Interaction in spoken word recognition models: Feedback helps

James S. Magnuson* **(james.magnuson@uconn.edu)**
Ted Strauss (ted.strauss@uconn.edu)
Harlan D. Harris (harlan.harris@uconn.edu)
Department of Psychology, University of Connecticut
Storrs, CT 06269-1020 USA

Abstract

A long-standing debate between autonomous and interactive models of spoken word recognition was given new life with the claim that not only are autonomous models more parsimonious, but feedback cannot aid recognition (Norris, McQueen, & Cutler, 2000). The claim was bolstered by simulations with the preeminent interactive activation model, TRACE (McClelland & Elman, 1986). When lexical feedback is turned off, as many words are recognized *more* quickly as are recognized *less* quickly (Frauenfelder & Peters, 1998), suggesting that feedback does not help recognition, even in the prototypical interactive model. However, these simulations used only a small subset of the lexicon, and did not address a primary motivation for interaction: to make the model robust in noise. We compared recognition of every item in a large TRACE lexicon with and without feedback under multiple levels of noise. With or without noise, most words were recognized more quickly with feedback than without. Feedback also makes the model resistant to noise: recognition is more accurate and faster with feedback than without when noise is added to the input. In short, feedback helps.

Word recognition: Interactive or autonomous?

The nature of information flow during perception and beyond is a central question in cognitive science. When should sensory data be integrated with prior knowledge? Does veridical perception require that sense data be protected from top-down knowledge, or can direct top-down interaction make bottom-up processing more efficient and robust in noise? Compelling arguments have been made for both positions. On the one hand, the case for early encapsulation follows from concerns that it would be difficult if not impossible to balance top-down and bottom-up information sources so as to allow veridical perception (Fodor, 1983); an organism that hallucinates tigers whenever it sees a flash of orange will not be able to act efficiently in its environment. Instead, veridical perception requires an encapsulated first-pass analysis of the bottom-up input that feeds its results forward to higher levels of representation where top-down knowledge can be integrated safely.

On the other hand, proponents of interaction argue that a system can learn to balance top-down and bottom-up interaction such that veridical perception is not sacrificed but top-down information is available to guide perception continuously from its earliest moments. On this view, early and continuous access to prior knowledge will make processing more efficient, essentially by tuning perceptual systems to prior probabilities based on experience (Knill & Richards, 1996).

While Fodor's (1983) arguments for modularity were meant to apply *between* rather than *within* input systems, similar arguments have been made for information flow within modalities. A notable example is the domain of speech perception and spoken word recognition, where this debate has recently taken center stage. The core phenomena at issue are lexical effects on the perception of sublexical units, e.g., phonemes (consonants and vowels), such as the word superiority effect (phonemes are detected more quickly in words than nonwords; Rubin, Turvey, & Van Gelder, 1976), and phoneme restoration (context dependent restoration of a phoneme replaced with noise or an ambiguous sound as a function of lexical or sentential context, e.g., Samuel, 1981).

Proponents of interactive models (e.g., McClelland & Elman, 1986) hold that early and continuous interaction not only accounts for effects of top-down knowledge, but does so in a way that is efficient and leads to robust performance given noise (McClelland & Rumelhart, 1981). Indeed, despite well-known deficiencies (McClelland & Elman, 1986; Norris, 1994), the TRACE interactive-activation model continues to hold its position as the model of speech perception and spoken word recognition that accounts for the broadest and deepest set of empirical phenomena.

Proponents of autonomous models (e.g., Norris, McQueen, & Cutler, 2000) argue that purely feedforward models can account for top-down effects via post-lexical interaction, precluding sublexical hallucination from feedback, and avoiding the arguably more complex machinery of feedback connections. Recently, the debate has been re-energized both empirically and theoretically. The empirical impetus came from evidence that apparently compelling evidence for interaction (that a lexically-restored phoneme could drive compensatory coarticulatory effects at the phoneme level, suggesting lexical feedback had truly modulated sublexical representations; Elman & McClelland, 1988) could arise from a potentially sublexical locus (diphone transitional probabilities) rather than from top-down lexical feedback (Pitt & McQueen, 1998). Later studies (e.g., Magnuson, McMurray, Tanenhaus, & Aslin,

* Also at Haskins Labs, New Haven, CT.

2003; Samuel & Pitt, 2003) replicated the original results using lexical contexts that could not be attributed to transitional probabilities, but not all parties are convinced (McQueen, 2003).

The theoretical impetus came from a paper by Norris et al. (2000), who argued that feedback is not necessary, and furthermore, it could never help word recognition – and might in fact hinder it. As Norris et al. put it:

> The best performance that could possibly be expected from a word recognition system is to reliably identify the word whose lexical representation best matches the input representation. This may sound trivially obvious, but it highlights the fact that a recognition system that simply matched the perceptual input against each lexical entry, and then selected the entry with the best fit, would provide an optimal means of performing isolated word recognition (independent of any higher-level contextual constraints), limited only by the accuracy of the representations. Adding activation feedback from lexical nodes to the input nodes (whether phonemic or featural) could not possibly improve recognition accuracy at the lexical level (p. 301).

Norris et al. bolstered this argument by pointing out that while feedback in TRACE does modify phonemic processing, Frauenfelder and Peeters (1998; FP98 from here on) showed that it does not have a "general beneficial effect on word recognition." FP98 compared the time it took a target word to reach a threshold with feedback set to the default level or turned off (set to 0.0; their Simulation 5). They compared recognition times for a set of 21 words that had been chosen for two other simulations. The expectation was that feedback would speed recognition, since the bottom-up input should be amplified by recurrence via top-down connections. However, FP98 found that about half their items were recognized more quickly without feedback than with feedback. The implication would seem to be that feedback in TRACE does not serve any useful purpose aside from accounting for top-down effects.

However, the FP98 simulation does not address a crucial motivation for feedback: it makes a model robust under degraded conditions, such as the presence of external or internal noise. This issue is addressed briefly in the original TRACE paper (McClelland & Elman, 1986), as it was addressed in depth in the first major interactive-activation paper (McClelland & Rumelhart, 1981). Noise is a crucial ecological consideration, given considerable internal noise in neural systems and the variable (and often literally noisy) conditions under which speech is experienced.

One way to construe feedback in interactive-activation models is that they make the models implicit Bayesians (cf. Movellan & McClelland, 2001). The simple addition of feedback gives the model early and continuous access to dynamic, context-sensitive prior probabilities at multiple windows of analysis without explicit representations of the probabilities. For example, simple diphone and longer n-phone transitional probabilities will emerge as a function of the structure of lexical neighborhoods: the more words there are with a particular sequence, the more feedback the component phonemes of that sequence receive. In the case of weak bottom up information (e.g., due to a low amplitude input signal or the presence of noise), feedback will help. Given roughly equivalent evidence for two sublexical alternatives, if one is contained in a word and the other is not, or one is contained in more words than the other, feedback will push the system towards that alternative. Given roughly equivalent bottom-up information for two lexical alternatives, if one has a higher prior probability (either in terms of lexical frequency [if it is implemented Dahan, Magnuson, & Tanenhaus, 2001] or sublexical frequencies implicit in the lexicon), this will be reflected in greater feedback and will push the system to favor the alternative more consistent with prior knowledge. Why, then, did FP98 fail to find a benefit of feedback? We argue that the apparent failure of feedback to help lexical recognition in TRACE stems from the failure to test the model in conditions where the bottom-up information does not perfectly identify a lexical alternative. We will also consider the possibility that the result may not generalize beyond the 21-word subset FP98 tested.

Simulations: Feedback and Noise

We reexamined the role of feedback in TRACE by comparing word recognition in TRACE with and without feedback, and under levels of increasing noise. This allows us to test the expectation that feedback in interactive-activation models should make them robust to noise. We also tested the generality of the FP98 failure to find a feedback advantage without noise by testing every word in a large, 901-word lexicon. FP98 only tested 21 words with homogenous characteristics (seven-phoneme words with uniqueness point at phoneme position four). These were chosen for specific reasons for their earlier simulations, but it is possible that they are not representative of the entire lexicon with respect to the effects of feedback.

Methods

Lexicon. We did not have access to FP98's "biglex" lexicon of 1024 words, so we generated our own ("biglex901") by following the procedures FP98 describe for compiling biglex: we scanned a large electronic dictionary (20,000 words) for all items that could be transcribed using only TRACE's 14 phonemes (/p/, /b/, /t/, /d/, /k/, /g/, /s/, /S/, /r/, /l/, /a/, /i/, /u/, /^/). This yielded a set of 462 words, so we substituted /^/ for schwa in the dictionary, which brought the total to 604. Collapsing across vocalic and consonantal liquids (substituting /l/ and /r/ for both) brought the total to 901.

TRACE parameters. We used the standard (McClelland & Elman, 1986) settings for all but two parameters. We assumed our lexicon would work best if we modified lexical inhibition and feedback as FP98 did to optimize performance with their biglex. We changed lexical inhibition from the standard 0.030 to 0.025, and of course, we manipulated feedback.

Feedback. We used three levels of feedback: none (0.00), the FP98 value for lexical feedback in large lexicons (0.015) and twice that (0.030, also the value used in the standard parameter set).

Noise. Gaussian noise was sampled from a normal distribution function and added to the input stimulus values (which range from 0.0 to 1.0). The mean of the distribution was kept constant at 0.0, while a 7-step continuum was created on the standard deviation of the noise in steps ranging from 0.0 to 1.5 in steps of 0.25.

Operationalizing recognition. TRACE solves the segmentation problem by reduplicating each phoneme and word unit. For example, there are copies of the template corresponding to the word "cat" aligned with the pseudo-spectral trace every 3 time cycles. In this way, a "cat" template will be closely aligned with any corresponding input over the entire input to the model. However, a modeler must decide how to interpret the bank of word units. FP98 based their interpretation on the method McClelland and Elman (1986) used for phoneme decisions: one simply chooses the unit known to be aligned with the input. FP98 point out that the unit immediately to the right of the perfectly aligned unit sometimes attains a higher activation, and therefore they summed the activation of the target unit perfectly aligned with the input and the unit immediately following it. One must also decide how to treat potential competitors. FP98 considered any unit with any overlap with the target as competitors. Response probabilities were then calculated at each TRACE processing step using the Luce (1959) choice rule:

$$R_i = \frac{e^{ka_i}}{\sum e^{ka_j}} \tag{1}$$

where R_i is the response probability for item i, a_i is that item's activation in TRACE, k is a constant (set to 20, as in FP98 simulations) that controls target-competitor separation, and the summed activations in the denominator include all target and competitor units. As in the FP98 simulations, an item was considered "recognized" when its response probability exceeded a threshold of 0.9.

While we have serious reservations about this decision rule (in particular, the selective nature of the target and competitor sets), we used it to maintain consistency with FP98. However, simulations with decisions rules that avoid these problems yield quite similar results (Magnuson, Strauss, & Harris, in preparation).

Simulation software. The simulations were conducted using jTRACE, a recent Java reimplementation of the original TRACE C code (Strauss, Magnuson & Harris, 2005, this volume). We have successfully replicated all attempted previous TRACE simulations with jTRACE, despite minor implementational differences (e.g., the original TRACE code depended heavily on pointer arithmetic, which is not available in Java). jTRACE is also augmented with an easy to use interface, and facilities for graphical representation, scripting and batch processing, and is available at http://maglab.psy.uconn.edu/jtrace.html. For our current purposes, jTRACE's ability to run large batches of simulations was crucial.

Procedure. Each word in the biglex901 lexicon was used as the target 21 times: 3 (levels of feedback) X 7 (levels of noise). We also repeated the simulations with smaller lexicons (including the original, 212-word TRACE lexicon known as *slex*), for reasons we discuss below. In each case, the decision rule described above was applied and recognition time and accuracy were recorded. Simulations were run for 100 cycles. Without noise, most words in the lexicon were recognized by cycle 95. Thus, 100 cycles provided adequate time for recognition; accuracy would not significantly increase at any feedback X noise level were we to run the simulations for more cycles. The items that were not correctly recognized typically fell short of the threshold, and could never reach it. This issue is addressed further below and by Magnuson et al. (in preparation).

Results

Feedback helps in the absence of noise. First, we examined whether we replicated the FP98 result that equal numbers of items are recognized more quickly with and without feedback. The left panel of Figure 1 is a scatter plot comparing reaction times with and without feedback. Items below the equality line were recognized more quickly with feedback than without. In fact, 73% of items in the lexicon were recognized more quickly with feedback.

How can we reconcile this with the FP98 report that feedback does not confer a general advantage? Recall that FP98 used a set of 21 words chosen to have particular characteristics important for earlier simulations. The 21 words they selected consisted of all the seven-phoneme words in their lexicon with uniqueness point at the fourth phoneme. We examined whether those items might have particular combinations of neighborhood, length, etc., that could lead to the FP98 result. In our 901-word lexicon, we only had 8 items that matched the FP98 characteristics. The right panel of Figure 1 shows the results for those 8 items. Just as FP98 reported, there is not a general feedback advantage for *items with these characteristics.*

So on the one hand, the items FP98 had chosen for other simulations happened to have characteristics that seem to counteract a feedback advantage. On the other, there remain a substantial proportion of items that are recognized more quickly without feedback. Our analyses so far show that items that are recognized more quickly *without feedback* fall into at least one but more often two or all three of the following sets: (a) long words – the longer a word is, the more items it overlaps with temporally, each of which can inhibit it; (b) items with multiple shorter words embedded within them; (c) items that share onsets with items that get early advantages from a "gang effect" – for example, "colleague" (/kalig/) is recognized more quickly without feedback, and its strongest competitor is "car" (/kar/), which receives feedback from several words beginning with /ka/

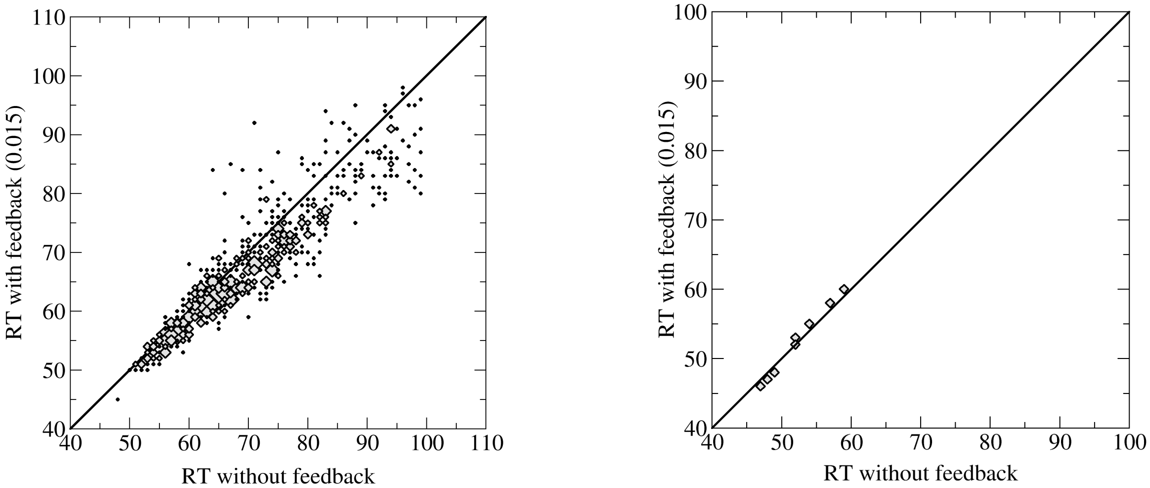

Figure 1: Response times in TRACE cycles with feedback (vertical axis) and without (horizontal). Items below the equality line are recognized more quickly with feedback. Left panel: all 901 items in the lexicon; size of a symbol indicates how many words lie at a point. Right panel: 8-item subset comparable to the FP98 set of 21 items.

and /kar/. In each of these cases, feedback boosts the activation of competitor items early on such that inhibition from those competitors slows the target's activation. We discuss these issues in more detail in Magnuson et al. (in preparation).

Feedback makes recognition resistant to noise. Again, the argument FP98 and Norris et al. (2000) make about the FP98 results – that there is no general benefit to feedback – neglects an important motivation for interactivity: robustness in noise (leaving aside for the moment the fact that we have now found a *general benefit of feedback*).

This motivated us to repeat the simulations with different levels of noise added to the input. As described above, Gaussian noise with a mean of 0.0 and standard deviation ranging from 0.0 (i.e., no noise) to 1.50 in steps of 0.25 was added to each input unit. All words in a lexicon were tested at each noise level. We also used three levels of feedback: none (0.0), the standard level for large lexicons (0.015) and the standard level for the original, 212-word lexicon (0.030). For each simulation, we measured accuracy and response time (time for the response probability of a unit to exceed the threshold).

Figure 2 shows accuracy and response time for each feedback condition at each level of noise. Regarding accuracy, note that there is a consistent advantage with feedback once the standard deviation of the noise reaches 0.5. The difference between the 0.015 and 0.030 levels of feedback demonstrates that too much feedback has a deleterious effect on accuracy, which suffers in the low-noise conditions.

Finally, we have the surprising result that accuracy without noise does not reach ceiling levels; it is a bit better than 85% with the lower level of feedback or without feedback. The items that are not recognized were always the most active item, but could not reach the FP98 threshold. These were typically short words or words in larger cohort groups.

To better understand why these words were problematic,

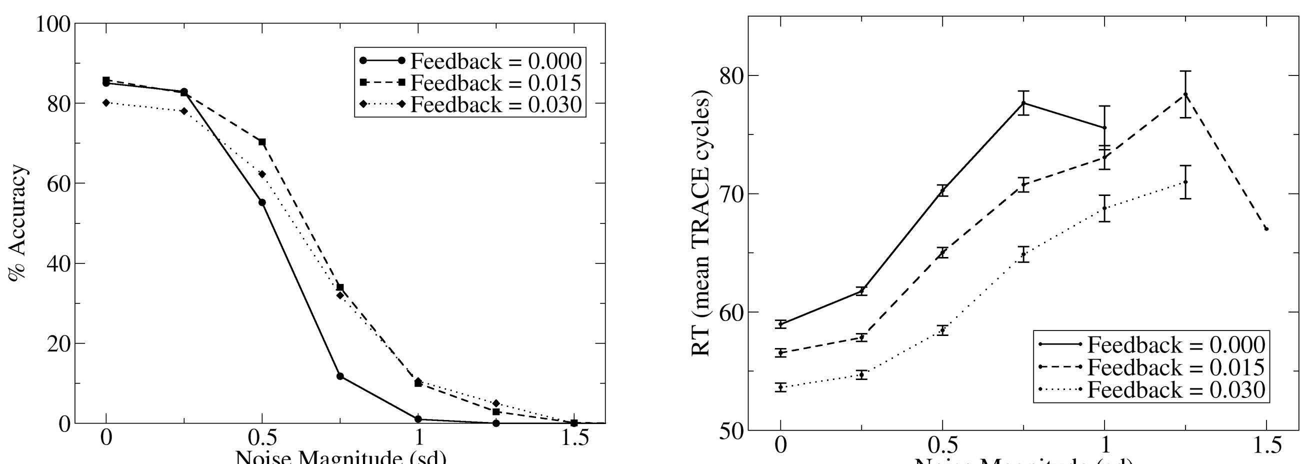

Figure 2: Mean accuracy (left) and response time for every item in the 901-word lexicon at each level of feedback and noise. Error bars in the right panel represent standard error (standard error is not meaningful for accuracy, as it is based on a binary measure – correct or not – for each item from a single simulation). The response time curves end at different noise levels for different levels of feedback because accuracy falls to 0 more quickly at different levels.

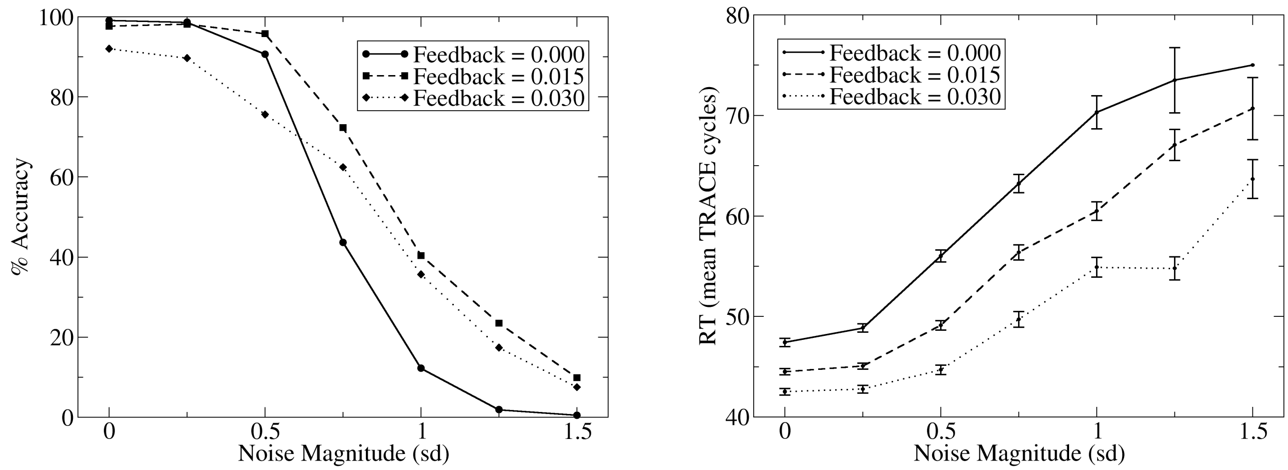

Figure 3: Mean accuracy (left panel) and response time for every item in the original 212-word TRACE lexicon (slex) at each level of feedback and noise. Error bars in the right panel represent standard error.

we explored several possible explanations. First, we tested whether the phoneme substitutions we used to get the lexicon up to 901 words might be responsible, as those substitutions could generate unnatural neighborhoods. We incrementally removed our substitutions, and repeated the full set of simulations with different lexicons. First, if we do not collapse across vocalic and consonantal liquids the lexicon is reduced to 604 words, but there is not a significant increase in accuracy. Second, if we do not substitute /^/ for schwa, the lexicon is further reduced, to 462 words, and there is a very modest improvement in accuracy. Reducing the lexicon back to the original 212 items and using the original parameter settings brings accuracy above 98% in the no- and low-feedback conditions. The results with this lexicon are shown in Figure 3.

Our current best explanation is that the accuracy decline with lexicon size reflects a scaling problem related to TRACE's small phoneme inventory. Neighborhoods become unrealistically dense as we increase the size of the lexicon, since we sample only from a small portion of the overall phonological space of the full English lexicon. Furthermore, a few of the vowel representations in TRACE seem to be insufficiently specific, as they occur in a larger than expected proportion of items that are not correctly recognized. For more detail, please see Magnuson et al. (in preparation).

The effects on response time are clear. Feedback speeds recognition (of those items recognized correctly), with an additional advantage to increasing feedback (with an accuracy trade-off, as increasing feedback also hurts accuracy at low levels of noise). A general benefit is observed, which increases with noise.

In short, *feedback helps*.

General Discussion

When the performance of TRACE is examined on a large number of items under clear and noisy conditions, the helpful role of feedback becomes apparent. Feedback is not a poorly-motivated mechanism that simply allows the model to account for top-down effects like word superiority. Instead, the underlying motivation for feedback in interactive activation models is to make performance robust under difficult conditions, such as speech in noisy or otherwise degraded or suboptimal conditions. Our simulations show feedback substantially boosts efficiency under clear and difficult conditions, both in accuracy and response time. This validates the original motivation for feedback, and demonstrates that the ability of a model like TRACE to account for top-down effects emerges from a mechanism with an important functional purpose.

Why does feedback help? As we discussed earlier, the top-down/bottom-up resonance that feedback generates makes an interactive activation model approximate a Bayesian analyzer; lexical representations encode sublexical patterns of varying grain sizes that guide the system as a whole towards the most likely cause of a particular input pattern. Given noisy input that is consistent with two sublexical patterns, one of which occurs in one or more lexical items but the other of which does not occur (e.g., a segment midway between /s/ and /S/ preceding /tr/), lexical feedback provides sublexical base-rate information, and guides the system to a rational response given the input.

Some (e.g., Norris et al., 2000) have claimed that feedback in a model like TRACE entails that the model will hallucinate clear inputs when it "hears" ambiguous ones. However, whether the model hallucinates or not is a function of the *amount* of feedback. Indeed, examples like Figures 13 and 30 in the original TRACE paper (McClelland & Elman, 1986) demonstrate how sensitive TRACE is to distorted input. In cases where ambiguous or incorrect phonemes are presented (e.g., /tluli/ instead of /truli/), TRACE recognizes a word, but one cannot claim it hallucinates the word or the restored phoneme. The word is recognized more slowly than a clear version, and the same holds at the phoneme level – e.g., in the /tluli/ example, /r/ is activated by feedback recurrence, but not as much as it would be given /truli/ as the input. Furthermore, the *trace* from which the model takes its name – the bank of memory units aligned with points in time – contains substantial information about the surface details of the input, and evidence that in fact /l/ rather than /r/ was "heard."

Some (e.g., Cairns et al., 1995; Norris et al., 2000) have argued that one could account for apparent top-down lexical effects sublexically. For example, one could build in sublexical base rate information by encoding diphone transitional probabilities at a phonemic level of representation. Proponents of this view often make the additional argument that this is how simple recurrent network models of spoken word recognition work – especially those without explicit lexical representations. However, Magnuson et al. (2003) report transitional probability analyses that demonstrate that for many lexical effects, the relevant base rate information is item-specific; that is, a transitional probability explanation for one item must appeal to diphones, while triphone or larger sequences are needed to account for other lexical effects. What is apparently needed is a dynamic, context-specific *n*-phone representation, where *n* equals uniqueness point or word length – which is exactly what lexical representations provide.

Magnuson et al. also argued that such representations are precisely what simple recurrent networks encode, even though the locus of the representation need not be an explicit lexical level; hidden units become sensitive to context-specific short- and long-range dependencies, providing a distributed lexical representation. Note that those representations depend on top-down feedback: the input to a simple recurrent network at a given time step includes bottom-up signal information as well as information about the states of the hidden units at previous time steps (via context units).

We have made a strong case that feedback *can* help recognition. But what of Norris et al.'s (2000) larger point – that feedback is never *necessary*? Norris et al. have abandoned the strong view contained in the extended quote we used earlier. They now acknowledge that there are times when prior knowledge – even lexical knowledge – has a prelexical influence (e.g., Norris, McQueen, & Cutler, 2003). We have argued against one solution – encoding diphone transitional probabilities at the phoneme level – as insufficient and certainly less efficient than lexical feedback for accounting for the dynamic, context-sensitive scope of relevant probabilities. Norris et al. have speculated that there is another solution that avoids online lexical feedback: *feedback for learning* (in analogy to backpropagation) could change feedforward phoneme-to-lexeme weights *after* processing. It remains to be seen whether such a mechanism can be implemented, but it keeps open the possibility that *online feedback* may not be necessary. However, given that online feedback *can* help – as we have demonstrated here – it has the potential to provide a more parsimonious account, if online feedback can also provide a basis for short- and long-term learning. Learning is beyond the current scope of TRACE, and we leave this question for future research.

Acknowledgments

Supported by National Institute on Deafness and Other Communication Disorders Grant DC-005765 to JSM.

References

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1995). Bottom-up connectionist modeling of speech. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds.), *Connectionist models of memory and language* (pp. 289–310). London: UCL Press.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001a). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42,* 317-367.

Frauenfelder, U. H., & Peeters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE.; In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101-146). Mahwah, NJ: Erlbaum.

Knill, D. C. & Richards, W. (1996). *Perception as Bayesian Inference*, Cambridge University Press, New York, NY.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science, 27*, 285-298.

Magnuson, J. S., Strauss, T. J., & Harris, H. D. (in preparation). Feedback in models of speech perception and spoken word recognition.

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

McClelland, James L.; Rumelhart, David E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*, 375-407.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review, 108*, 113-148.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, 87-104. Cambridge: MIT.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioural and Brain Sciences, 23*, 299-370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204-238.

Pitt, M. A. and McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon. *Journal of Memory and Language*, *39*, 347-370.

Rubin, P., Turvey, M. T., & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychphys., 19*, 384–398.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474-494.

Samuel, A. G. & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory & Language, 48*, 416-434.

Strauss, T. J., Magnuson, J. S., & Harris, H. D. (2005). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition.