priority in different languages (Cutler, Mehler, Norris, & Segui, 1986; Cutler & Otake, 1994; Sebastian-Galles, Dupoux, Segui, & Mehler, 1992; Tabossi, Collina, Mazzetti, & Zoppello, 2000).

Given such data, some authors suggested that no "primary" unit exists. Instead, we should focus on *interactions* among levels of perceptual representation (Jusczyk, 1986). For example, recent theories have attempted to *organize* the disparate data, seeking optimal segmentation strategies across languages with different sound-patterns (McQueen, Norris, & Cutler, 1994). In this article, we suggest that Grossberg's (1980) *adaptive resonance theory* (ART) may provide such organization, with elegant ties to more global cognitive processes.

2. Toward a synthesis: attention, unitization, and adaptive resonance

Considered collectively, 30 years of speech-unit research has generated little apparent progress. If the goal was to decide a "winner", the enterprise has clearly failed: Despite dozens of studies, the candidate list has actually grown.[2] From at least one theoretical perspective, the classic question of speech units seems ill conceived. Specifically, the usual question (phonemes or syllables?) presumes a bottom-up system of cognitive processing that is likely incorrect (see commentaries accompanying Norris, McQueen, & Cutler, 2000). Moreover, theories incorporating top-down processes (e.g., McClelland & Elman, 1986) typically presume fixed units, reified to fit intuitions about speech (e.g., feature, phoneme, and word nodes in TRACE). By contrast, an alternate view—self-organization through adaptive resonance—simply nullifies the "units" question. In doing so, it has the ironic effect of finding new value in speech-unit studies that were previously dismissed.

In this section, we briefly review basic principles of ART, focusing on aspects relevant to speech units. These ideas derive from Grossberg and colleagues (Grossberg, 1980, 1999, 2003 [this volume]; Grossberg, Boardman, & Cohen, 1997; also Hinton & Shallice, 1991; Plaut, McClelland, & Seidenberg, 1996). In speech perception, a fundamental hypothesis of ART (actually, a specific version called ARTPHONE) is that conscious percepts, and speech units of all possible grain sizes, are emergent products of resonant brain states. This is a fully interdependent system: Perception occurs when bottom-up and top-down knowledge sources bind into a stable state. Processing in ART begins when featural input activates *items* (feature clusters) in working memory. Items, in turn, activate *list chunks* in memory. These are products of prior learning (perhaps prototypes) that may correspond to any combination of features. Possible chunks therefore include phonemes, syllables, and words.

[2] We will not belabor the point, but the speech-unit literature raises another red flag: The "file-drawer" problem. Given a literature with oscillating patterns of results, we imagine that many null results have accumulated over the years. If the ratio of failed to published experiments is 1:1, all published articles should adopt a statistical criterion of $p<0.025$. Nevertheless, the practice of piloting and refining experiments is likely very common in this challenging domain of inquiry.

Once items activate list chunks, a feedback cycle begins. Items feed activation upward through synaptic connections; input-consistent chunks will return activation. If items receive sufficient top-down confirmation, they continue sending activation upward (as in familiar PDP models). Within limits, this feedback loop (a *resonance*) is self-perpetuating, binding the respective activation patterns into a coherent whole. The bottom-up pattern that initiates interactive activation need not perfectly match its resultant feedback for resonance to occur. Cooperation between "levels" and competition within "levels" smoothes out small mismatches, but large mismatches prohibit resonance.

Chunks in ART are learned *attractor states* (Grossberg, 1980; Plaut et al., 1996). Of particular importance, chunks can be learned at multiple scales of analysis, and naturally form nesting relations (Vitevitch & Luce, 1999). In spoken words, acoustic features are equally consistent with many potential chunks, including segments, diphones, triphones, syllables, and words (see *subsymbols* in Van Orden & Goldinger, 1994). As ART processes a spoken word, small "local" resonances begin to form, perhaps with chunks approximating phonemes. But these local dynamics occur in the context of larger, global dynamics: Phonemes participate in syllable-sized chunks, which participate in word-sized chunks, etc. As in PDP models, high-level representations become active early, then exert influence on lower building-blocks. However, ART does *not* assume such units directly in its architecture, and "high-level" chunks have no direct connections to "low-level" chunks. Instead, all structures self-organize, in parallel, through competitive dynamics. Unitization is a flexible, opportunistic process: The most predictive units will dominate behavior.

To illustrate this hypothesis, and to underscore the difficulty of speech units, a comparison with reading may be helpful. Despite historical difficulties, some theorists remain confident that research will isolate reliable, bottom-up speech units (e.g., Nearey, 1997). Intuitively, such a discovery would resolve the speech-unit issue. However, the literature on reading suggests otherwise. Printed English is a man-made system: Its pristine, separated letters provide ideal bottom-up support. Nevertheless, abundant evidence suggests that readers *impose* complex unitization onto letter strings: Various experiments have implicated bigrams, onsets and rimes, and numerous phonetically motivated graphemes (Healy, 1994; Kessler & Treiman, 2001). Moreover, studies of *repetition blindness* suggest that people can shift attention between such units, as task demands require (Harris & Morris, 2001). If myriad units arise in an ideal bottom-up domain, it suggests they are *created* by perceptual dynamics. In this regard, ART has three features that relate to speech units:

(1) Attractor states are stable, learned memory structures. They "exist" as contributors to larger perceptual events. However, their *psychological reality* in any perceptual event requires local resonance, momentarily independent of the global resonances that define speech communication. This amounts to a race, with attractors of all sizes competing for resonance, based on consistency with the input.

"Psychological reality" denotes a state of perceptual consequence, such that a person is momentarily cognizant of the attractor (e.g., a phoneme). Intuition suggests that, in normal communication, words and sentences have psychological reality. On sublexical levels, syllables and phonemes may achieve psychological reality, but only occasionally. (Foss & Swinney, 1973, suggested this occurs when fluent processing fails, as when someone stammers.) Note that psychological reality has little bearing on *functional* utility: Attractors approximating phonemes may serve vital perceptual functions, stabilizing local dynamics so global resonances can form

(Van Orden & Goldinger, 1994). In ART, however, their "existence" is a stochastic function of self-organization, rather than design.

(2) Competitions between local and global resonances do not reflect activation levels of specific units. Instead, they reflect *masking*, which helps maintain the coherent experience of perception. In typical resonance, larger chunks (e.g., words) mask smaller chunks (e.g., phonemes). This allows ART to maximize information transmission, as more predictive attractors assume priority. Consider the spondee JIGSAW: In processing, many potential subdivisions will approach local resonance, including individual phonemes (/dʒ/) and syllables (/dʒIg/). In this case, both syllables also form words, which are natural units of perceptual organization. Nevertheless, common experience is hearing the entire word, not a loose collection of parts.

As explained by Grossberg and Myers (2000; also Boardman, Grossberg, Myers, & Cohen, 1999), temporally distributed signals are resolved in ART as resonant waves, capable of integrating information backward and forward in time. This is required by coarticulatory cues in speech, and successfully reproduces trading relations, cue integration, and similar phenomena. Simulations show that ART naturally exhibits "delayed commitment" in speech perception (Bard, Shillcock, & Altmann, 1988). ART also naturally resolves ambiguity: Just as phonemes and syllables are masked by word-level dynamics, ambiguous words will evoke multiple local resonances that are masked by global, contextually coherent states (Gottlob, Goldinger, Stone, & Van Orden, 1999). As Swinney (1979) showed, such transient states can be detected indirectly, but are absent from normal experience. This empirical profile clearly resembles that of phoneme or syllable perception.

(3) When bottom-up and top-down knowledge sources achieve resonance, *attention* is drawn, creating conscious experience. Grossberg (1980) stated that "…adaptive resonances are the functional units of cognitive coding", and they are the basis for episodic memory. Although resonances naturally draw attention, ART also allows selectivity. If a specific stimulus is anticipated (as in phoneme monitoring), relevant top-down structures are preactivated. This greatly accelerates resonance, even with noisy bottom-up data, as in phoneme restoration (Samuel & Ressler, 1986). Also, because of competitive dynamics, such expectancies may impede or destabilize resonances for unexpected stimuli.

3. Implications of adaptive resonance

Taking these characteristics together, ART provides an elegant, unifying account of prior monitoring data (and many indirect procedures). In this section, we combine the principles sketched above, allowing natural predictions to emerge. In some cases, the predictions are self-evident and data already stand in support. In other cases, the theory suggests that accurate prediction will be quite difficult, and the literature agrees. We then briefly summarize five recent experiments (Goldinger & Azuma, 2003), demonstrating that symmetric constraints jointly affect the apparent primacy of different speech units. In no specific order, ART carries at least three implications for formal assessments of speech units:

(1) The *creation* of reality in monitoring experiments. As reviewed, research on speech units began with phoneme- and syllable-monitoring experiments. These methods remain, and similar procedures have been developed, such as spotting words embedded in longer words (McQueen et al., 1994). However, such procedures have been criticized because the apparent units change

with selective attention. However, from an ART perspective, attention does not undermine monitoring data. According to ART, McNeill and Lindig (1973) were absolutely correct: By choosing perceptual targets ahead of time, listeners can facilitate resonance. Thus, monitoring participants can effectively manifest the "reality" of any unit, as early studies showed. However, where McNeill and Lindig saw an indictment of monitoring procedures, ART sees a special case of normal functioning.

(2) Continuing on the topic of monitoring tasks, ART makes two predictions that have broad support in the literature. The first (easier) prediction is that, all things being equal, larger units will typically "win". Because syllables typically mask their constituent phonemes, RTs are faster to syllables than to phonemes (Savin & Bever, 1970). In similar fashion, words mask syllables, sentences mask words, etc. These masks are never absolute: People can reconstruct sentences from memory. However, masking fields typically keep attention at the most functional levels of analysis. Thus, people are equally capable of monitoring for syllables or phonemes, but phonemes will typically be disadvantaged in RTs (although this pattern changes with carefully selected foils; Norris & Cutler, 1988). This leads to the second (harder) prediction:

(3) Because all things are *not* always equal, different potential units will occasionally "win" in virtually any experimental procedure. As described, adaptive resonance requires bottom-up and top-down knowledge sources to coalesce. However, there is no a priori requirement of equal contributions (which would severely limit perceptual ability). Processing in ART is self-optimizing, allowing rapid coherence across variable situations (Grossberg, 1980). Strong bottom-up information can support resonance with minimal top-down matching, as when clearly spoken non-words are readily understood.[3] Conversely, distorted bottom-up signals are readily identified with top-down support, as in phoneme restoration (Samuel, 2001).

With respect to speech units, such flexible processes predict flexible results. It is well known that acoustic cues to phonemes often spread throughout syllables. Thus, bottom-up support for phonemes and syllables are often roughly equivalent. In ART, processing will typically favor larger units. However, there must be occasions when phonemes carry strong, independent cues, as in temporally extended fricatives or prevoiced stops. These segments may "win the race" against their own carrier syllables (Healy & Cutting, 1976; Swinney & Prather, 1980). Alternatively, experimenters may select priming relations, helping specific segments achieve resonance before carrier syllables can exert masking effects. Conversely, experimenters may create top-down expectancies favoring one unit over another. Many combinations of tasks and stimuli could be fashioned to tilt the syllable-phoneme balance back and forth.

4. Pushing people around: the social psychology of speech units

To demonstrate symmetric influences on unitization, we briefly summarize five recent experiments that contrasted phoneme and syllable monitoring.[4] Our goal was not to determine which unit is perceptually favored. Instead, we sought to establish a baseline pattern, then push

[3] Notably, after just a few repetitions, nonwords show evidence of top-down matching, a laboratory example of unitization (Feustel, Shiffrin, & Salasoo, 1983).

[4] The experiments summarized here are part of a more comprehensive study (Goldinger & Azuma, 2003). Full details are not appropriate for this special issue, but will be available in our forthcoming manuscript, or upon request.

the result in both directions. We set ourselves a strict requirement: To test the ART symmetry hypothesis, we required separate bottom-up and top-down manipulations. This ideal is violated by many procedures, such as priming, wherein both knowledge sources are affected at once (Goldinger, Luce, Pisoni, & Marcario, 1992). In fact, we admit that true independence may be impossible. For example, imagine boosting the energy of specific segments for a phoneme-monitoring test. Although this is a bottom-up manipulation, participants may quickly develop implicit or explicit expectancies. Despite this difficulty, progress may be achieved by *juxtaposing* one manipulation against another. This was our strategy in Experiments 1–3.

In Experiment 1, participants listened to two lists of 120 bisyllabic non-words, with each list divided into 12 sub-lists of 10 non-words. Before each block of 10 trials, a target (phoneme or syllable) was shown on computer, and remained visible throughout all 10 trials. Twelve different targets were thus specified per list. Each list contained 20 target items (requiring a "yes" response), always with target sounds in initial position. To lend variation to the stimuli, eight sub-lists each contained two targets; the remaining four sub-lists each contained one target.

Across participants, the same items were equally used in phoneme and syllable monitoring. However, non-target (foil) stimuli were arranged differently, depending upon task. Specifically, each sub-list was arranged such that phoneme targets had two near neighbors (differing only in place or manner). For example, the target /b/ was mixed with foils beginning with /d/ and /p/, and other more remote foils. In similar fashion, each syllable target was mixed with two initial-phoneme foils, syllables sharing initial phonemes with targets, but neither vowels nor codas. For example, the target /bʌg/ was mixed with foils beginning with /bæp/ and /biz/, and other more remote foils). To avoid unequal "yes" responses, foils were arranged differently for phoneme and syllable monitoring, so each task required exactly 20 positive responses.

Thirty students were tested individually in the experiment. In counterbalanced fashion, participants made both phoneme and syllable detections to the stimulus lists. The stimuli were recorded and edited by a naïve volunteer, and the experiment was conducted by a naïve research assistant. No instructions were shown on the computer: The assistant was responsible for explaining all procedures, to wit: Participants were shown visual cues, representing auditory targets. Upon hearing targets, they pressed a "yes" key as quickly as possible. Given the prior literature, we expected faster RTs to syllables than phonemes.

In Experiment 2, most aspects were the same, except for the stimulus recordings. In this case, four students were separately approached (by the first author) with a very effective pick-up line: "*You know, you have a great voice. I would love to record you for an experiment on speech perception*". When the volunteers came to the laboratory, they were told different background stories about the experiment (see Intons-Peterson, 1983). All were told a simplified story about the "syllable-phoneme" question (motivated in terms of speech-recognition machines), and that our new method apparently solves the long-standing issue. In fact, "*our manuscript got positive reviews, but we need one more control experiment before it gets published in a major journal…*"

Beyond this point, the stories differed: Two volunteers learned that our data implicated *phonemes* as the fundamental building-blocks of speech. We presented this as the intuitive outcome, noting that phonemes resemble letters, noting minimal word pairs, etc. The other two volunteers learned that our data implicated *syllables* as fundamental units. This was presented as the intuitive outcome, emphasizing that speech is rhythmic, noting that children can count syllables but cannot count phonemes, etc. In every case, the helpful background talk ended with a

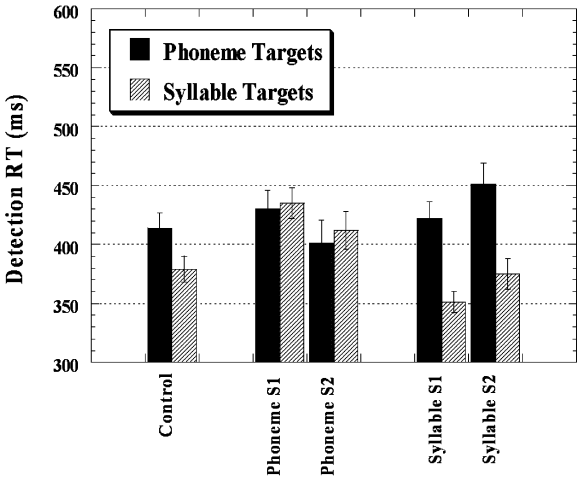

Fig. 1. Phoneme and syllable detection times in Experiment 1 (labeled "control") and Experiment 2. The label "Phoneme S1" refers to the condition with tokens recorded by phoneme-biased speaker 1, and so forth. Standard errors of means are shown.

reminder to *"take your time, stay about 6 inches away from the microphone, and produce your phonemes (syllables) clearly."*

After the desired outcome was made abundantly clear, each volunteer recorded the same non-words used in Experiment 1. The same research assistants (who remained naïve) edited the stimuli and conducted the listening experiments, as before. Experiment 2 consisted of four sub-experiments (one per speaker), with 30 participants apiece. The results of Experiments 1 and 2 are shown in Fig. 1.

In all experiments, error rates were very low (below 3%); only RT data will be discussed. As shown by the left-most bars, Experiment 1 ("control") reproduced the familiar pattern from Savin and Bever (1970). Initial syllables were detected 36 ms faster than initial phonemes [$F(1,29) = 16.8$, $p < 0.01$].[5]

The remainder of Fig. 1 shows separate data for each speaker in Experiment 2. The effect of biasing the stimulus recorders was clear. When speakers believed that phonemes should be more fundamental units, their tokens apparently encouraged such an effect: The baseline syllable advantage was reversed in those conditions. Although neither phoneme advantage was reliable (condition "phoneme S2" was $p = 0.057$), their interactions with the control condition were robust [S1: $F(1,58) = 10.7$, $p < 0.01$; S2: $F(1,58) = 13.1$, $p < 0.01$]. As shown, both syllable-biased speakers produced tokens encouraging large syllable advantages [S1: $F(1,29) = 40.2$, $p < 0.01$; S2: $F(1,29) = 53.9$, $p < 0.01$]. These conditions also produced significant interactions when contrasted to control [S1: $F(1,58) = 19.5$, $p < 0.01$; S2: $F(1,58) = 27.0$, $p < 0.01$]. The interactions within Experiment 2 (comparing Unit X Bias) were obviously reliable as well.

The results of Experiment 2 were striking, showing that motivated speakers can selectively handicap the syllable-phoneme race. Listening to the tokens revealed fairly systematic differences:

[5] Although we only report analyses by participants, all reported effects were reliable by items.

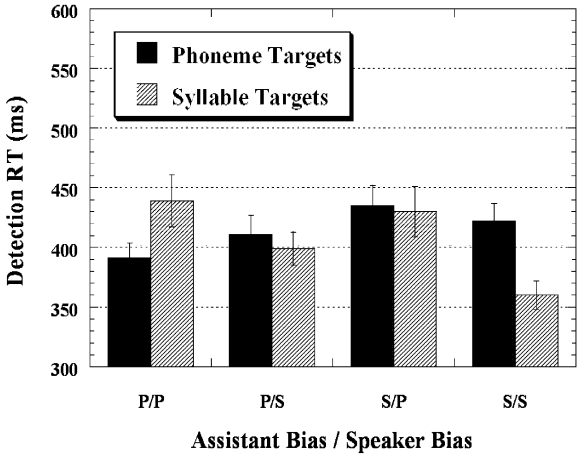

Fig. 2. Phoneme and syllable detection times in Experiment 3. The labels on the X-axis refer to the biases imposed on research assistants and token recorders (e.g., "P/S" indicates that the assistant was biased toward phonemes, and the speaker was biased toward syllables). Standard errors of means are shown.

Phoneme-biased speakers tended to talk more slowly, to use prevoicing, to lengthen fricatives, and to accentuate stop releases. Syllable-biased speakers tended to exaggerate prosodic cues, and to "swallow" initial consonants (i.e., they were typically short, casual in manner of articulation, and spoken with low intensity). As a final step, the experimental ruse was described to 24 students, who then listened to tokens from each speaker (all tokens were used equally across listeners). After 25 tokens, they tried to guess which hypothesis, phoneme or syllable, the speaker believed. Accuracy rates for the four speakers were 88%, 91%, 92%, and 91%.

The manipulation in Experiment 2 was intended to selectively boost phonemes or syllables from the bottom-up. However, people may develop hypotheses during testing, making the manipulation partly top-down. Although this is not easily assessed, we can contrast the putative bottom-up manipulation with an obvious top-down manipulation. In Experiment 3, the materials and procedures from Experiment 2 were used again; the only change involved the *assistants* who collected the data. Four assistants were recruited, all in different semesters. Each received upper-division research credit, and was motivated to produce "good" data. When the assistants received laboratory training, the critical variable was our stated hypothesis. The stories from Experiment 2 were used again. Finally, each assistant ran an experiment with 30 listeners, as before. For two assistants, experimental tokens came from speakers who had received the same biasing instructions. For the other two, tokens came from speakers with the opposite bias.

The results (see Fig. 2) suggest a blending of bottom-up and top-down influences. When both the research assistant and stimulus recorder were biased toward phonemes, a significant phoneme advantage emerged [$F(1,29) = 28.8$, $p < 0.01$]. When biases conflicted in either direction, no reliable differences emerged. When both the assistant and stimulus recorder were biased toward syllables, a large syllable advantage emerged [$F(1,29) = 64.1$, $p < 0.01$]. Together with Experiment 2, the data suggest that monitoring results can be influenced bi-directionally. Given such a fluid perceptual system, an inconsistent empirical literature naturally follows.

5. Episodic influences on psychological reality

As our review suggests, ART provides an attractive theory of speech perception, combining the robust behavior of interactive models with mechanisms for self-organization. Among those mechanisms, top-down matching is especially interesting, as it offers a connection to *episodic memory* effects in perception. Many recent studies suggest that detailed, episodic traces are created in speech perception (Bradlow, Nygaard, & Pisoni, 1999; Church & Schacter, 1994; Sheffert, 1998). Moreover, those traces seemingly affect later perceptual events (Goldinger, 1996; Nygaard & Pisoni, 1998).

Goldinger (1998) proposed that the mental lexicon may be a collection of detailed memory traces, rather than abstract units. To illustrate episodic perception, he applied a pure exemplar model, Hintzman's (1988) MINERVA 2, to lexical access. For every known word, a potentially vast collection of partially redundant traces resides in memory. These traces encode conceptual, perceptual, and contextual details of their original encoding events. Given a stimulus word, an *analog probe* activates all traces in parallel, each in proportion to their mutual similarity. The weighted average of activated traces constitutes an *echo* sent to "consciousness" from long-term memory. By virtue of past traces, echoes contain information not present in the probe (such as word meanings), associating new stimuli to prior knowledge. By separately testing echo intensity and content, Goldinger (1998) found that MINERVA 2 qualitatively simulated word-shadowing data, including both RTs and the sounds of people's voices.

In its core processes, MINERVA 2 shares common ground with ART: Both are predicated on resonant dynamics that pass between working memory and long-term memory. However, MINERVA 2 is quite limited, relative to ART. For example, MINERVA 2 has no "front-end" perceptual mechanisms; it merely assumes feature-vector representations. This simplification is helpful when testing core processes, but a more complete theory is required. In this regard, ART is especially promising: Its bottom-up processes are well articulated, and its top-down processes are completely compatible with episodic influences. Moreover, it provides natural mechanisms to *create* detailed memory traces. In ART, raw features and top-down chunks enter into resonance, drawing attention and enabling memory encoding. By avoiding hypotheses about explicit "word units", ART allows every episode of word perception to be a little different. All activated features, perceptual and conceptual, will shape the eventual resonance.

In brief, ART is a good candidate theory to move beyond MINERVA 2, combining perceptual processes with a potential for episodic effects. Experiments 4 and 5 tested episodic memory effects in phoneme and syllable monitoring. As in prior studies (Goldinger, 1996), the method involved presenting and later repeating stimuli. To gauge effects of specific episodes, repetitions sometimes changed in voice, or in their required responses. In Experiment 4, the procedure included study and test phases, separated by 5 min. During study, participants identified 80 bisyllabic non-words, clicking labeled boxes on the computer screen (four choices per trial). The non-words were evenly divided between a female and male voice (randomly presented within sessions, counterbalanced across participants). This was intended to familiarize listeners with voice-specific tokens, and to have non-words processed holistically. According to ART, this should accelerate global resonances for repeated non-words, especially when voice cues match between study and test. If so, this should increase top-down masking, increasing the asymmetry between phonemes and syllables.

For simplicity, separate phoneme- and syllable-monitoring tests were conducted, with 60 participants each. The test materials were 40 study tokens (same voice), 40 different voice tokens, and 40 new non-words. Twenty-four non-words were selected as targets, with eight per condition. The assignment of non-words to conditions was counterbalanced, allowing collection of baseline RTs and repetition effects for all targets. In other regards, the procedures resembled those of Experiment 1.

The results (see Fig. 3) contained expected and unexpected findings. In baseline trials, a 30-ms syllable advantage was observed [$F(1,119) = 9.5$, $p < 0.01$]. Given same-voice repetitions, this difference nearly tripled to 87 ms [$F(1,119) = 33.8$, $p < 0.01$]. The increased asymmetry was expected, due to a presumed acceleration of global resonances in same-voice trials. This may also explain the (unexpected) slowdown among same-voice trials, which were 73 ms slower than baseline trials [$F(2,118) = 29.9$, $p < 0.01$]. Finally, because different-voice trials present less familiar bottom-up cues, they should diminish the masking potential of study episodes. Indeed, the 43-ms syllable advantage was reliable, but was statistically equivalent to the baseline difference, and was smaller than the same-voice difference [$F(2,118) = 14.0$, $p < 0.01$].

Experiment 4 showed one direction of episodic influence: Globally coherent episodes seemingly imposed masking effects on smaller potential units. This pattern is not easily classified as "bottom-up" or "top-down". Instead, specific bottom-up inputs selectively activate specific top-down traces, a purely bidirectional effect. In Experiment 5, we further assessed the symmetry of "episodic perception" by adding a more blatant top-down manipulation. Whereas Experiment 4 involved non-word identification for the study phase, Experiment 5 involved monitoring tasks for both study and test. Our goal was to have participants encode voice-specific tokens, and to also encode specific *responses* for each stimulus. Previous research suggests that task-specific responses can greatly affect episodic traces, as estimated by visual repetition priming (Goldinger, Azuma, Kleider, & Holmes, 2002). In Experiment 5, participants first completed 60 phoneme-monitoring trials and 60 syllable-monitoring trials (Round 1). As before, non-words were presented in two voices. After Round 1, participants completed Round 2, with everything repeated. This allowed assessment of repetition priming under ideal conditions, and presumably increased the episodic encoding of each token-response combination.

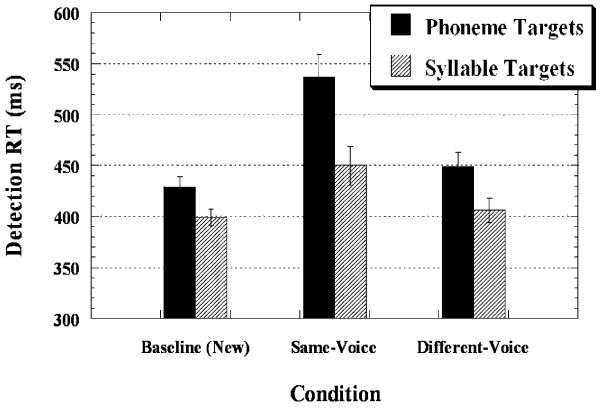

Fig. 3. Phoneme and syllable detection times in Experiment 4, shown as a function of item repetition status (new non-words, old same-voice, old new-voice). Standard errors of means are shown.

In Round 3, participants again performed phoneme and syllable monitoring, but something changed. Four between-subject conditions were tested (40 students each), with factorial changes in voices and tasks. In one condition (same voice, same task), unchanged stimuli were presented in the same tasks yet again. In a second condition (different voice, same task), the same nominal stimuli were presented in each task, but their voices were switched, relative to the preceding rounds. In a third condition (same voice, different task), all items were presented in their original voices, but the previous phoneme-monitoring tokens were switched to syllable monitoring, and vice versa. This task switch is easily characterized as a top-down manipulation, logically separate from the auditory input. In the fourth condition (different voice, different task), all items from previous rounds were repeated, with both voices and tasks changed.

Results for all conditions are shown in Fig. 4 (with data from Rounds 1 and 2 combined across groups). In Round 1, the results resembled previous findings, with a reliable syllable advantage [$F(1,159) = 21.6$, $p<0.01$]. RTs in Round 2 were faster than Round 1 [$F(1,159) = 48.8$, $p<0.01$], and the syllable advantage was somewhat reduced. Results for Round 3 varied across conditions: When prior episodes were repeated yet again (SV-ST), RTs were faster than Round 2, and the syllable advantage vanished, possibly due to a floor effect. The remaining conditions showed disruptions of this fluent performance. In the ST-DV condition, RTs slowed down, although not to the degree seen in Experiment 4. In this case, RTs were faster than baseline [$F(1,39) = 30.2$, $p<0.01$], and were equivalent to those in Round 2. However, as in Experiment 4, a significant syllable advantage [$F(1,39) = 8.6$, $p<0.05$], returned in these trials.

In the DT-SV group, a large slowdown occurred, with RTs far exceeding baseline [$F(1,39) = 85.9$, $p<0.01$]. Hearing perfect matches to previous tokens is apparently helpful when task requirements are constant, but is quite disruptive when requirements change (Goldinger et al., 2002). The bottom-up and top-down contributions to this effect are seen by comparison with the DT-DV condition: In this case, RTs were reliably slower than the ST-SV condition [$F(1,79) = 31.8$, $p<0.01$], but were equivalent to baseline, and were significantly faster than the DT-SV condition [$F(1,79) = 59.4$, $p<0.01$].

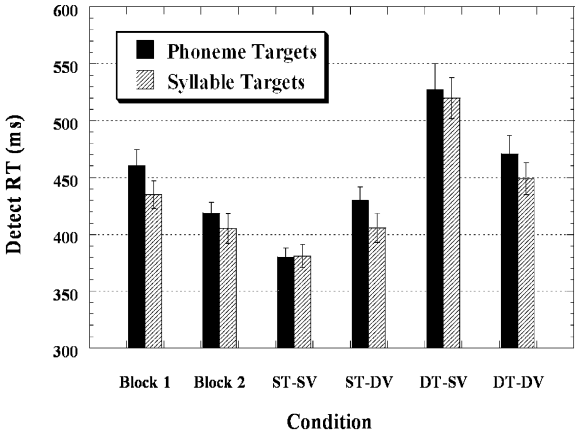

Fig. 4. Phoneme and syllable detection times in Experiment 5, shown by condition. "ST-SV" denotes "same task-same voice", and so forth. Standard errors of means are shown.

Experiment 5 further suggests that monitoring performance reflects bottom-up and top-down constraints. It also suggests that episodic traces can mediate top-down matching: When aspects of prior experiences are changed, fluency is reduced. Contrasting all groups, task changes were more disruptive than voice changes. Of particular interest, same-voice tokens *exacerbated* the difficulty imposed by different-task trials. By our account, their familiar bottom-up cues facilitate resonance with study episodes, including their associated responses. More generally, it seems that monitoring follows the principles of adaptive resonance: Factors that should facilitate or inhibit resonance have parallel effects on monitoring RTs.

6. Conclusion: adaptive resonance as a unifying principle

We began by noting Kuhn's (1962) view that *puzzle-solving* is a central activity of normal science. Kuhn emphasized that theoretical gravity has little bearing on good puzzles, which need only be challenging and ultimately solvable. Regarding speech units, there is a long-standing puzzle with little apparent progress toward a solution. As other authors have argued, this suggests that the original question—what is *the* fundamental unit?—is misguided.

As an alternative, we outlined basic principles of Grossberg's (1980) ART, with special focus on its resolution of the speech-unit hypothesis. By re-casting "units" as self-organizing dynamic states, ART both nullifies the question and helps connect years of speech-unit data to broader cognitive theory. The principles of adaptive resonance have proven very versatile, providing cogent accounts of visual perception, learning and memory, attention, decision making, and other cognitive domains (Carpenter & Grossberg, 1991). Grossberg et al. (1997) specifically modeled speech perception, but the principles of adaptive resonance naturally unify speech perception with many related areas, such as spoken and printed word perception, sentence processing, and episodic memory. Similar enthusiasm is often expressed for connectionist models, which conceptually derive from adaptive resonance (Rumelhart & McClelland, 1986).

Regarding units of speech perception, ART organizes prior data, predicts new results, and helps resolve a long-standing puzzle. Beyond speech units, we believe the same resonance mechanisms can explain repetition blindness (Harris & Morris, 2001) and deafness (Miller & McKay, 1996). Similarly, ART seems to predict an illusion discovered by Frost, Repp, and Katz (1988), wherein the simultaneous presentation of printed words and envelope-shaped noise gives the impression that speech was actually heard. By using competitive dynamics to create coherent experience, ART may account for McGurk effects (McGurk & MacDonald, 1976), and their varying strengths across different consonants. Finally, ART may explain how normal perceptual processes allow fluent interpretation of sinewave speech, including the provision that people perform better with just a little background knowledge (Remez, Rubin, Pisoni, & Carrell, 1981; Remez, 2003 [this volume]). We look forward to testing resonance principles in these and related domains.

Acknowledgements

This research was supported by grant R01-DC04535-03 from the National Institute of Deafness and Communicative Disorders (NIH). We thank Paul Luce, Greg Stone, and Guy Van Orden for many helpful discussions on the topics of resonance and speech units.

References

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, *44*, 395–408.

Boardman, I., Grossberg, S., Myers, C., & Cohen, M. (1999). Neural dynamics of perceptual order and context effects for variable speech-rate syllables. *Perception & Psychophysics*, *61*, 1477–1500.

Bradlow, A., Nygaard, L., & Pisoni, D. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*, 206–219.

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

Church, B., & Schacter, D. (1994). Perceptual specificity of auditory priming: Memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 521–533.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, *20*, 55–60.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.

Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, *33*, 824–844.

Decoene, S. (1993). Testing the speech unit hypothesis with the primed matching task: Phoneme categories are perceptually basic. *Perception & Psychophysics*, *53*, 601–616.

Feustel, T., Shiffrin, R., & Salasoo, A. (1983). Episodic and lexical contributions to the repetition effect in word recognition. *Journal of Experimental Psychology: General*, *112*, 309–346.

Foss, D. J., & Swinney, D. A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, *12*, 246–257.

Frost, R., Repp, B. H., & Katz, L. (1988). Can speech perception be influenced by simultaneous presentation of print? *Journal of Memory and Language*, *27*, 741–755.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Goldinger, S. D., & Azuma, T. (2003). Changing priorities among speech units: symmetric bottom-up and top-down effects, manuscript in preparation.

Goldinger, S. D., Azuma, T., Kleider, H. M., & Holmes, V. (2002). Font-specific memory: More than meets the eye? In J. S. Bowers, & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 157–196). Oxford: Oxford University Press.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1210–1238.

Gottlob, L. R., Goldinger, S. D., Stone, G. O., & Van Orden, G. C. (1999). Reading homographs: Orthographic, phonologic, and semantic dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 561–574.

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

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

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*(3/4), doi:10.1016/S0095-4470(03)00051-2.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 483–503.

Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review*, *107*, 735–767.

Harris, C. L., & Morris, A. L. (2001). Illusory words created by repetition blindness: A technique for probing sublexical representations. *Psychonomic Bulletin & Review*, *8*, 118–126.

Healy, A. E. (1994). Letter detection: A window to unitization and other cognitive processes in reading text. *Psychonomic Bulletin & Review*, *3*, 333–344.

Healy, A. E., & Cutting, J. (1976). Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior*, *15*, 73–83.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Hintzman, D. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Intons-Peterson, M. J. (1983). Imagery paradigms: How vulnerable are they to experimenters' expectations? *Journal of Experimental Psychology: Human Perception & Performance*, *9*, 394–412.

Jusczyk, P. W. (1986). A review of speech perception research. In L. Kaufman, J. Thomas, & K. Boff (Eds.), *Handbook of perception and performance* (pp. 27–57). New York: Wiley.

Kessler, B., & Treiman, R. (2001). Relationship between sounds and letters in English monosyllables. *Journal of Memory and Language*, *44*, 592–617.

Kolinsky, R., Morais, J., & Cluytens, M. (1995). Intermediate representations in spoken word recognition: Evidence from word illusions. *Journal of Memory and Language*, *34*, 19–40.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, *101*, 653–675.

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, *79*, 124–145.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McNeill, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. *Journal of Verbal Learning and Verbal Behavior*, *12*, 419–430.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 621–638.

Mehler, J., Segui, J., & Frauenfelder, U. (1981). The role of the syllable in language acquisition and perception. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 295–305). Amsterdam: North-Holland.

Miller, M., & McKay, D. (1996). Relations between language and memory: The case of repetition deafness. *Psychological Science*, *7*, 347–351.

Mills, C. B. (1980). Effects of the match between listener expectancies and coarticulatory cues on the perception of speech. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 528–535.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*, 3241–3254.

Norris, D., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics*, *43*, 541–550.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral & Brain Sciences*, *23*, 299–325.

Nygaard, L., & Pisoni, D. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*, 355–376.

Plaut, D. C., McClelland, J. L., & Seidenberg, M. S. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Plomp, R. (2002). *The intelligent ear: on the nature of sound perception*. Hillsdale, NJ: Erlbaum.

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

Remez, R.E. (2003). Establishing and maintaining perceptual coherence: Unimodal and multimodal evidence. *Journal of Phonetics*, *31*(3/4), doi:10.1016/S0095-4470(03)00042-1.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–950.

Rubin, P., Turvey, M. T., & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics*, *19*, 394–398.

Rumelhart, D., & McClelland, J. L. (Eds.) (1986). *Parallel distributed processing*. Cambridge: MIT Press.

Samuel, A. (1989). Insights from a failure of selective adaptation: Syllable-initial and syllable-final consonants are different. *Perception & Psychophysics*, *45*, 485–493.

Samuel, A. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*, 348–351.

Samuel, A. G., & Ressler, W. H. (1986). Attention within auditory word perception: Insights from the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 70–79.

Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, *9*, 295–302.

Sebastian-Galles, N., Dupoux, E., Segui, J., & Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language*, *31*, 18–32.

Shand, M. (1976). Syllabic vs. segmental perception: On the inability to ignore "irrelevant" stimulus parameters. *Perception & Psychophysics*, *20*, 430–432.

Sheffert, S. (1998). Voice-specificity effects on auditory word priming. *Memory & Cognition*, *26*, 591–598.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–659.

Swinney, D. A., & Prather, P. (1980). Phonemic identification in a phoneme monitoring experiment: The variable role of uncertainty about vowel contexts. *Perception & Psychophysics*, *27*, 104–110.

Tabossi, P., Collina, S., Mazzetti, M., & Zoppello, M. (2000). Syllables in the processing of spoken Italian. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 758–775.

Van Orden, G. C., & Goldinger, S. D. (1994). Interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1269–1291.

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

Warren, R. M. (1971). Identification times for phonemic components of graded complexity and for spelling of speech. *Perception & Psychophysics*, *9*, 345–349.