Recognition of continuous speech requires top-down processing

Kenneth N. Stevens

Department of Electrical Engineering and Computer Science and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139. **stevens@speech.mit.edu**

Abstract: The proposition that feedback is never necessary in speech recognition is examined for utterances consisting of sequences of words. In running speech the features near word boundaries are often modified according to language-dependent rules. Application of these rules during word recognition requires top-down processing. Because isolated words are not usually modified by rules, their recognition could be achieved by bottom-up processing only.

In this commentary, I will address a question that is related to the problem under discussion here, but is somewhat more general: Does lexical access during running speech utilize top-down information from hypothesized lexical units to influence the processing of the speech signal at the sublexical level? The evidence in the target article of Norris et al. is based on psycholinguistic experiments with isolated words, and does not address the recognition of word sequences. The recognition of word sequences can present problems different from those for isolated words because when words are concatenated the segments can undergo modifications that are not evident in utterances of isolated words.

We begin by assuming that a listener has access to two kinds of language-specific knowledge. The language has a lexicon in which each item is represented in terms of a phoneme sequence, with each phoneme consisting of an array of distinctive features. The listener also has knowledge of a set of rules specifying certain optional modifications of the lexically-specified features that can occur in running speech. These modifications frequently occur at word boundaries, and are less evident in single-word utterances. (There are, of course, also obligatory morphophonemic rules.)

As acousticians with a linguistic orientation, we take the following view of the process of human speech recognition (Stevens 1995). There is an initial stage in which landmarks are located in the signal. These landmarks include acoustic prominences that identify the presence of syllabic nuclei, and acoustic discontinuities that mark consonantal closures and releases. The acoustic signal in the vicinity of these landmarks is processed by a set of modules, each of which identifies a phonetic feature that was implemented by the speaker. The input to a module is a set of acoustic parameters tailored specifically to the type of landmark and the feature to be identified. From these landmarks and features, and taking into account possible rule-generated feature modifications, the sequence of words generated by the speaker is determined. This process cannot, however, be carried out in a strictly bottom-up fashion, since application of the rules operates in a top-down manner. A typical rule specifies a lexical feature that potentially undergoes modification, it states the modified value of the feature, and it specifies the environment of features in which this modification can occur (cf Chomsky & Halle 1968). Thus it is necessary to make an initial hypothesis of a word sequence before rules can be applied. This initial hypothesis must be made based on a partial description of the pattern of features derived from the feature modules.

As an example, consider how the words can be extracted in the sentence "He won those shoes," as produced in a casual style. The /ð/ is probably produced as a nasal consonant, and the /z/ in "those" is usually produced as a palato-alveolar consonant, and may be devoiced. Acoustic processing in the vicinity of the consonantal landmarks for the word "those" will yield a pattern of features that does not match the lexically-specified features for this word. The feature pattern may, however, be sufficient to propose a cohort of word sequences, including the word "nose" as well as "those." Application of rules to the hypothesized sequence containing "those" will lead to a pattern of landmarks and features that matches the pattern derived from the acoustic signal. One such rule changes the nasal feature of the dental consonant from [−nasal] to [+nasal] when it is preceded by a [+nasal] consonant (Manuel 1995). (Close analysis will reject the word "nose," since the rule that creates a nasal consonant from /ð/ retains the dental place of articulation.) Another rule palatalizes the final /z/ when it precedes the palatoalveolar /š/ (Zue & Shattuck-Hufnagel 1979).

We conclude, then, that a model for word recognition in running speech must be interactive. That is, the process must require analysis by synthesis (Stevens & Halle 1967), in which a word sequence is hypothesized, a possible pattern of features from this sequence is internally synthesized, and this synthesized pattern is tested for a match against an acoustically derived pattern. When the utterance consists of isolated words, as in the experiments described in Norris et al.'s target article, there is minimal application of rules, and the acoustically based features match the lexically specified features. Consequently isolated word recognition can be largely based on bottom-up or autonomous analysis, as proposed by the authors.

No compelling evidence against feedback in spoken word recognition

Michael K. Tanenhaus, James S. Magnuson, Bob McMurray, and Richard N. Aslin

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627. **{mtan; magnuson; mcmurray}@bcs.rochester.edu aslin@cvs.rochester.edu www.bcs.rochester.edu**

Abstract: Norris et al.'s claim that feedback is unnecessary is compromised by (1) a questionable application of Occam's razor, given strong evidence for feedback in perception; (2) an idealization of the speech recognition problem that simplifies those aspects of the input that create conditions where feedback is useful; (3) Norris et al.'s use of decision nodes that incorporate feedback to model some important empirical results; and (4) problematic linking hypotheses between crucial simulations and behavioral data.

Norris et al. have provided a valuable service to the field by organizing and evaluating the literature concerning lexical influences on phonemic decisions in spoken word recognition. We believe their analysis will sharpen the discussion of issues in spoken word recognition and help shape the future research agenda in the field. Nonetheless, we find their case against feedback unconvincing for the following reasons.

1. Occam's razor has a double-edged blade. Norris et al. invoke Occam's razor to support their a priori claim that models without feedback should be preferred to models with feedback. Occam's razor, however, applies only when there is no empirical basis for preferring one model over another. In fact, there is considerable evidence for feedback connections in various cortical areas and for feedback in perceptual and cognitive processes. In visual perception, where the links between brain mechanisms and perception are best understood, there is evidence for feedback connections and processing interactions at both high and low levels (Churchland et al. 1994; Wandell 1995). There is also evidence for feedback at auditory levels presumably preceding phonemic processing (Yost & Nielsen 1977). Moreover, as Norris et al. acknowledge in section 3.5, feedback is likely at higher levels in language comprehension. Why, then, should sub-lexical processing be uniquely devoid of feedback? Given the ubiquitous nature of feedback in the brain, it is simpler to hypothesize feedback than to make sublexical processing a special case.

2. Feedback is surely helpful. Norris et al. argue that feedback cannot improve the efficiency of word recognition. This is only true given the sort of idealized input representation they use, consisting of noise-free discrete phonemes. Speech, however, is characterized by noise and variability (due to coarticulation, talker differences, etc.). Given a more realistic characterization of the input, feedback would be helpful, as it is in higher level language processing.

3. Feedback is required by the data and is incorporated into Merge. Norris et al. admit that lexical effects on phonemic decisions in non-words provide evidence against autonomous models of spoken word recognition. Merge allows for this feedback and differs from other autonomous models by adding phonemic decision nodes where phonemic and lexical information can be integrated. Although lexical information influences phonemic decisions in Merge, the autonomy of phonemic processing is preserved, because information at the lexical units is unaffected by the phonemic decision units. Parsimony issues aside, distinguishing interaction at the decision level from interaction at the "perceptual" level is at worst impossible and at best requires clearer linking assumptions between the model and the data.

4. Simulations against TRACE and in support of Merge are problematic. Although we are not trying to defend TRACE as a fully plausible model, it is important to note that the simulations challenging TRACE and supporting Merge depend upon particular parameter settings and questionable linking assumptions between the data and the models. Consider the subcategorical mismatch simulations that play a central role in Norris et al.'s arguments. The relevant Merge activations are shown in Figure 2 in section 5.2.1.

Compare the target activations for W1W1 from Figure 2A with the activations for N3W1 and W2W1 (the correct target is W1 for all three conditions). Clearly, the activations follow different timecourses. W1W1 precedes N3W1, which precedes W2W1. The puzzle, however, is that mean lexical decisions are fastest to W1W1 and slower (but equivalent) to N3W1 and W2W1. Marslen-Wilson and Warren (1994) reported that TRACE does not predict this pattern, but rather predicts the W1W1 < N3W1 < W2W1 ordering that is present in the activation functions. Merge is able to capture the empirical lexical decision pattern, despite showing similar activation patterns as TRACE, but only when a particular decision threshold (.20) is assumed. Activations for W1W1 cross this threshold at Cycle 8, and activations for N3W1 and W2W1 cross together at Cycle 10. With a slightly lower threshold, say .19, N3W1 would be faster than W2W1.

Norris et al. would like to conclude that this is compelling evidence for Merge and against TRACE. Their argument is that feedback in TRACE prevents the model from getting the activations just right; in their simulations with a mock-up of TRACE, they could not find a set of parameters that would yield a threshold where N3W1 and W2W1 will be treated the same without introducing other deviations from the actual lexical decision data. Their simulations of the sub-categorical mismatch findings might be a powerful argument against TRACE, if we had strong independent reasons to believe that (1) a particular all-or-none decision threshold of precisely .20 is correct, and (2) the feedback parameter in TRACE creates a fatal flaw which makes it impossible to find a threshold that would correctly simulate the lexical decision data. We find both of these assertions implausible.

More crucially, we should ask why lexical decisions are not mirroring the highly similar activation patterns predicted by both TRACE and Merge. Why do activations for W2W1, which lag behind activations for N3W1, have similar mean lexical decision times? The answer lies in the activation patterns and the linking hypotheses between the activations and lexical decision times. Early on in W2W1, W2 becomes quite active, following the same trajectory as W1W1 through Cycle 8. If one assumes that faster lexical decisions tend to be affected by earlier states of the system than slower lexical decisions or that the system is affected by noise, the distribution of lexical decisions in the W2W1 condition will contain a small proportion of fast "yes" times, based on activation of W2, as well as some slow "yes" responses based on the activation of W1. Whereas the means might be similar for the N3W1 and W2W1 conditions, the distributions are likely to differ in ways that are clearly testable but not revealed by mean lexical decision times alone.

More generally, we believe that arguments about model architecture on the basis of simulations of the type appealed to by Norris et al. are extremely important. However, the arguments are only as strong as the linking hypotheses between the model and the data. Norris et al. have simply not made a compelling case that feedback is unnecessary in the architecture or in the simulations used to support their Merge model.

ACKNOWLEDGMENT
This work was supported by a grant NSF SBR-9729095.

Why not model spoken word recognition instead of phoneme monitoring?

Jean Vroomen and Beatrice de Gelder

Department of Psychology, University of Tilburg, 5000 LE Tilburg, The Netherlands. **j.vroomen@kub.nl**
cwis.kub.nl/~fsw_1/psychono/persons/jvroomen/index.htm

Abstract: Norris, McQueen & Cutler present a detailed account of the decision stage of the phoneme monitoring task. However, we question whether this contributes to our understanding of the speech recognition process itself, and we fail to see why phonotactic knowledge is playing a role in phoneme recognition.

Psycholinguistics is a strange research domain. Once, the noble aim was to understand human language processing, or, more in particular, to understand how humans recognize words when they hear sounds. There was no obvious way to tackle that question because spoken language processes themselves were not particularly designed for introspection or any other direct method. Psycholinguists therefore invented clever tasks like phoneme monitoring and lexical decision. These tasks, so was the idea, would allow one to tap the underlying processes and deliver the data on which models of speech recognition could be built. TRACE (McClelland & Elman 1986), and indeed Shortlist (Norris 1994b) are an example of that. With the present work of Norris et al. though, it seems that the focus has been shifted from trying to understand spoken word recognition toward trying to understand the ingenious methods that psycholinguists come up with. We wonder whether this move will lead towards a deeper understanding of the speech recognition process.

A decade ago, the relation between data and theory was straightforward. For example, in TRACE there was a bank of phoneme detectors that mediated between articulatory features and words. The (too) strong assumption was that the activation level of a particular phoneme was reflected in the time a subject needed to detect that specific phoneme. One could have anticipated that this assumption was a bit of an oversimplification. At that time, it was already well known that the phoneme was, at least to some extent, an invention, and not so much a natural concept. Different populations with little knowledge about the alphabet (young children, dyslexics, illiterates, Chinese, and other non-alphabetic readers) were unable to explicitly represent speech as a concatenation of phonemes, yet did not have any apparent difficulty recognizing spoken words (see, e.g., Bertelson 1986 for a review). A task like phoneme monitoring requiring an explicit decision about the presence of a phoneme could thus be expected to be related with alphabetic reading instruction, but not so for spoken word recognition.

Norris et al. now formalize this distinction in a model that segregates recognition of phonemes from decisions about phonemes. They make a strict distinction between phoneme recognition units and phoneme decision units. Decision units are very different from recognition units. Decision units are strategic, they are made on the fly, they receive information from the word level, and they have inhibitory connections. None of those properties is shared by phoneme recognition units. Phoneme recognition units are what