Magnuson, J. & You, H. (2018). Feedback in the Time-Invariant String Kernel model of spoken word recognition. In Chuck Kalish Martina Rau, Jerry Zhu & T. Rogers (Eds.), CogSci 2018 (pp. 732–737).

Feedback in the Time-Invariant String Kernel model of spoken word recognition

James S. Magnuson
james.magnuson@uconn.edu
Department of Psychological Sciences and
CT Institute for the Brain and Cognitive Sciences
University of Connecticut, Storrs, CT 06269-1020, USA

Heejo You
heejo.you@uconn.edu
Department of Psychological Sciences and
CT Institute for the Brain and Cognitive Sciences
University of Connecticut, Storrs, CT 06269-1020, USA

Abstract

The Time-Invariant String Kernel (TISK) model of spoken word recognition (Hanngan et al., 2013) is an interactive activation model like TRACE (McClelland & Elman, 1986). However, it uses orders of magnitude fewer nodes and connections because it replaces TRACE's time-specific duplicates of phoneme and word nodes with time-invariant nodes based on a string kernel representation (essentially a phoneme-by-phoneme matrix, where a word is encoded as by all ordered open diphones it contains; e.g., *cat* has /kæ/, /æt/, and /kt/). Hannagan et al. (2013) showed that TISK behaves similarly to TRACE in the time course of phonological competition and even word-specific recognition times. However, the original implementation did not include feedback from words to diphone nodes, precluding simulation of top-down effects. Here, we demonstrate that TISK can be easily adapted to lexical feedback, affording simulation of top-down effects as well as allowing the model to demonstrate *graceful degradation* given noisy inputs.

Keywords: Computational models; neural networks; spoken word recognition; interaction; feedback

To feedback or not to feedback

Theories of spoken word recognition agree on 3 principles: (1) incrementally (as a word is heard), (2) words in memory are activated as a function of similarity to the input and prior probability (e.g., word frequency), and (3) activated words compete for recognition. Theories differ in how they map phonetic inputs to lexical items and mechanisms that they propose to account for the dynamics of lexical competition (Magnuson, Mirman & Harris, 2012). Notable differences include proposals for or against lexical inhibition or top-down (lexical-to-phoneme) feedback (McClelland & Elman, 1986 vs., respectively, Marslen-Wilson & Warren, 1994 or Norris, Cutler & McQueen, 2000, 2016). The best-known model of spoken word recognition (SWR) is the interactive-activation model, TRACE (McClelland & Elman, 1986), which uses explicit lexical-phonemic feedback to account for top-down effects in SWR (several are described below). In contrast, Norris et al. (2000; see also 2016) have argued that anything a feedback system can do can be done in a system without feedback

Top-down effects in SWR include the *Ganong effect* (Ganong, 1980) effect, where phoneme identification is biased according to lexical knowledge. For example, compared to a nonword continuum between *iss* and *ish*, where participants are asked to identify the final consonant, identification shifts towards /s/ if the continuum is instead between a word and nonword pair like *kiss-kish* or towards /ʃ/ if the continuum is instead between a nonword-word pair like *fiss-fish*. Another important top-down effect is *phoneme restoration* (Samuel, 1981a,b, 1996, 1997), where a phoneme replaced by noise is perceived (or at least identified) consistently with lexical context (e.g., the same noise, #, is heard as /t/ in /æf#^r/ but as /f/ in /æ#t^r/). Participants typically report hearing all phonemes in the obscured word, and have difficulty identifying the phonemic position of the noise. Crucially, if a phoneme is replaced with silence, restoration does not occur and participants easily identify which phoneme is missing.

Norris et al. (2000, 2016) have argued that direct feedback from words to phonemes (what they call *activation feedback*) cannot benefit speech processing. They claim that any system employing activation feedback can be matched by a purely feedforward system wherein top-down effects emerge from post-lexical integration of lexical and phonemic representations (rather than online modulation of phoneme representations by lexical feedback). Norris et al. (2000) demonstrated that an autonomous (feedforward) network with post-lexical integration could simulate top-down effects like those described above. They further argued that a system tuned to optimally identify each phoneme could not be improved by top-down feedback.

However, this ignores an important motivation for feedback in parallel-distributed processing (PDP) models: *graceful degradation* (for example, given noise). Magnuson, Mirman, Luthra, Strauss and Harris (2018; see also Magnuson, Strauss & Harris, 2005) have demonstrated beneficial effects of feedback in TRACE. Magnuson et al. compared accuracy and recognition time for every word in the original 211-word TRACE lexicon as well as a larger, 907-word lexicon with and without feedback. As noise was added, feedback preserved accuracy and recognition times were faster with feedback than without.

Feedback and TISK

Hannagan, Magnuson and Grainger (2013) introduced the *Time-Invariant String Kernel* (TISK) model of spoken word recognition. We will describe TISK in more detail in the next section. For now, we note that Hannagan et al. did not include lexical-to-N-phone feedback in the original TISK implementation, for purposes of simplicity. Our goal in this paper is to examine whether it is possible to implement feedback in TISK without impeding its ability to simulate the phenomena covered by Hannagan et al. (2013) while endowing it with the ability to simulate familiar top-down effects and with the robustness in noise (graceful

degradation) demonstrated for TRACE by Magnuson et al. (2018). We also search for parameters that would allow TISK to exhibit graceful degradation without feedback.

Representing sequences for word recognition

Two fundamental challenges for models of spoken word recognition are representing temporal order and representing repeated elements. To illustrate these challenges, consider the simple network diagrammed in Figure 1. In this network, phoneme nodes feedforward to word nodes. Each word has incoming connections from each of its constituent phonemes. However, this network cannot encode temporal order. The phoneme sequences corresponding to ACT, CAT and TACK (as well as nonwords such as /tkæ/, /ktæ/, or /ætk/) would generate the same amount of activation for the three corresponding word nodes. The network is also unable to encode distinct codes for words with repeated phonemes. The input /dæd/ would equally activate nodes for DAD and ADD. The second /d/ in /dæd/ would simply be more evidence that /d/ had occurred; the network cannot represent two instances of /d/ in specific temporal positions.

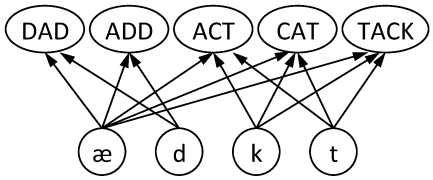

Figure 1: A simple word recognition network where phonemes feed to words, but neither order nor repeated elements can be represented. Reproduced with permission: https://doi.org/10.6084/m9.figshare.5852532.v1.

This model is not a caricature; a model like this can be used productively to explore the dynamics of competition where order does not matter (e.g., cases where amount of overlap rather than temporal distribution of overlap matters). Indeed, the Merge model (Norris et al., 2000) has exactly this structure. But of course, ultimately, models of spoken word recognition (SWR) must go beyond this simplifying assumption and grapple with the representation of order and repeated elements.

The TRACE model (McClelland & Elman, 1986) takes an infamously brute-force approach to the problem. TRACE essentially translate time to space, by creating time-specific duplicates of feature, phoneme, and word nodes. A template for CAT is maximally activated by strongly activated /k/, /A/, and /t/ phonemes aligned with the word node, which must be activated by appropriately aligned pseudo-spectral inputs on the feature level (see Figure 2). As "time" progresses in a TRACE simulation, inputs aligned with specific time points activate aligned features, phonemes, and words. This time-specific "reduplication" strategy -- aligning copies of each feature, phoneme, and word with specific time points -- allows TRACE to represent temporal order, and repeated elements. The first /d/ of DAD and the second will activate independent /d/ nodes.

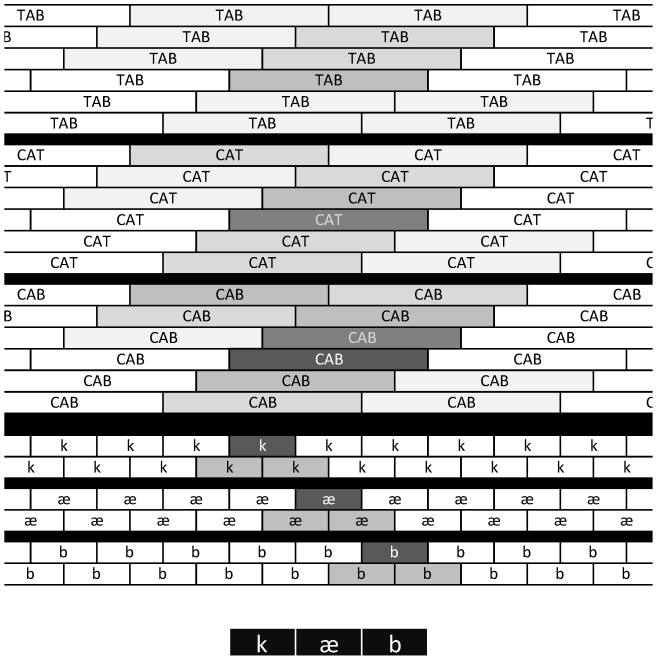

Figure 2: Schematic of TRACE's time-as-space encoding. At the bottom of the figure, inputs (/k/, /æ/, /b/,) have specific alignments (in TRACE, these would be distributed representations of over-time pseudo-spectral features). Inputs activate phoneme nodes aligned with them, which in activate aligned word nodes. Darkness of shading indicates approximate degree of activation. Reproduced with permission: https://doi.org/10.6084/m9.figshare.5852556.v1.

This reduplication strategy has been criticized many times, beginning with the original TRACE paper (McClelland & Elman, 1986, p. 77). We do not agree with claims that this architecture is completely implausible (see responses and counterarguments in Hannagan et al. [2013] and Magnuson [2015]). However, estimates of how large TRACE would have to be to accommodate realistic phonological and lexical inventories raise the question of whether more efficient solutions might be possible. Hannagan et al. estimated that extending TRACE to accommodate 40 phonemes and 20,000 words would require ~1.3 million nodes and more than 40 billion connections. This is because of the large number of *time-specific* nodes TRACE requires (copies of each phoneme and word node aligned at many time slices in TRACE's memory). Hannagan et al. developed a solution that replaces almost all time-specific nodes with *time-invariant* nodes – e.g., just one instance of each word node.

The way TISK does this is by using a variant of *open diphone coding*. Open diphones are phoneme pairs (which can be ordered or unordered; we use ordered pairs) that occur in a string whether they are adjacent or not. For example, the phonemes of *act* are /ækt/. Its ordered open diphones are /æk/, /kt/, and /æt/. We list several examples in Table 1 that should give an intuitive sense that enumerating open diphones could provide distinctive codes for similar words. It might also seem problematic that the number of diphones will grow with word length; how do we compare a word with one open diphone (2 phonemes long) to one with

6 (4 phonemes long) or 10 (5 phonemes long)? This is where kernel operations come in. We can represent each word as a phoneme x phoneme matrix, where each cell represents an ordered phoneme, and its value is (for example) the count of the appropriate diphone. (If we include a "blank" for the second position, we can also encode each single phoneme in a word, crucially providing a means for including words consisting of a single phoneme.) Then kernel operations – e.g., matrix similarity – can be applied independently of word length, as the matrix provides a length-independent representation format.

Table 1: Examples of ordered open diphones.

Word	Ordered open diphones
CAT	kæ, kt, æt
TACK	tæ, tk, æk
ACT	æk, æt, kt
DAD	dæ, dd, æd
ADD	æd
SOUL	so, sl, ol
SOLO	so x 2, sl, ol, oo

TISK does not use simple open-diphones, however. It uses a *symmetry network* that weights diphone activation by the distance between the two phones (such that /st/ would be less activated by SPOT than STOP). Hannagan and Grainger (2012) discuss biological plausibility of such coding, and behavioral and brain imaging results consistent with open bigram coding for visual word recognition. Work by Hannagan et al. (2011) suggests that similar coding may emerge in trained connectionist models. See Hannagan et al. (2013) for finer details of TISK. To use TISK, see You and Magnuson (2018), and the TISK Python repository at https://github.com/maglab-uconn/TISK1.0.

TISK 1.1: Adding lexical feedback

As we discussed above, there are several motivations for adding feedback to TISK. First, without feedback, an interactive activation model cannot simulate well-replicated findings of top-down lexical effects on sublexical processing (and to be clear, while feedback in an interactive model achieves those lexical effects through direct lexical influence, this remains controversial; see Norris et al., 2000 and 2016 for arguments that top-down influences can apply post-perceptually without feedback). The second reason appears to be less familiar to most cognitive scientists, even though it is a primary motivation for feedback: feedback allows graceful degradation (for example, when noise is added to speech). This gives us a very clear 5-point agenda in adding feedback to TISK, formulated as **5 questions**:

1. TISK without feedback had similar time course and item-specific RTs as TRACE; does TISK with feedback?
2. Can we find a parameter set (that includes top-down lexical-to-N-phone feedback) that allows TISK to simulate top-down effects while preserving its ability to simulate phenomena it has already been tested on (Hannagan et al., 2013)?
3. Are its top-down effects plausible (comparable to human performance?
4. Does feedback allow TISK to exhibit graceful degradation in the face of noise?
5. Can we find parameters that afford graceful degradation in the face of noise *without* feedback?

Table 2: TISK and TISKfb (with feedback) parameters. N-phone includes both single phones and diphones. There is positive feedback to words' constituents and inhibitory feedback from words to non-constituents (units not contained in the word).

Class	Parameter	TISK	TISKfb
Decay	Input phoneme decay	0.010	**0.001**
	N-phone decay	0.010	**0.100**
	Word decay	0.050	0.050
Gain	Input to N-Phone	0.100	0.100
	Diphone to word	0.050	0.050
	Single phone to word	0.010	0.010
	Word to word	-0.005	**-0.010**
Feedback	+Word to N-Phone	0.000	**0.100**
	-Word to N-phone	0.000	**-0.050**

Table 2 lists parameters for the original TISK model and for **TISKfb** (*with feedback*). The original parameters were determined via trial and error by Hannagan et al. (2013), and are stable to modification (a fairly wide range of values can be used for each parameter). We found that stable performance with feedback requires both positive feedback from words to constituents (component diphone and single phone units in the N-phone layer) and weaker negative feedback to inhibit non-constituents, as well as stronger decay for N-phone units. In order to isolate effects of feedback, we compare TISKfb to TISK with all feedback parameters set to zero, but with the same changes in decay and inhibition shown in Table 2. There is not space in this paper to report full details of our explorations of the parameter space, but we did find that these parameter changes actually make TISK (with or without feedback) more robust. We now turn to the 5 questions.

Figure 3 addresses part of **question 1** (*are item-specific RTs similar in TISK with feedback as in TRACE and TISK?*) by plotting item-specific RTs for TISK without feedback, TISKfb (TISK with lexical feedback), and TRACE for the original 211-word TRACE lexicon. Clearly, item-specific RTs are similar.

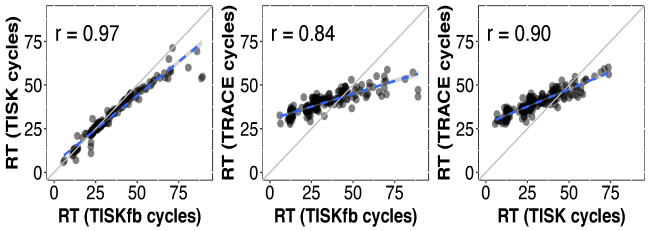

Figure 3: RT correlations for TISK, TISKfb (TISK with feedback), and TRACE. Solid line is the identity line; dashed line is linear best fit; 'shadow' (so narrow it is difficult to see) indicates 95% CI.

Figure 4 addresses the other part of **question 1 (***is the time course of phonological competition similar in TISKfb as in TISK and TRACE*?) as well as **question 2** (*does TISKfb account for everything reported in Hannagan et al., 2013?*). The rank ordering of competitor types remains the

same, although each is damped somewhat. (Note that TISK and TRACE differ in that 0.0 is the lowest possible activation in TISK; hence, rank order is the crucial concern.)

Figure 5 further addresses **question 2** (*does TISKfb account for everything reported in Hannagan et al., 2013?*) and plots RT in the three models as a function of "lexical dimensions:" length, different types of competitors, and one "external" count – the number of other words a target word embeds into (see caption). We observe a remarkable degree of similarity among the models in the strength and direction of each predictor's relationship to item-specific RT.

Figures 6 and 7 address **question 3** (*are TISKfb's top-down effects plausible?*). Figure 6 explores the *Ganong effect* (Ganong, 1980). We begin with a continuum from one phoneme to another (e.g., changing in steps from /s/ to /ʃ/, i.e., *ess* to *esh*) and establish a baseline identification /s/-rate at each step of the continuum. If we alter the continuum such that one endpoint corresponds to a word, while the other corresponds to a nonword (e.g., from *bus* /b^s/ to **buhsh* /b^ʃ/, or from **russ* /r^s/ to *rush* /r^ʃ/), and measure identification again, we will find that the /s/-/ʃ/ decision boundary shifts towards the lexical endpoint. To test TISK and TISKfb on this, we created a continuum from /s/ to /ʃ/ and tested it without lexical context (top row of Figure 6) with and without feedback (left and right panels) to establish a baseline. Then we placed this continuum in the word-nonword contexts *bus-*buhsh* (middle row) and nonword-word contexts **russ-rush* (bottom row). The contexts have no effect in the original TISK (left panels), but shift "identification" in the same direction it would be shifted for human subjects with feedback (right panels).

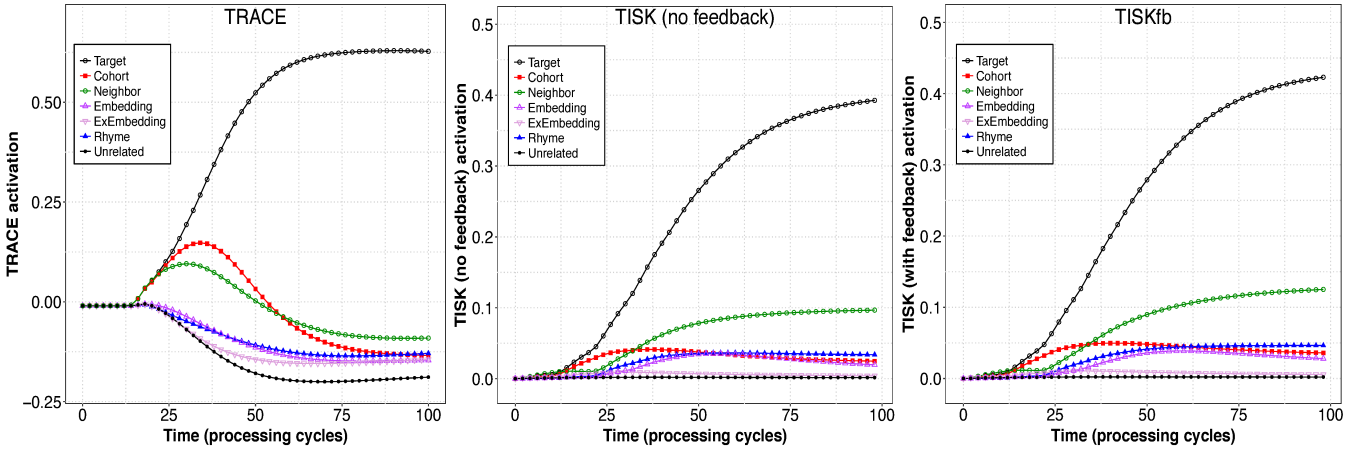

Figure 4: Mean time course for targets and different classes of competitors in TRACE, TISK, and TISKfb. Each line represents the mean for a class of items over all 211 words in the original TRACE lexicon. Cohorts overlap in the first two phonemes. *Rhyme* items overlap in all but the first phoneme. *Neighbors* differ by a single phonemic deletion, addition, or substitution. *Embeddings* are words embedded within a target, while *exEmbeddings* are words a target is embedded within (e.g., *cat* has *at* embedded within it, *at* "ex-embeds" in *cat*).

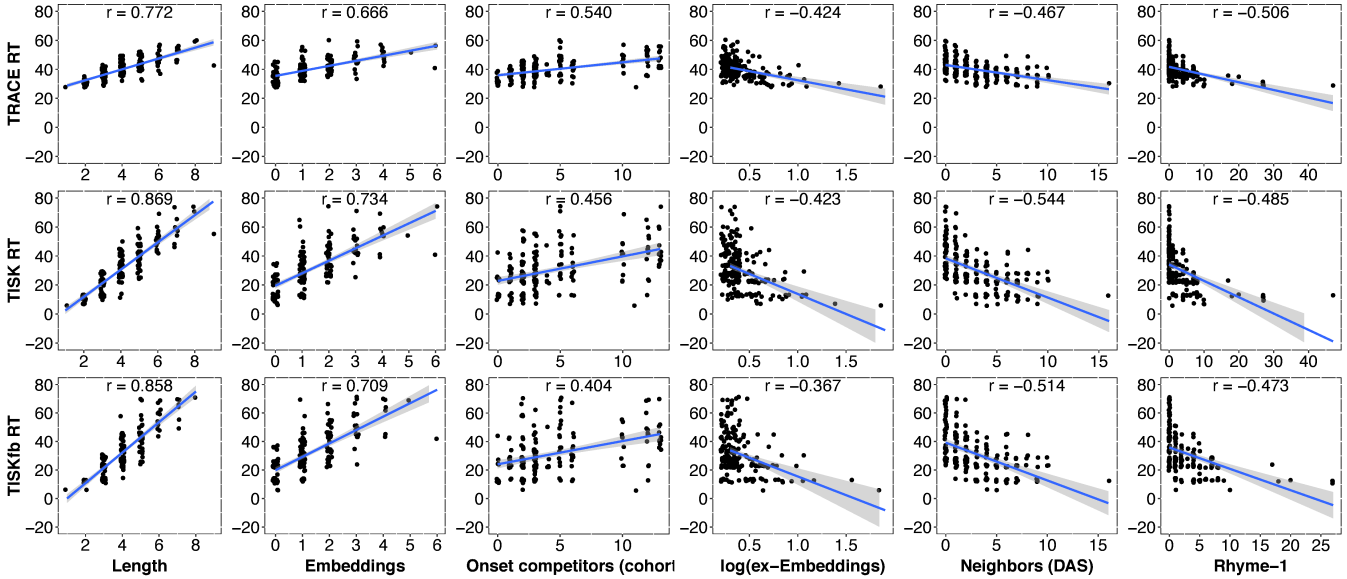

Figure 5: item-specific RTs in TRACE, TISK without feedback, and TISKfb (with feedback) as a function of lexical dimensions for the 211-word TRACE lexicon. *Length* is length in phonemes. Other dimensions are described in the caption for Figure 4. Solid lines in each plot indicate linear best fit, and shadows on that line indicate 95% confidence intervals.

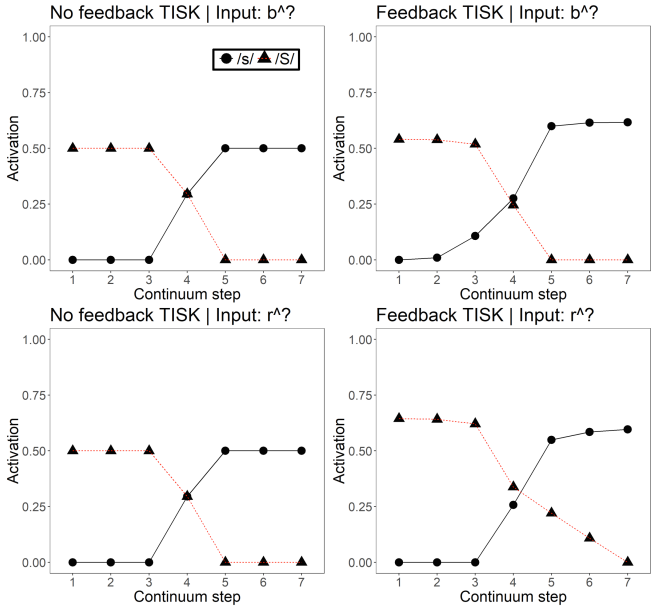

Figure 6: Lexical effects on phoneme activations. Top: input is /b^?/, where /?/ is a continuum between /s/ and /ʃ/. On the left is the result with TISK without feedback, with activations plotted for /s/ and /ʃ/; activations change approximately linearly across the continuum. On the right, results are plotted with TISKfb; crucially, activations of /s/ increase, as they are consistent with the lexical item *bus*, while no changes is seen for /ʃ/, which corresponds to a nonword ending with *sh* but the same onset as *bus*. On the bottom row, the opposite pattern is observed, as /ʃ/ is consistent with the word *rush*, while /s/ would make the nonword **russ*.

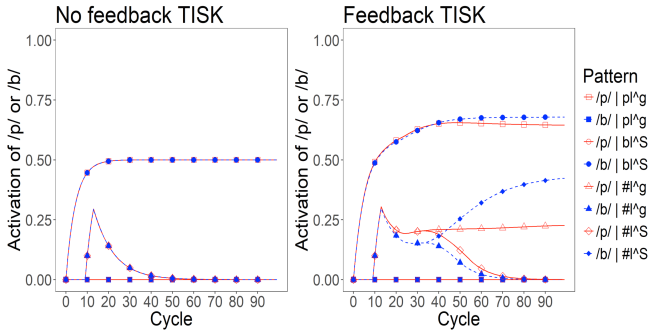

Figure 7: Retroactive phoneme restoration by following context. The input items are /pl^g/, /bl^ʃ/ (complete words *plug* and *blush*) and /#l^g/ and /#l^ʃ/ (*plug* and *blush* with the first phoneme replaced with noise). On the left, without feedback, there is robust activation of /p/ and /b/ given clear words, and transient, equal activation given noise replacement. On the right, we see that feedback enhances the activation of intact phonemes, and sustains context-appropriate phonemes after the transient response to noise. The difference in the activation of restored phonemes (/b/ given /#l^ʃ/ is more activated than /p/ given /#l^g/) is due to differences in neighborhood (*blush* has fewer competitors than *plug*).

In Figure 6, we see proactive effects of feedback; preceding context modulates later phoneme activations. In Figure 7, we examine potential *retroactive* influences in a Ganong experiment. Here, the input is either the intact word *plug*, the intact word *blush*, or a phoneme stimulus that is perfectly ambiguous between /p/ and /b /followed by *–lug* or *–lush* (denoted as /#l^g/ vs. /#l^S/ in Figure 7). Thus, only the final phoneme disambiguates. Without feedback (left panel), there is no context effect. In TISKfb (right panel), we see two effects of lexical context. First, there is differential activation of /b/ and /p/ after the second phoneme occurs (step 20), because there are more /p/-onset words than /b/-onset words in the TRACE lexicon. Second, we see the effect we predicted: the final phoneme drives lexical disambiguation effects on the first phoneme's activation. The smaller impact for /p/ follows from its denser competition neighborhood.

Questions 4 and 5 are whether we will observe benefits of feedback (graceful degradation) like those observed with TRACE (Magnuson et al., 2018) with TISKfb, and whether there are parameter configurations can allow a model without feedback to exhibit graceful degradation. In Figure 8 (following page), we present mean accuracy and response time (for correctly recognized items) for TISK with the original Hannagan et al. (2013) parameters, TISKfb, and TISK with all the same parameter changes as TISKfb as the amount of noise we add (to every phoneme in a word) increases. As Magnuson et al. observed with TRACE, feedback promoted graceful degradation of accuracy as noise increased, though there was little variation in RT for correctly recognized items. The new parameters in Table 2 provided substantial robustness against noise with or without feedback compared to the original parameters. Crucially, though, feedback provides a substantial benefit beyond those conferred by changes in decay and inhibition (as can be seen in the right panel of Figure 8).

A crucial question is whether the results without feedback could be improved with different parameters. While we have not yet searched the parameter space exhaustively, we have heuristically searched a fairly broad range of what appear to be the critical dimensions. The results in Figure 8 represent approximately the best results we have been able to obtain with and without feedback.

Conclusions

The answers to our first four questions are clearly "yes". ; TISKfb parameters can be selected that promote stable, TRACE-like performance while providing a basis for modeling top-down lexical effects. The answer to the fifth question is a qualified "yes": changes in decay and inhibition parameters provide substantial improvement in graceful degradation, but not to a degree that matches TISK with feedback. Thus, our results converge with those of Magnuson et al. (2018) in demonstrating the beneficial role of feedback in promoting graceful degradation, contra claims by Norris et al. (2000, 2016) that feedback in interactive activation models can provide no benefit.

Acknowledgments

We thank Thomas Hannagan for helpful advice and comments. This work was supported in part by NSF IGERT grant DGE-1144399 (J. Magnuson, PI) and NSF grant 1754284 (J. Magnuson, PI).

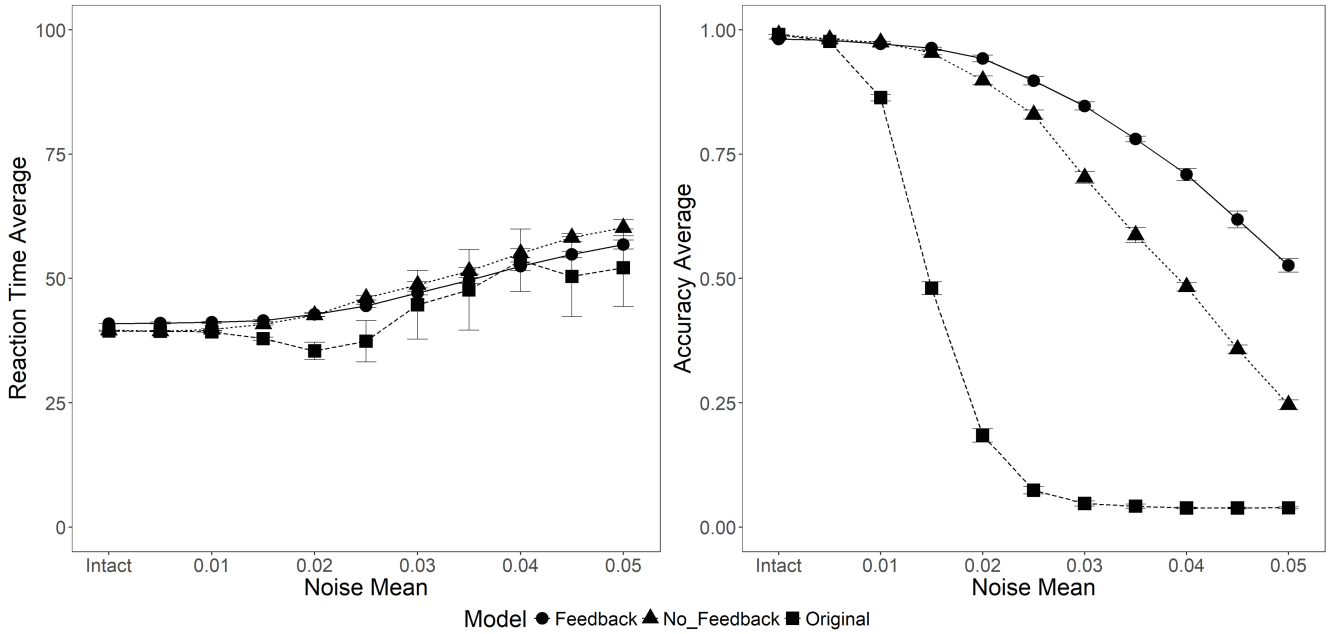

Figure 8: Graceful degradation under noise. Progressively larger amounts of Gaussian noise were added to input units. At each level of noise, the mean of the distribution is indicated; the standard deviation was one half of the mean; after addition of noise to inputs, values less than 0 or greater than 1 were replaced with 0 and 1, respectively. Left panel: RT for correctly recognized items under increasing levels of noise added to input nodes for TISK with original (Hannagan et al., 2013) parameters, with TISKfb parameters from Table 2, and for TISK "no feedback" with the new parameters from Table 2 except with feedback parameters set to zero. Effects on RT were minimal. Right: two interesting effects were observed in accuracy. The new TISK parameters provided substantially greater graceful degradation with and without feedback compared to the original parameters. However, feedback provided a large additional benefit.

References

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110-125.

Hannagan, T., Dandurand, F., & Grainger, J. (2011). Broken symmetries in a location invariant word recognition network. *Neural Computation, 23*, 251–283.

Hannagan, T., & Grainger, J. (2012). Protein analysis meets visual word recognition: a case for String kernels in the brain. *Cognitive Science, 36*, 575–606. doi:10.1111/j.1551-6709.2012.01236.x

Hannagan, T., Magnuson, J. S. & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, 4:563. doi:10.3389/fpsyg.2013.00563.

Magnuson, J. S. (2015). Phoneme restoration and empirical coverage of interactive activation and adaptive resonance models of human speech processing. *J. Acoust. Soc. America*, 137, 1481-1492. doi:10.1121/1.4904543.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics*, pp. 76-103. Cambridge University Press.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, 9:369. doi:10.3389/fpsyg.2018.00369

Magnuson, J. S., Strauss, T. J., & Harris, H. D. (2005). Interaction in spoken word recognition models: Feedback helps. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1379-1384.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology,* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

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

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4-18.

Samuel, A. (1981a). The role of bottom-up confirmation in the phonemic restoration illusion. *J. Exp. Psych.: Human Perception and Performance, 7*, 1124–1131.

Samuel, A. (1981b). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 474–494.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*, 28-51.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32,* 97-127.

You, H. & Magnuson, J.S. (2018). TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*. doi:10.3758/s13428-017-1012-5