Phoneme restoration and empirical coverage of interactive activation and adaptive resonance models of human speech processing

James S. Magnuson[a)]
Department of Psychology, University of Connecticut, Storrs, Connecticut 06269

(Received 26 October 2014; revised 17 November 2014; accepted 25 November 2014)

Grossberg and Kazerounian [(2011). J. Acoust. Soc. Am. **130**, 440–460] present a model of sequence representation for spoken word recognition, the cARTWORD model, which simulates essential aspects of phoneme restoration. Grossberg and Kazerounian also include simulations with the TRACE model presented by McClelland and Elman [(1986). Cognit. Psychol. **18**, 1–86] that seem to indicate that TRACE cannot simulate phoneme restoration. Grossberg and Kazerounian also claim cARTWORD should be preferred to TRACE because of TRACE's implausible approach to sequence representation (reduplication of time-specific units) and use of non-modulatory feedback (i.e., without position-specific bottom-up support). This paper responds to Grossberg and Kazerounian first with TRACE simulations that account for phoneme restoration when appropriately constructed noise is used (and with minor changes to TRACE phoneme definitions), then reviews the case for reduplicated units and feedback as implemented in TRACE, as well as TRACE's broad and deep coverage of empirical data. Finally, it is argued that cARTWORD is not comparable to TRACE because cARTWORD cannot represent sequences with repeated elements, has only been implemented with small phoneme and lexical inventories, and has been applied to only one phenomenon (phoneme restoration). Without evidence that cARTWORD captures a similar range and detail of human spoken language processing as alternative models, it is premature to prefer cARTWORD to TRACE. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4904543]

[ADP] Pages: 1481–1492

I. INTRODUCTION

Grossberg and Kazerounian (2011) present a computational model of word recognition, the cARTWORD model, and claim that it solves the problem of representing temporal order for speech-like inputs. They demonstrate that it can account for the basic phenomenon of phoneme restoration (reviewed in some detail below), and present simulations appearing to demonstrate that TRACE, an interactive activation model of speech and word recognition (McClelland and Elman, 1986), cannot. In this paper, I review the problem of representing temporal order, rebut criticisms Grossberg and Kazerounian (2011) make of the TRACE architecture, point out fundamental aspects of spoken word recognition that cARTWORD appears not to account for, and demonstrate that when proper noise stimuli are used (along with minor changes to TRACE phoneme definitions), the TRACE model provides a robust basis for phoneme restoration, and that TRACE accounts for additional aspects of phoneme restoration cARTWORD was not tested on.

II. REPRESENTING TEMPORAL ORDER FOR SPEECH AND THE PLAUSIBILITY OF TRACE

Representing ordered sequences is a fundamental problem in neuroscience, and is particularly salient in the case of speech. Some orderings of the phonemes /k/, /æ/, /t/ result in words—/kæt/, /tæk/, /ækt/ (CAT, TACK, and ACT)—while others do not (/ktæ/, /tkæ/, /ætk/). The same is true at the word level; DOG CHASES CAT is different from CAT CHASES DOG or *CHASES DOG CAT. Thus, models of speech processing must distinguish temporal orderings. Models must also distinguish repetitions of elements; the second /d/ in /dæd/ must be encoded as a second /d/ event, not just further evidence that /d/ has occurred. The same is true for word sequences, such as DOG EATS DOG.

Some computational approaches to speech processing in the psychological literature ignore temporal order (Luce and Pisoni, 1998; Norris *et al.*, 2000), or have not been tested in detail with (simplifications of) speech [such as simple recurrent networks (SRNs) (Elman, 1990, 1991, though see Gaskell and Marslen-Wilson, 1997 and Magnuson *et al.*, 2003)]. Only one model provides truly deep and broad coverage of phenomena in human speech perception and spoken word recognition while providing a basis for representing temporal order including repeated elements: the TRACE model (Elman and McClelland, 1986; McClelland, 1991; McClelland and Elman, 1986; Strauss *et al.*, 2007).

TRACE employs a specialized mechanism to represent serial order: a memory where units at each of its representational levels—features, phonemes, words—are reduplicated at successive time slices, providing a spatial medium for representing temporal sequences. Thus, there is a /d/ detector aligned with memory onset, and another every three time slices, to the end of the memory bank. Feature and word units

a)Author to whom correspondence should be addressed. Electronic mail: james.magnuson@uconn.edu

are similarly reduplicated. These independent detectors represent repetitions as independent events, allowing TRACE to represent complex words with phoneme repetitions, and to parse and recognize word sequences, including sequences with repeated words. McClelland and Elman (1986) themselves pointed out weaknesses of reduplication (pp. 76–78), but demonstrated that it allowed TRACE to achieve broad, deep coverage of phenomena in human speech perception and spoken word recognition.

Grossberg and Kazerounian critique several aspects of TRACE. First, they deride TRACE as "embody[ing] … general properties in a way that is inconsistent with basic properties of human speech perception, or indeed of any real-time physical model" (Grossberg and Kazerounian, 2011, p. 456). They also argue that TRACE is "not a real time model" and "does not include a plausible representation of time that can be used in any physical process," and that it is not clear how TRACE's representations could emerge in a learning system. McClelland *et al.* (2014) remind us that TRACE is not meant to provide a neural-level solution: "The structure of the TRACE model should not be viewed as a literal claim about the neural mechanism. Instead it provides as a way of capturing the relative rather than absolute constraints between phoneme and word level information: If there is a /k/ at a particular time, it supports the word 'cat' starting in the same time, and the word 'ticket' starting two phonemes earlier (among many other possibilities), and these constraints are captured in the connections between units for the corresponding items in the corresponding positions. Activation in this array of units formed a dynamic memory trace of the results of a spoken input, hence the name of the model."

That said, there are at least two ways that the seemingly implausible reduplication scheme of TRACE could be addressed. The first is with the programmable blackboard model, which McClelland (1986) designed explicitly to demonstrate how TRACE-like (interactive activation) computations could be achieved without reduplicated units. Second, consider a model of echoic memory based on a frequency-by-time matrix (1 to 4 s in duration, the approximate duration of echoic memory for speech; Connine *et al.*, 1991; Watkins and Watkins, 1980), with the simplifying assumption that time is discretized into steps. Now as auditory input is encountered at time 0, the time = 0 frequency vector would encode the input on position 0. At time = 1, the time = 0 vector would shift to position 1, and the new input would be encoded at slot 0 (and at some point, the system would have to "wrap," recycling position 0, etc.). It is a small step to imagine a similar memory where frequency vectors would be replaced or augmented by phonemic vectors, or some other phonetic or phonemic recoding. As each phoneme is processed,[1] the matrix corresponding to the previous phonological state could be shifted on the memory matrix, replaced with the current one aligned at slot 0. Thus, while TRACE can be fairly criticized for its reduplication mechanism,[2] it is not wildly implausible.

It is crucial to note that no other model comes close to simulating the breadth and depth of results that TRACE does with high accuracy. In addition to the more than one dozen diverse, central phenomena of speech perception and spoken word recognition simulated in the original paper (McClelland and Elman, 1986), TRACE simulates the time course of lexical activation and competition at a very fine grain (e.g., eye tracking data from a paradigm developed a decade after TRACE; Allopenna *et al.*, 1998; Dahan *et al.*, 2001a; Dahan *et al.*, 2001b; see Strauss *et al.*, 2007, for a summary of the phenomena TRACE simulated in the original paper, as well as later, fine-grained time course simulations conducted with TRACE).

Grossberg and Kazerounian (2011) also contend that TRACE's "silence" phoneme unit is problematic. What seems intuitively odd about the silence phoneme is that it is activated by a special input pattern. TRACE's feature level consists of vectors of 63 elements, comprised of seven acoustic-phonetic dimensions each with nine levels, eight of which represent different aspects of each dimension. The ninth level of each feature is reserved for representing silence. To activate the silence phoneme, the input consists of activations for the ninth level of each feature and zeros on the first eight levels of all features. The silence "phoneme" at the phonological level can inhibit all other phonemes when the silence pattern is encountered in the input, providing a qualitatively different model state when silence input is encountered. Because the silence phoneme activates the silence "word," similar inhibition can occur at the lexical level. The rationale of having silence lead to inhibition is not laid out explicitly in the original TRACE paper, but is consistent with the inclusion in interactive activation models in general of detectors that code *absence* of features, not just detectors that code presence of features (e.g., McClelland, 2013; McClelland and Rumelhart, 1981). The rationale follows from the hypothesis that relatively extended absence of input should lead to "recognition" of silence—that is, the system should arrive in a qualitatively different state reflecting absence of input, analogous to subjective experience when portions of edges are absent in a visual stimulus (McClelland, 2013; Warren, 1970).[3]

The special silence pattern in TRACE is not as artificial or *ad hoc* as it might seem. TRACE silence units code for the absence of input, like off-cells. This function might be achieved more realistically. For example, one could implement a bias node with a positive connection to an off/silence detector, with that input cancelled via inhibition whenever feature detectors receive sufficient activation. The off/silence unit would dominate in the absence of robust featural input (due to its constant input from the bias node), but otherwise would be inhibited; the silence phoneme in TRACE is simply a computational shortcut for achieving such a function. As we shall see in Sec. IV, whether one considers the silence phoneme or "true" silence (i.e., actual absence of input) as the appropriate baseline for evaluating phoneme restoration is immaterial, and so resolving this issue is not crucial here.

Grossberg and Kazerounian (2011) also take issue with TRACE's lack of absolute constraints on top-down feedback (specifically, they argue that top-down feedback must not be allowed in the absence of any bottom-up support). They cite a passage from McClelland and Elman (1986) (p. 75) where those authors *speculated* about how feedback might be used

in a learning variant of TRACE. Grossberg and Kazerounian (2011) argue that the mechanism outlined there would lead to unstable learning. This is tangentially pertinent; such a learning mechanism was not implemented, and someone developing a learning variant of TRACE could certainly take into account the issues Grossberg and Kazerounian (2011) raise and avoid unstable learning, *if* such instability were demonstrated to occur, and *if* the developer implemented a mechanism akin to the one McClelland and Elman proposed. But the real crux of the critique by Grossberg and Kazerounian (2011) on this issue is that they claim that the reason TRACE fails (in their simulations) to provide a basis for simulating human phoneme restoration is that it permits non-modulatory feedback (feedback in the absence of bottom-up support). In their view, non-modulatory feedback in TRACE allows greater restoration when a phone is replaced by "true" silence (rather than the silence phoneme) than when it is replaced by noise. As I demonstrate in Sec. IV, absence of modulatory control on feedback in TRACE has nothing to do with their simulation results; their results follow instead from a flawed approach to creating TRACE analogs of stimuli used with human subjects.

Finally, it is useful to review why McClelland and Elman adopted the reduplication approach despite noting the challenges to linking it to known neural mechanisms: it allowed the model to simultaneously (a) represent serial order of features, phonemes, and words, (b) represent/decode sequences including multiple instances of one phoneme (/dæd/) or word (DOG EATS DOG) including embeddings, and (c) account for a wide and broad range of phenomena in human speech perception and spoken word recognition. No other model comes close to the depth and breadth of TRACE's coverage (Magnuson *et al.*, 2012).

III. THE PLAUSIBILITY OF cARTWORD

Grossberg and Kazerounian (2011) present an alternative mechanism for sequence encoding: chunk-mediated gating. With this organization, a list chunk representation for a sequence of inputs such as ABC can be hard-wired so that the chunk is more strongly activated by that ordering of elements than another (e.g., CBA). A stipulated wiring scheme leads an ABC chunk to "expect" stronger input from earlier elements: A > B > C. A chunk for CBA would similarly expect (respond maximally to) C > B > A. Earlier elements achieve greater activation due to their ability to inhibit units responding to later arriving inputs. Grossberg and Kazerounian (2011) show that in combination with the adaptive resonance theory matching rule (which prevents a unit from receiving top-down support in the absence of bottom-up support; Carpenter and Grossberg, 1987), their model successfully simulates phoneme restoration when a phoneme is replaced with noise, and restoration failure when the phoneme is replaced with silence (the crucial pattern of results in human listeners; e.g., Samuel, 1981a,b, 1996, 1997). However, there are serious problems with this mechanism.

First and most crucially, cARTWORD can represent sequences, but *cannot* represent sequences that contain repeated elements; the framework cannot distinguish ABA from AB—the second A in ABA would just be more evidence for A having occurred. Additional evidence for B in an input pattern like ABCB might lead to equivalent activation of A and B, creating ambiguity as to the identity of the first element in the sequence. Were cARTWORD extended beyond its very small two-word "lexicon," it would not be able to tell /tu/ (too) from /tut/ (toot), or /IlIsIt/ (illicit) from the nonword /IlsItI/ ("ihl-sih-tih"). This rules out cARTWORD as a plausible model of sequence encoding for word recognition, as it cannot represent vast numbers of words [29% of lemmas in the Brown corpus (Kucera and Francis, 1967) include at least one repeated phoneme].

Grossberg and Kazerounian might counter that cARTWORD could be combined with the item-order-rank (IOR) model of working memory proposed in other domains (Silver *et al.*, 2012). The framework is laid out in Fig. 2 of Silver *et al.*; outputs from a bank of time-invariant nodes (one per element, not one per element per time slice) for each possible element (e.g., each phoneme, if extended to this domain) combine with outputs from counting (clock) cells to activate relative time-specific representations in the IOR memory. As shown in their figure, there is another IOR node for each possible element, but divided like pie slices for different relative temporal positions. For example, if only four time steps were tracked, there would be a /p/ IOR node with four slots; if /p/ occurs when the clock is in position 3, the /p/ IOR node would be activated at slice 3. This is an intriguing approach, but it is not clear how to extend it sufficiently for speech, where each IOR node might need to be capable of coding dozens of time steps (again, reflecting perhaps the duration of echoic memory). Furthermore, to handle noisy inputs and provide a basis for recovering from misparsings, substantial, time-specific inhibitory connectivity between IOR nodes would likely be required, as in TRACE. Indeed, extending the IOR approach in these ways would result in a mechanism for encoding temporal sequences not terribly different from that employed by TRACE, since each "pie slice" in an IOR phoneme node becomes a time-specific representation of that phoneme—exactly the reduplication problem cARTWORD claims to avoid. Similar problems necessarily arise at supra-phonemic levels of encoding (cARTWORD's list chunks and lexical nodes); repeated words will be as problematic as repeated phonemes, and the IOR framework must again be extended to include many time-specific representations for words.

Another concern with cARTWORD is that a very simple model (with five abstract "acoustic" detectors, and just two "words") has been applied to only one phenomenon: phoneme restoration. While Grossberg and Kazerounian (2011) provide a demonstration proof that cARTWORD handles the basic details of phoneme restoration, we have no idea whether the behavior of the model will remain robust if the model is extended with a larger phoneme inventory, more phonetically detailed representations, varied word length, a larger lexicon, etc. Whether the model's behavior will remain stable with expansion of several orders of magnitude is an empirical question. Furthermore, until such a scaled-up cARTWORD has been tested with the phenomena

from the original TRACE paper and the more recent time-course data TRACE accounts for, there is no basis for a valid comparison of the models. Demonstration proofs are insufficient; like other literatures, the speech literature is rife with examples of very plausible modeling predictions that do not hold up when simulations are actually conducted (Magnuson *et al.,* 2012). Grossberg and Kazerounian would do a tremendous service to the field by carrying out such simulations with a scaled-up version of cARTWORD incorporating any computational mechanisms they claim are necessary to account for spoken word recognition. In the meantime, TRACE simply provides vastly superior coverage of the empirical literature. I next consider the claims of Grossberg and Kazerounian that TRACE fails to provide a basis for phoneme restoration.

IV. COMPARING TRACE AND cARTWORD EMPIRICALLY

In their Sec. VI A, Grossberg and Kazerounian (2011) present TRACE simulations communicated to them by a reviewer, as well as their own follow-up simulations (all conducted with the jTRACE java reimplementation of TRACE, the source for which is not cited: Strauss *et al.,* 2007). Their reviewer's simulations compared activations of an /l/-unit aligned with the onset of "luxury" when /l/ was either intact, replaced by TRACE's "silence" phoneme, or replaced by a "noise phoneme," defined as all 63 feature values (including those meant to signal silence) replaced by a value of 0.4. As shown in Grossberg and Kazerounian (2011) Fig. 8, /l/ became more active when replaced by the noise phoneme than when replaced by the silence phoneme. Grossberg and Kazerounian (2011) argued that the silence phoneme is an implausible way to simulate silence (since it expects a special input pattern orthogonal to all other phonemic inputs, rather than absence of input). They replaced the silence phoneme with what they called a "true silence phoneme" (one with all features set to 0.0), and found that /l/ became more active given the true silence phoneme than given their noise phoneme (Grossberg and Kazerounian, 2011, Fig. 9a).

Grossberg and Kazerounian (2011) conclude that TRACE is falsified with respect to the phenomenon of phoneme restoration. However, any apparent model failure must be inspected to determine what type of failure it represents (Magnuson *et al.,* 2012). The least interesting possibility is a failure to construct a valid linking hypothesis between stimulus and task constraints imposed on human subjects and the model. Unfortunately, the Grossberg and Kazerounian (2011) failed TRACE simulations are of this sort. These problems can be remedied by better linking of TRACE noise to stimulus noise.

First, note that Grossberg and Kazerounian (2011) used a single level of "noise" in their TRACE simulations, but that noise was implemented as an actual phoneme—the 0.4 noise phoneme described above, added using the phoneme editor in jTRACE. This is a poor analog for noise inserted into a real sound file. Creating a new phoneme in jTRACE makes it a "full member" of the phonological level, with inhibitory links to all other phonemes. Thus, it is not at all surprising that such noise would lead to extremely low /l/ activation, since the noise input would very strongly activate the 0.4 noise phoneme and simultaneously drive other phoneme activations to very low levels via inhibition. Furthermore, this noise is not constant, but ramps on and off, like any TRACE phoneme. Consider the TRACE inputs used by Grossberg and Kazerounian (2011), plotted in Fig. 1. The top left panel shows "intact" *luxury,* with the initial /l/ in place. The top right panel shows *-uxury,* that is, /l/ replaced by the TRACE silence phoneme. The bottom left

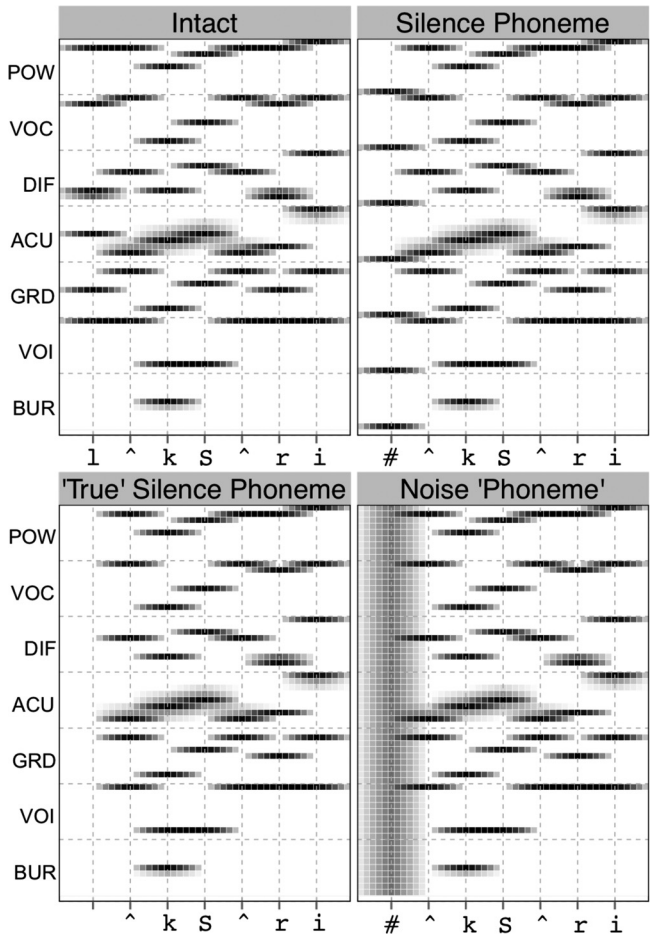

FIG. 1. Stimuli used in the TRACE simulations of "luxury" reported by Grossberg and Kazerounian (2011). The top left panel shows the original. The *x* axis is TRACE feature slices, with dashed horizontal lines indicating phoneme centers (occurring every six slices), and phoneme labels added for clarity. The *y* axis consists of the 63 rows of the TRACE input matrix; there are seven features [power, vocalic, diffuse, acute, gradual (labeled "consonantal" in the original paper, but labeled this way in the code), voicing, and burst], each of which has nine levels; horizontal dashed lines indicate boundaries between features. The two inputs on the top row are normal TRACE inputs: luxury with all phonemes intact, and luxury with the silence phoneme (/-/) substituted for /l/. The bottom two inputs were created by Grossberg and Kazerounian (2011) using the jTRACE phoneme editor. They created a "true silence" phoneme with all feature values set to 0. Note that this true silence does not disrupt the second phoneme; it just removes /l/ features. They also created a "noise phoneme", with a target pattern of 0.4 on all features. As can be seen in the figure, because this is a phoneme, it ramps on and off, and summates where it overlaps with the second phoneme. Problems with these true silence and noise phonemes are discussed in the text.

panel shows luxury with the first phoneme replaced with their true silence phoneme, and the bottom right panel shows /l/ replaced with their noise phoneme. Note that the TRACE silence phoneme pattern ramps on and off, as do the noise inputs in the bottom right panel. In the true silence phoneme (bottom left) used by Grossberg and Kazerounian (2011), there is nothing to ramp on and off, but note that the full /^/ phoneme in second position is intact. The key point is that the stimuli in Fig. 1 are extremely different from phoneme restoration stimuli used with human subjects, in which stretches of the speech waveform are excised (silence replacement) or replaced with noise. The ramping on and off of the noise phoneme in particular is radically different from real noise replacement stimuli.

We can create better analogs by actually replacing stretches of TRACE input with Gaussian noise or zeroes [true silence, but spliced in so that not just the /l/ is replaced, but so are temporally overlapping ("coarticulated") slices of adjacent phonemes]. Such materials are shown in Fig. 2. These materials were created by editing "feature files" exported from TRACE; they cannot be created using the phoneme editor in jTRACE. Noise portions were created as matrices of appropriate dimensions with Gaussian noise with mean set to the noise level and standard deviation set to 25% of the noise level (serving to constrain the noise values to be greater than zero; virtually identical results are observed if instead noise values are shifted positively to avoid negative values, or even when every feature is set to a constant value).

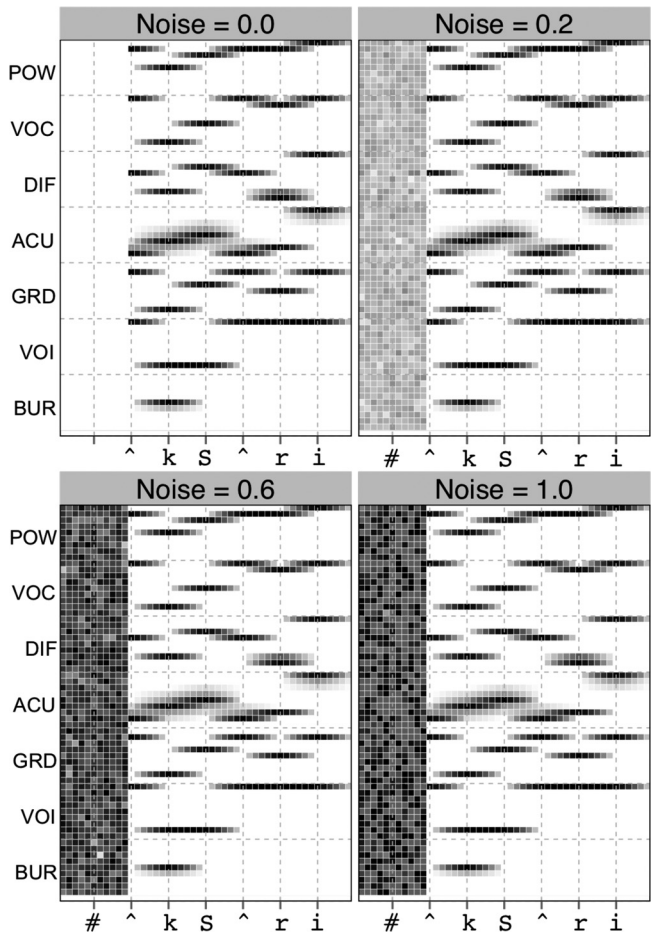

FIG. 2. Stimuli used in the current simulations; see Fig. 1 for axis details. The true silence input at top right (noise = 0.0), unlike the Grossberg and Kazerounian (2011) true silence phoneme, actually sets all bottom-up input values to 0 over the entire region previously spanned by /l/, including values corresponding to the first few slices of the first vowel. In the other panels, the /l/ region is replaced by Gaussian noise of increasing magnitude. Splicing in silence (zeroes) or noise results in TRACE stimuli that are appropriately analogous to materials used in human phoneme restoration experiments.

I conducted simulations with TRACE with each of these inputs (in the context of the original 212-word TRACE lexicon) for the word luxury with its initial (/l/), medial (/S/), or final (/i/) segment replaced. In each case, the word "luxury" was recognized. Activations of /l/ aligned at word onset given each of these stimuli are shown in the left panel of Fig. 3. Maximal activation is observed with intact /l/, less with true silence, and even less with the silence phoneme (indeed, inhibition from the silence phoneme drives /l/'s activation well below it is −0.10 resting level), as Grossberg and Kazerounian (2011) found. However, the results with noise-replaced stimuli are dramatically different. Grossberg and Kazerounian (2011) found that /l/ became more active given their noise *phoneme* than the silence phoneme (because the silence phoneme is orthogonal to all other core TRACE phonemes, but their noise phoneme overlapped in several features with /l/), but not more active than when /l/ was replaced with their true silence phoneme (again, because the noise *phoneme* would actively *inhibit* /l/), and, furthermore, the differences in their simulations mainly showed up in the very late time course.

One can see in Fig. 3 that /l/ exhibits a burst of transient activation in the *early* time course given noise replacement (rather than *noise phoneme* substitution). Activation given noise (noise > 0.0) is greater than activation given true silence (noise = 0.0). Noise levels of 0.6 or greater also lead to greater activation of /l/ in the later time course than do true silence or the silence phoneme. At all levels of non-zero noise shown, /l/ becomes robustly more active in the early time window (prior to cycle 25) than when /l/ is replaced with true silence.[4] The middle and right panels plot results when the medial (/S/) and final (/i/) segments of luxury are replaced, and we also see a substantial advantage for the noise-replaced phoneme early on (∼30 cycles after replacement onset). The early time window is the critical region for restoration; what matters is whether there is a basis for differential behavior *as the word is being experienced,* rather than many time steps after the noise (e.g., approximately 30 slices after replacement onset, where the 0.0 noise case catches up to the noise replacements).

However, a reviewer pointed out the need for additional checks that Grossberg and Kazerounian (2011) did not include in their jTRACE simulations. One must check whether replaced phonemes become more activated than other phonemes; that is, the early jump in activity as noise is presented might reflect uniform transient activation of all phonemes in response to noise. Similarly, one must also check whether activation in the *early time window* is guided by lexical feedback (the reviewer noted that the early window might be too early for lexical feedback to have had time

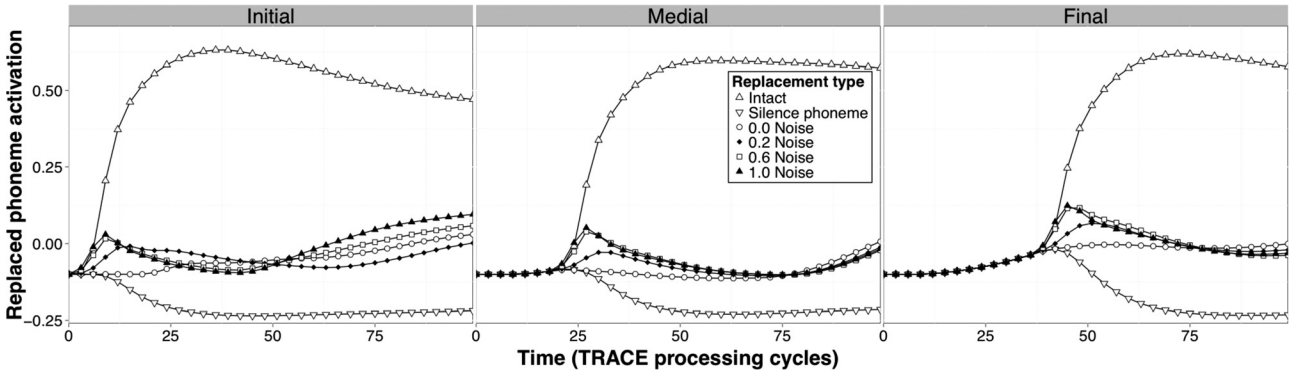

FIG. 3. Replaced phoneme activations for initial, medial, and final phonemes in "luxury" when replaced by the silence phoneme, true silence (0.0 noise), and varying levels of noise.

to impact phoneme activation). Figure 4 plots activations for position-specific phonemes for initial (/l/), medial (/S/) and final (/i/) phoneme replacements in luxury with noise set to 0.0 or 1.0. It is clear that there is a problem for initial and medial positions with 1.0 noise; /k/ and /g/ are most activated in the early time window, and the replaced phoneme only catches up later. For final position, we see a clear top-down effect, with /i/ pre-activated and most active throughout the simulation, although /k/ and /g/ become nearly as activated when the noise comes on. To determine whether the problematic results for initial and medial position were specific to luxury, I conducted additional simulations with all three-phoneme ($n = 73$), five-phoneme ($n = 31$), and seven-phoneme ($n = 8$) words in the TRACE lexicon. The results are shown in Fig. 5. Although the average responses in Fig. 5 appear to show greater restoration given noise than silence, in approximately 75% of simulations with 1.0 noise, the replaced phoneme was not the most active in the early time window (for nearly all initial and medial cases). In the majority of these cases, /k/ or /g/ was most active.

The top panel of Fig. 6 displays the original feature definitions of each phoneme in TRACE and provides a clue as to what is going on. Phonemes use different numbers of features, ranging from 7 to 12 of the 63 possible feature units. The more feature units that are "on" for a phoneme, the more similar it will be to broadband noise; and indeed, /k/ and /g/ have more features on than other phonemes. There is no theoretical commitment to the phoneme definitions in TRACE; the aim is for them to reflect similarity among actual phonemes. We can preserve phoneme similarity while equalizing similarity to noise with slight revisions, displayed in the lower panel of Fig. 6. Now each phoneme takes just one value for each feature, such that each has exactly seven features on. Hierarchical clustering solutions for original and modified phoneme definitions are virtually indistinguishable. I tested TRACE with these new features on the 12 core simulations from McClelland and Elman (1986), which are included in the "gallery" of simulations with jTRACE; there were negligible quantitative differences, but critical details of all simulations remained intact with the slightly revised phoneme definitions.

Next, I checked luxury again with the new feature definitions; the results are shown in Fig. 7 (replaced phoneme activations by position and noise level) and Fig. 8 (all position-specific phoneme activations for 0.0 and 1.0 noise for initial, medial, and final phoneme replacements). For initial and medial replacements, the replaced phoneme does not dominate initially in the early window, but the result for final position is very clear; /i/ is preactivated by top-down feedback, and when noise comes on, it is boosted more than other phonemes (which have virtually identical activations). Why does the replaced phoneme not similarly dominate in initial

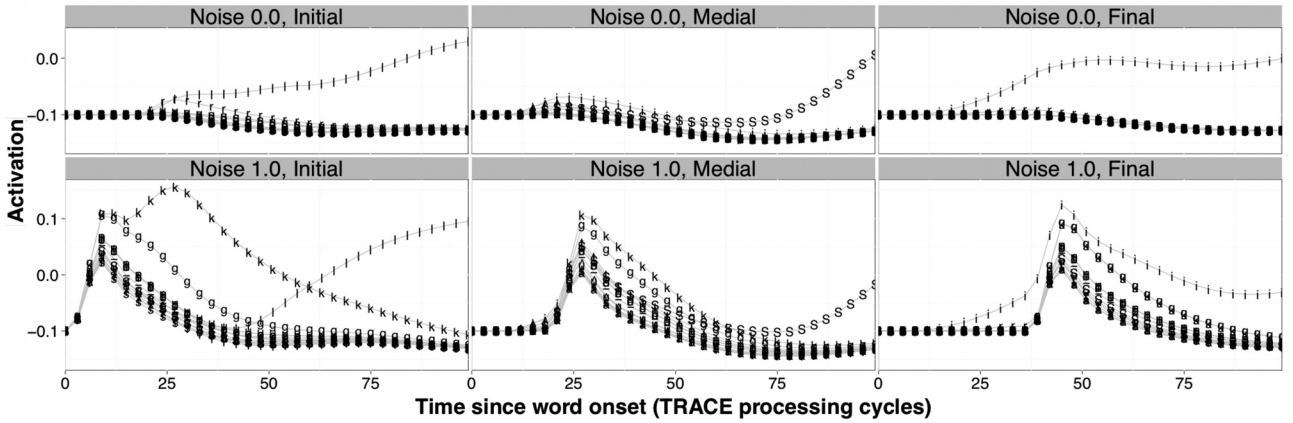

FIG. 4. Phoneme activations given true silence (top row) or noise = 1.0 (bottom row) for initial, medial and final phoneme replacements for luxury with the full TRACE lexicon, original phoneme feature definitions, and standard parameters.

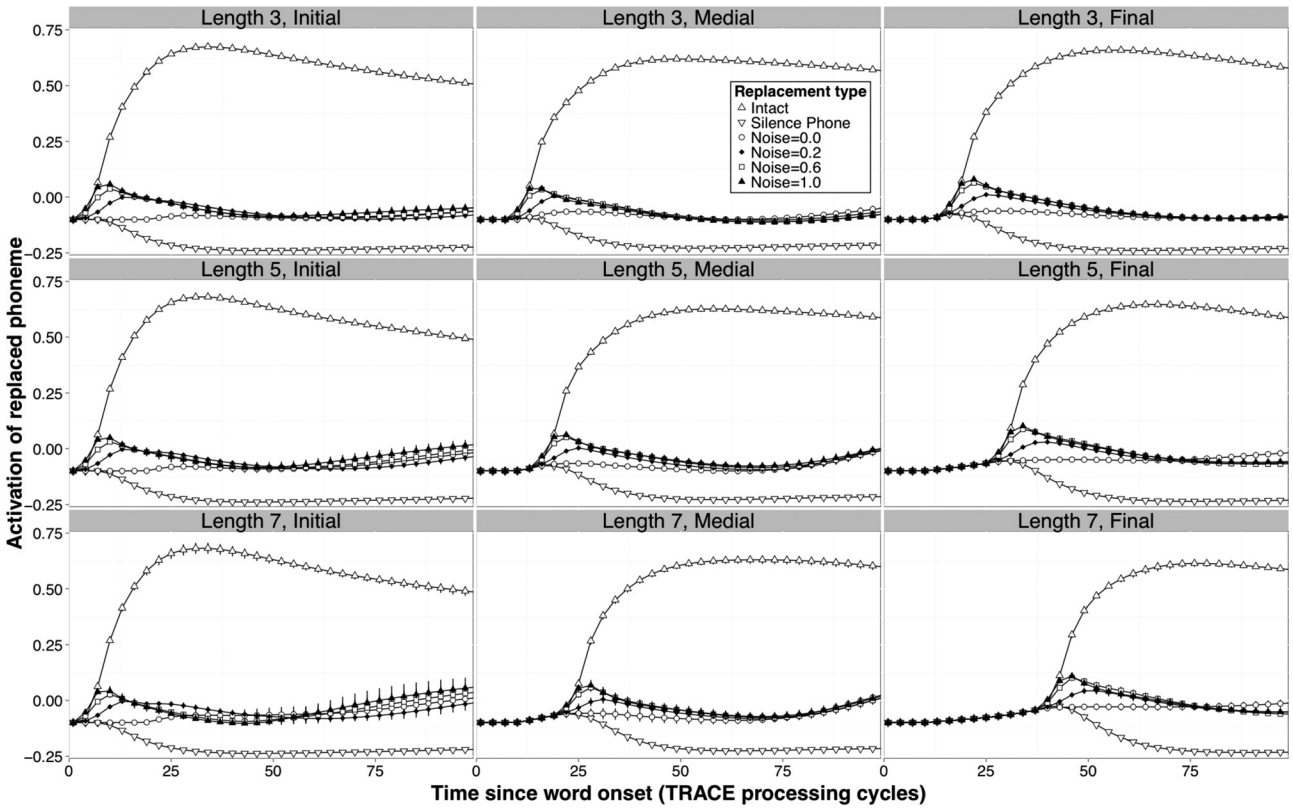

FIG. 5. Mean critical phoneme activations for all three-, five-, and seven-phoneme words (top, middle, and bottom, respectively), with critical phonemes in initial, medial, or final position (left, middle, and right columns, respectively), with original phoneme intact, replaced with the silence phoneme, replaced with true silence, or replaced with Gaussian noise with mean of 0.2, 0.6, or 1.0. Error bars indicate 95% confidence intervals; however, variability is so low most error bars are masked by plotting symbols, indicating high consistency between words.

and medial positions? Obviously, the lexical node for luxury is less activated at word onset or word medially than at word offset. But we can also see that at initial position, /l/ has to overcome greater activation of /r/, and this raises an important point: the replaced phoneme is not necessarily the most probable (e.g., given /#æt/, where # indicates noise, the obscured word could be bat, cat, fat, hat, lat, mat, gnat, pat, rat, etc.), nor is it necessarily likely to receive more top-down activation than other words. Indeed, given /#^/ at word onset, /r/ is more likely than /l/, as four words in the TRACE lexicon begin with /r^/ vs three that begin with /l^/. Feedback is also passed down by words that partially match the input after word onset (e.g., /pl^g/ or /kr^S/); in fact, the bigram /r^/ is more likely than /l^/ (occurring in 17 vs 10 words, respectively). A similar situation holds for medial position, where /S/ has to overcome activation of /i/, which is activated strongly by the word /l^ki/ (lucky); /i/ has an advantage both because /ki/ is more frequent in the TRACE lexicon than /kS/ (three vs one occurrences) and because of the short-word advantage that emerges naturally in TRACE [benefiting lucky vs the longer word luxury; TRACE exhibits an early short-word activation advantage (short words activate more quickly than long words), and a late long-word advantage (long words become more active than short words); both effects were documented in human perceivers by Pitt and Samuel, 2006; see Magnuson *et al.,* 2013, for analysis].

Figure 9 shows replaced phoneme activations by word length and noise level, while Fig. 10 shows mean activation of replaced phonemes across word lengths and replacement positions, as well as the next-most activated phoneme (with revised feature definitions). Figure 9 looks much like Fig. 5. Figure 10 shows that in initial position, the most activated phoneme aside from the replaced phoneme (labeled "next" in the figure) has greater activation in the early time window than the replaced phoneme. However, an analysis of phonotactic probabilities based on the TRACE lexicon revealed that the next-most activated phoneme in these cases was more (usually) or equally probable compared to the replaced phoneme based on the frequency of the first bigram in the word. Thus, the behavior of TRACE mirrors lexical probabilities, demonstrating robust top-down lexical restoration effects. The same is true for medial position, where the next-most activated phoneme tends to be consistent with bigrams that are slightly to substantially more likely than bigrams in the target word.

Note that this means that a pure test of lexical restoration in TRACE is virtually impossible in a lexicon with even 200 words, if we take a pure example of restoration to mean that the replaced phoneme is the most active as soon as top-down feedback is available; if we conduct tests where we replace the initial phoneme in all three-phoneme words, the majority of the time, there will be phonemes with greater phonotactic probability than the replaced phoneme. We can conduct a pure test by reducing the lexicon to a single word. The results for luxury when it is the sole lexical item are shown in Fig. 11, replicating greater activation for replaced phonemes given noise than given silence, and Fig. 12,

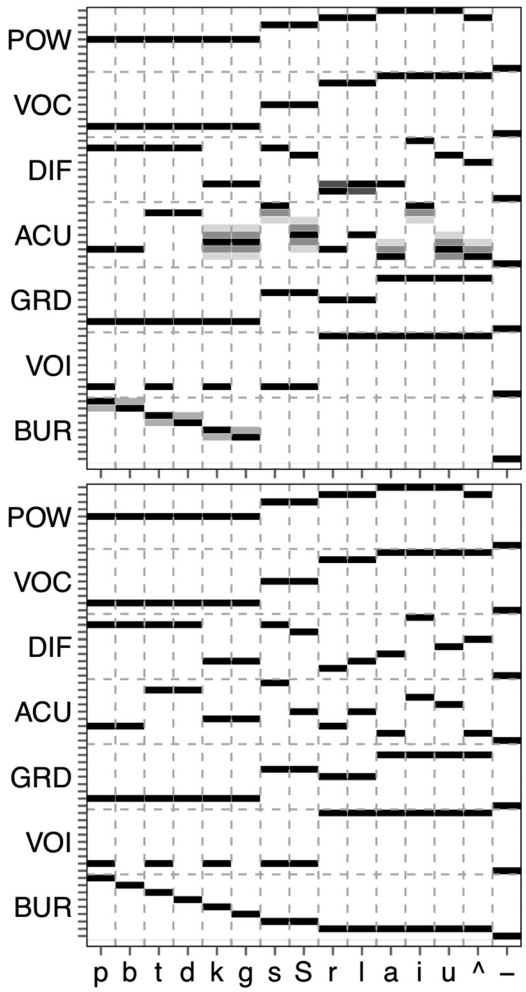

FIG. 6. Heatmap depiction of original TRACE phoneme definitions (top) and revised definitions (bottom). For original features, cell values range from 0.0 (white) to 1.0 (black), with a few gradient values (0.2, 0.3, 0.5) in the original definitions. For revised features, values are 0.0 or 1.0, and each phoneme has exactly seven features on. Each column corresponds to one phoneme center; features spread and ramp on and off to these peak values as seen in Figs. 1 and 2.

demonstrating that the replaced phonemes now receive selective benefits of top-down feedback (with clear pre-activation from top-down feedback for medial and final positions).[5]

Consider again Fig. 9, showing how the results generalize across words in the standard TRACE lexicon with revised phoneme feature definitions. Although there is little variation across length and position, Fig. 13 reveals an interesting difference (using the revised feature values shown in Fig. 6). Figure 13 plots the average early peak (prior to cycle 50) of replaced phonemes as a function of word length and replacement position. There are clear effects of word length, replicating larger effects with longer words with human subjects (Samuel, 1981a,b, 1996, although the trend is carried primarily by final position replacements. There are also clearly stronger effects for later replacement positions, although there is no trend for initial position and a weaker trend for medial position. While increased restoration with increased word length is attested in the psychological literature (Samuel, 1981a,b, 1996), effects of replacement position have been more mixed, sometimes with progressively stronger effects for later positions (Samuel, 1981b) and sometimes progressively *weaker* effects for later positions (Samuel, 1996, experiment 1.1), but perhaps most often, with a pattern where restoration is roughly equally strong for initial and final position, and somewhat weaker for medial position (e.g., Samuel, 1981a; Samuel, 1996, experiment 1.2). These discrepant patterns could be the result of interactions with word length. However, the studies that have manipulated both length and position present summary data only at the level of main effects, precluding examination of interaction patterns. Furthermore, as Samuel (1981b, 1996) points out, participants may well employ a strategy where they are biased to respond "intact" (restored). Thus, the unbiased/non-strategic predictions from TRACE might require modulation by a decision process [with appropriate linking hypotheses (Tanenhaus *et al.*, 2000)]. It is difficult to intuit what cARTWORD would predict, but comparable simulations would provide a basis for comparing the models empirically.

To summarize, noise replacements lead to greater phoneme activation than (true) silence or silence phoneme replacements in TRACE (Figs. 3–5, 7–12). However, with the original TRACE feature definitions (Fig. 6, top), some phonemes are more similar to broadband noise than others, and the bottom-up activations of these phonemes mask top-down effects, which emerge later. Replacing the original feature definitions with new ones (Fig. 6, bottom) that preserve phonemic similarity structure but equalize similarity to noise does not change the behavior of the TRACE model in general, but makes clear that noise-replacements selectively

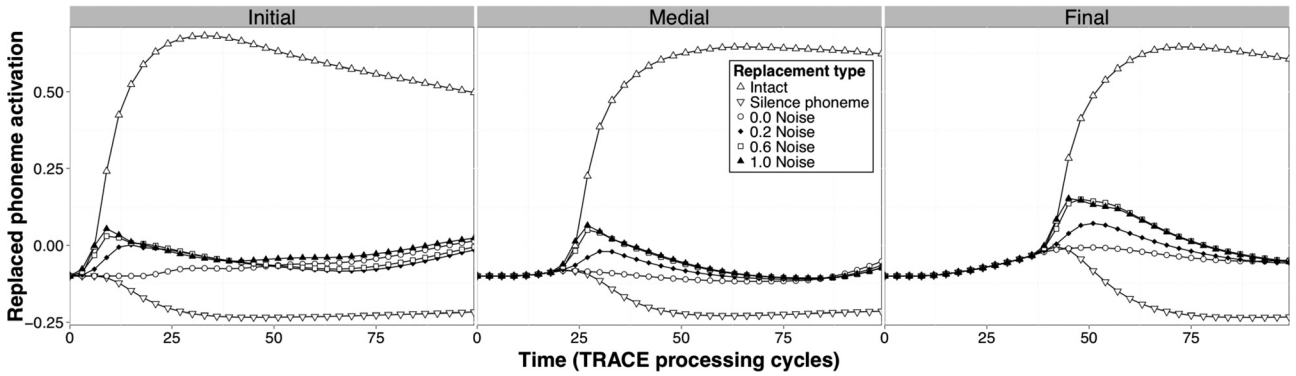

FIG. 7. Luxury with replacements at multiple positions, with revised feature values from Fig. 6 and original TRACE parameters.

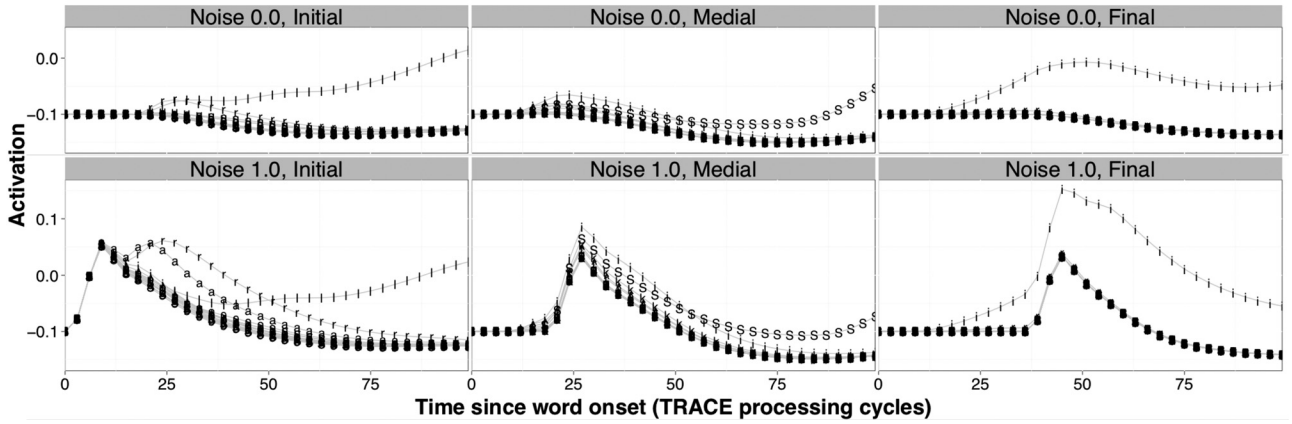

FIG. 8. Phoneme activations at initial, medial, and final replacements sites given the input luxury, for noise = 0.0 and noise = 1.0, with revised feature values and original TRACE parameters.

activate phonemes that are consistent with lexical feedback (though not necessarily just the intended word; lexical feedback operates quickly enough to boost phonemes consistent with high frequency biphones in the lexicon; Figs. 7 and 9). A purer test of lexical feedback can be obtained by constructing a lexicon with just one word; in this case, clear, early effects of top-down activation are observed (Fig. 11; see also footnote 5). Thus, TRACE does in fact provide a basis for predicting phoneme restoration given noise replacement vs (true) silence—or silence phoneme—replacement. Furthermore, its predictions are consistent with reports of greater restoration with increasing word length, and possibly could provide new insight into mixed results of replacement position in the perceptual literature. Thus, the empirical case Grossberg and Kazerounian (2011) make against TRACE evaporates with these results. With appropriately constructed input patterns, TRACE exhibits patterns found with human listeners.

V. COVERAGE VS IN-PRINCIPLE ARGUMENTS

TRACE stands out as the model of spoken word recognition with the broadest and deepest coverage, and it is crucial that any competing model be tested on a similar range of phenomena. Like cARTWORD, Shortlist/Merge (Norris *et al.*, 2000) relies on in principle arguments, rather than comprehensive coverage demonstrated by simulation. After presenting simulations of a small subset of phenomena the

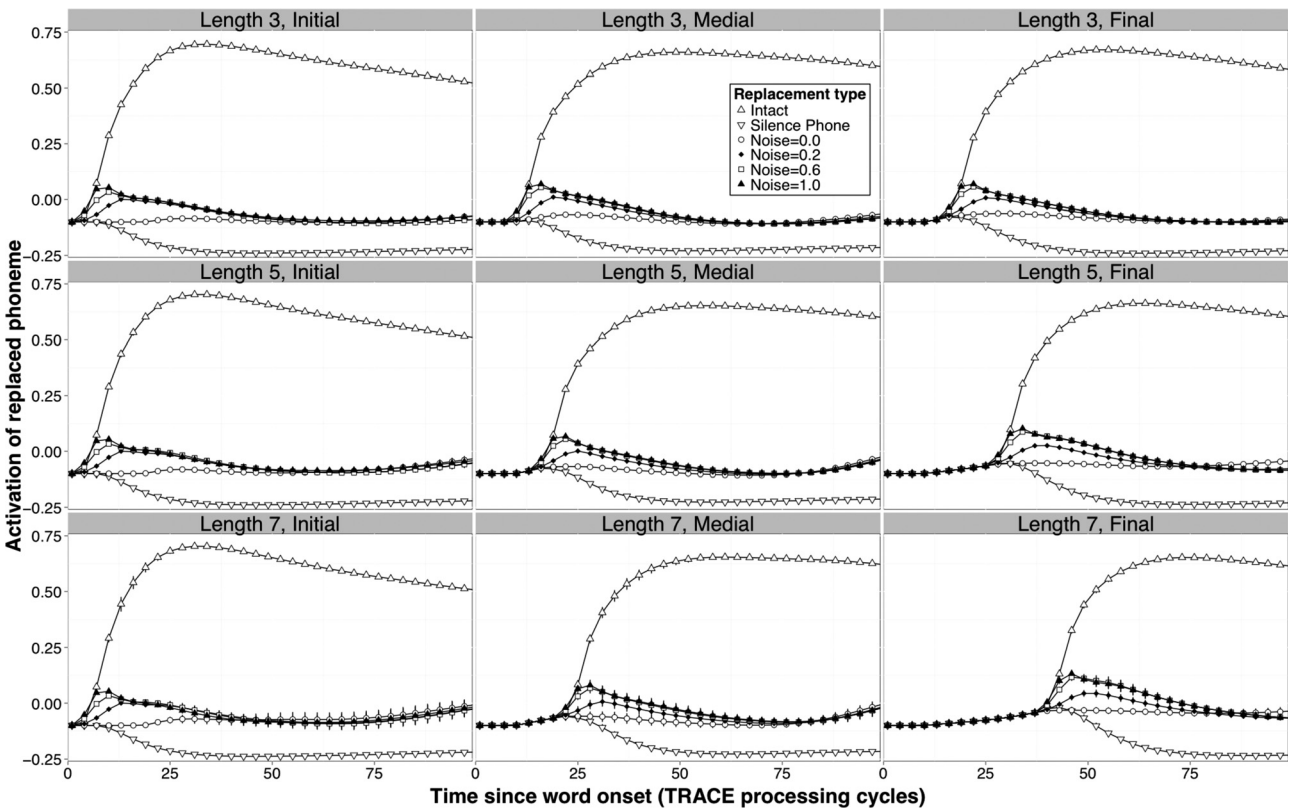

FIG. 9. Replaced phoneme activations for all three-, five-, and seven-phoneme words, with revised feature values. Error bars represent 95% confidence intervals.

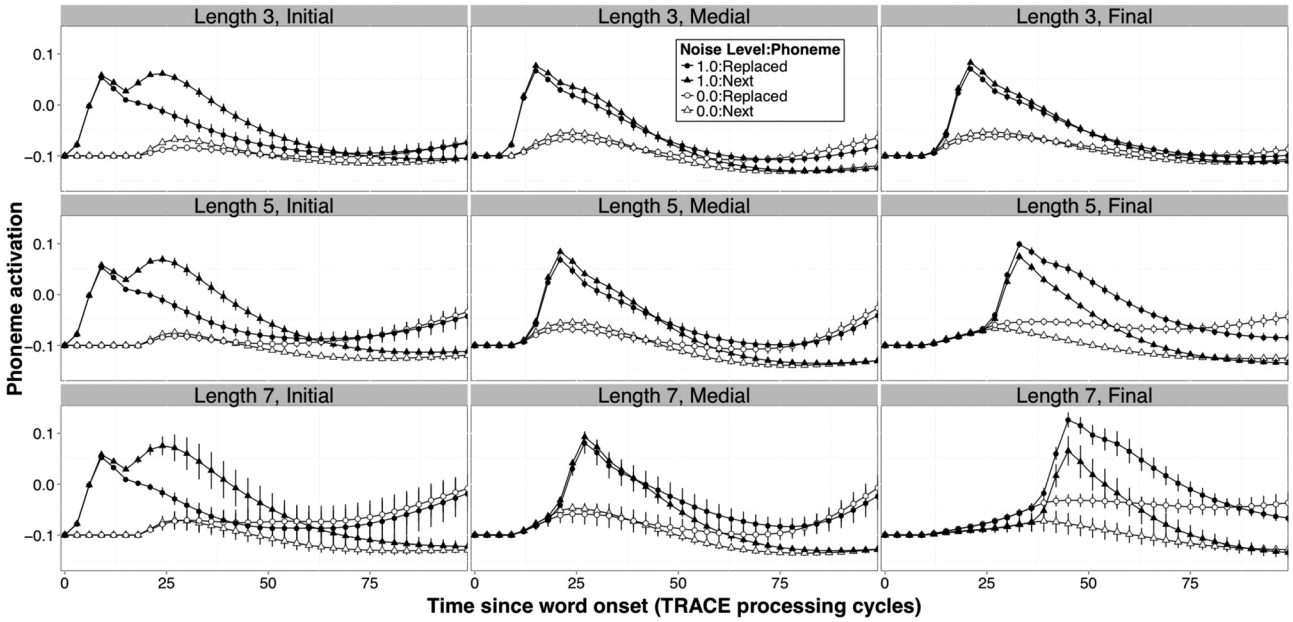

FIG. 10. Replaced phoneme and "next" (next-most activated) phonemes for all three-, five-, and seven-phoneme words, with revised feature values. Error bars represent 95% confidence intervals.

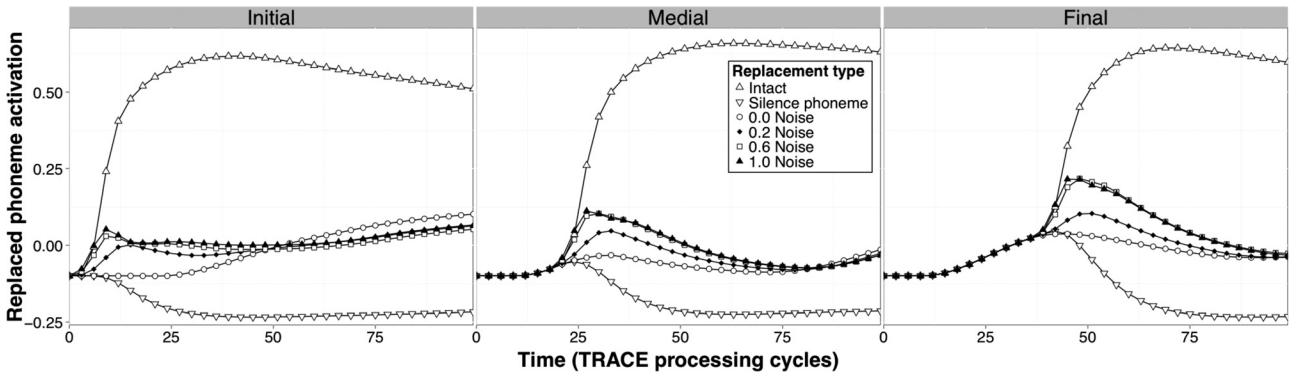

FIG. 11. Luxury with replacements at multiple positions, with revised feature values, but no lexical competitors (luxury is the only word in the lexicon).

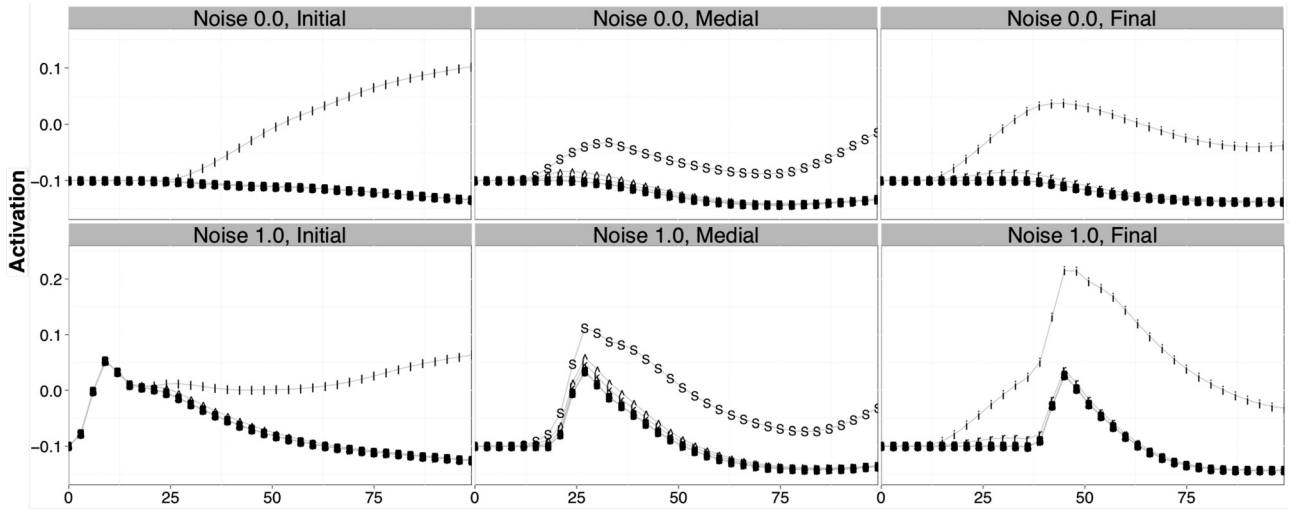

FIG. 12. Position-specific phoneme activations for luxury, for noise = 0.0 and noise = 1.0, with revised feature values, and no lexical competitors (luxury is the only word in the lexicon). The "pure" lexical restoration effect is now evident at all three replacement positions.

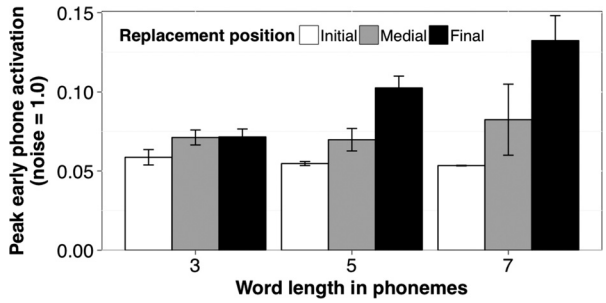

FIG. 13. Early peak activation by word length and replacement position for the full TRACE lexicon with revised feature definitions. Error bars represent 95% confidence intervals.

TRACE model handles, Norris *et al*. asserted, "It should be clear without report of further simulations, however, that Merge can also explain the other basic lexical effects observed in the literature" (p. 318). The set of unsimulated effects included phonemic restoration. But the fact that a framework logically *could* account for phenomena A, B, and C separately does not guarantee that parameter combinations exist that would allow the model to account for all three simultaneously. TRACE stands alone in its demonstrated ability to account for phoneme restoration across a range of word lengths and replacement positions, using the *same parameters* that have already been used to account for many phenomena in speech perception and word recognition [although demonstrating clear phoneme restoration effects in TRACE required minor revisions to phoneme definitions (Fig. 6), which do not alter TRACE's ability to account for the phenomena simulated in the original TRACE paper (McClelland and Elman, 1986)]. Challenging TRACE requires competing models to provide similar coverage without resorting to unique parameters for different phenomena.

VI. CONCLUSIONS

The plausibility and generality of cARTWORD is limited by its inability to represent sequences with repeated elements, while the fact that it has been implemented with only five abstract inputs and two words and has not been tested on a similar range of phenomena as models such as TRACE precludes comprehensive comparison. Simulations with TRACE following the methods of Grossberg and Kazerounian (2011)—but with better analogs to noise-replacement stimuli used in experiments with human subjects—demonstrated greater activation of noise-replaced phonemes than (true) silence—or silence phoneme—replaced phonemes. These results were replicated with the original TRACE phoneme definitions as well as slight modifications that made all TRACE phonemes equally similar to broadband noise. These successful TRACE simulations of phoneme restoration demonstrate that the failures Grossberg and Kazerounian (2011) to simulate those phenomena with TRACE were the result of a flawed approach to simulating noise replacement, not a failure of the model. As I discussed above, their critiques of reduplicated units and non-modulatory feedback in TRACE are not compelling: reduplicated units provide a plausible means for constructing an echoic memory [possibly quite similar to what IOR memory (Silver *et al.*, 2012) would be like were it scaled up to handle many phonemes and words] or could be eliminated with a programmable blackboard approach (McClelland, 1986), and non-modulatory feedback in TRACE was not the basis for the apparent failure of TRACE to correctly simulate phoneme restoration; rather, that failure was due to poor analogs of noise- and silence-replaced stimuli. I hope that these arguments and simulations will lead to the further elaboration of the cARTWORD model needed before valid comparisons to other models can be made.

ACKNOWLEDGMENTS

I thank Ted Strauss for suggestions regarding jTRACE simulations, and Dan Mirman, Paul Allopenna, and Jay McClelland for helpful discussions. Preparation of this manuscript was supported by NSF CAREER Grant No. 0748684, NIH P01-HD00199 to Haskins Laboratories (J. Rueckl, PI), and NIDCD Grant No. R15DC011875-01 to SUNY New Paltz (N. Viswanathan, PI).

[1]In the absence of a solution to the lack-of-invariance problem in speech perception, models and theories of spoken word recognition commonly simplify the input to unrealistically simple analogs of acoustic-phonetic patterns (as in TRACE) or even abstract and discrete analogs of phonemes (Grossberg and Kazerounian, 2011; Norris *et al.*, 2000).

[2]See Magnuson *et al.* (2012) for discussion; also see Hannagan *et al.* (2013) for an estimate of the number of nodes and connections needed to scale TRACE to a realistic English lexicon, as well as a novel string kernel approach to sequence encoding in the interactive activation framework that replaces TRACE's reduplication strategy and allows massive reductions in numbers of nodes and connections.

[3]I thank Jay McClelland for bringing to my attention this quote from the initial report of phoneme restoration by Warren (1970): "Silent intervals have functions akin to phonemes, requiring their accurate identification and localization for speech comprehension."

[4]I also conducted simulations with a large range of other noise values not displayed in Fig. 3 (0.1, 0.3, 0.4, 0.5, 0.7, 0.8, 0.9, 1.1, 1.2,…). The trends for early activation given noise seen in Fig. 3 were observed for all these values (albeit weakly for 0.1). Increasing noise beyond 1.0 has no further impact, as TRACE treats values greater than 1.0 as 1.0.

[5]To assess whether the single-word lexicon results would generalize to other items, I constructed four-word lexicons where all words were three-, five-, or seven-phonemes long, designed such that the initial, medial, and final phonemes of all words would be the most likely given the small lexicon. The results are similar to those for luxury when it is the one word in the lexicon; effects of top-down feedback are made very clear by immediate, selective activation of the replaced phoneme when noise is encountered. As in Fig. 9, some late advantages for 0.0 noise vs 1.0 noise are observed for final position replacements. I assume this is due to phoneme inhibition effects; noise activates phonemes other than the replaced phoneme, of course, and there is a period after the initial top-down boost where the replaced phoneme's activation lags due to lateral inhibition. To test this hypothesis, I reduced phoneme lateral inhibition from its default level of 0.4 to 0.3. This resulted in virtually identical early time window results, but wiped out the late, small advantage for 0.0 noise, both for the small lexicons and the full lexicon. I also confirmed that with phoneme inhibition reduced in this way, the core 12 simulation results from the original TRACE paper are replicated. Presenting these results would require several more figures, but there is little point; again, what is crucial is what happens *at the time of replacement,* not the late, post-word offset time course. However, I describe these results for readers concerned about the slight, late advantage for 0.0 noise for some conditions.

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (**1998**). "Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models," J. Mem. Lang. **38**, 419–439.

Carpenter, G. A., and Grossberg, S. (**1987**). "A massively parallel architecture for a self-organizing neural pattern recognition machine," Comput. Vision Graph. Image Process. **37**, 54–115.
Connine, C. M., Blasko, D. G., and Hall, M. (**1991**). "Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints," J. Memory Lang. **30**, 234–250.
Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (**2001a**). "Time course of frequency effects in spoken-word recognition: Evidence from eye movements," Cogn. Psychol. **42**, 317–367.
Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (**2001b**). "Tracking the time course of subcategorical mismatches: Evidence for lexical competition," Lang. Cogn. Process. **16**(5/6), 507–534.
Elman, J. L. (**1990**). "Finding structure in time," Cogn. Sci. **14**, 179–211.
Elman, J. L. (**1991**). "Distributed representations, simple recurrent networks, and grammatical structure," Mach. Learn. **7**, 195–224.
Elman, J. L., and McClelland, J. L. (**1986**). "Exploiting the lawful variability in the speech wave," in *Invariance and Variability of Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Lawrence Erlbaum, Hillsdale, NJ), pp. 360–380.
Gaskell, M. G., and Marslen-Wilson, W. D. (**1997**). "Integrating form and meaning: A distributed model of speech perception," Lang. Cogn. Process. **12**, 613–656.
Grossberg, S., and Kazerounian, S. (**2011**). "Laminar cortical dynamics of conscious speech perception: Neural model of phonemic restoration using subsequent context in noise," J. Acoust. Soc. Am. **130**, 440–460.
Hannagan, T., Magnuson, J. S., and Grainger, J. (**2013**). "Spoken word recognition without a TRACE," Front. Psychol. **4**, 563.
Kucera, H., and Francis, W. N. (**1967**). *Computational Analysis of Present-Day American English* (Brown University Press, Providence), pp. 1–424.
Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear Hear. **19**, 1–36.
Magnuson, J. S., Mirman, D., and Harris, H. D. (**2012**). "Computational models of spoken word recognition," in *The Cambridge Handbook of Psycholinguistics*, edited by M. Spivey, K. McRae, and M. Joanisse (Cambridge University Press, Cambridge, UK), pp. 76–103.
Magnuson, J. S., Mirman, D., and Myers, E. (**2013**). "Spoken word recognition," in *The Oxford Handbook of Cognitive Psychology*, edited by D. Reisberg (Oxford University Press, New York), pp. 412–441.
Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (**2003**). "The time course of spoken word recognition and learning: Studies with artificial lexicons," J. Exp. Psychol. Gen. **132**(2), 202–227.
McClelland, J. L. (**1986**). "The programmable blackboard model of reading," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by J. L. McClelland, D. E. Rumelhart, and the PDP research group (MIT Press, Cambridge, MA), Vol. II, pp. 122–169.
McClelland, J. L. (**1991**). "Stochastic interactive processes and the effect of context on perception," Cogn. Psychol. **23**, 1–44.
McClelland, J. L. (**2013**). "Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review," Front. Psychol. **4**, 503.
McClelland, J. L., and Elman, J. L. (**1986**). "The TRACE model of speech perception," Cogn. Psychol. **18**, 1–86.
McClelland, J. L., Mirman, D., Bolger, D. J., and Khaitan, P. (**2014**). "Interactive activation and mutual constraint satisfaction in perception and cognition," Cogn. Sci. **38**(6), 1139–1189.
McClelland, J. L., and Rumelhart, D. E. (**1981**). "An interactive activation model of context effects in letter perception: Part 1. An account of basic findings," Psychol. Rev. **88**, 375–407.
Norris, D., McQueen, J. M., and Cutler, A. (**2000**). "Merging information in speech recognition: Feedback is never necessary," Behav. Brain Sci. **23**, 299–325.
Pitt, M. A., and Samuel, A. G. (**2006**). "Word length and lexical activation: Longer is better," J. Exp. Psychol. Hum. Percept. Perform. **32**, 1120–1135.
Samuel, A. (**1981a**). "The role of bottom-up confirmation in the phonemic restoration illusion," J. Exp. Psychol. Hum. Percept. Perform. **7**, 1124–1131.
Samuel, A. (**1981b**). "Phonemic restoration: Insights from a new methodology," J. Exp. Psychol. Hum. Percept. Perform. **4**, 474–494.
Samuel, A. G. (**1996**). "Does lexical information influence the perceptual restoration of phonemes?," J. Exp. Psychol. Gen. **125**, 28–51.
Samuel, A. G. (**1997**). "Lexical activation produces potent phonemic percepts," Cogn. Psychol. **32**, 97–127.
Silver, M. R., Grossberg, S., Bullock, D., Histed, M. H., and Miller, E. K. (**2012**). "A neural model of sequential movement planning and control of eye movements: Item-Order-Rank working memory and saccade selection by the supplementary eye fields," Neural Netw. **26**, 29–58.
Strauss, T. J., Harris, H. D., and Magnuson, J. S. (**2007**). "jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition," Behav. Res. Methods **39**, 19–30.
Tanenhaus, M. K., Magnuson, J. S., Dahan, D., and Chambers, C. (**2000**). "Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing," J. Psycholing. Res. **29**, 557–580.
Warren, R. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.
Watkins, O. C., and Watkins, M. J. (**1980**). "The modality effect and echoic persistence," J. Exp. Psychol. Gen. **109**(3), 251–278.