Vroomen, J., & de Gelder, B. (2000). Why not model spoken word recognition instead of phoneme monitoring? Behavioral and Brain Sciences, 23, 349-350.

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 they always were: they are assumed to mediate between the speech signal and words. In fact, almost nothing is said in Norris et al. about recognition units that has not been said previously. In our view, this is disturbing if the ultimate goal is to understand speech recognition, and not phoneme monitoring, lexical decision, or whatever other task psycholinguists have invented or will invent in the future.

One can of course argue that it pays to understand the tools one is working with. In this particular case, it was the decision stage in the phoneme monitoring task that troubled our view. Basically Norris et al. argue that we have been misled and that many of the feedback phenomena occurred at a task-specific decision stage. This may well be correct, but it should be realized that this task specific decision stage is also the least interesting part of the word recognition process. Indeed, the phoneme decision stage is in fact superfluous. One can recognize words without phoneme decision units: Decision units only exist because the experimenter told a subject to perform a task with phonemes. In our view, there is a distinction between being critical about a task and knowing its weaknesses versus modelling its weaknesses. Why should one model that aspect of a task which is ultimately the least informative? Would it not be better to try instead to model spoken word recognition?

The ultimate question, in our view, is what has been learned from Norris et al.'s model about speech recognition itself. The architecture they propose is a straightforward one: Phonemes activate words, and words compete. The main argument for the absence of feedback from word recognition units to phoneme recognition units is a logical one: Phonemes are already recognized fast and accurately, and sending information back from words to phonemes simply does not improve word recognition. So far, this may well be correct, but Norris et al. make a surprising exception to this strictly bottom-up process. They allow "lower"-order statistical knowledge about transitional phoneme probabilities to play a role in phoneme recognition. To us, this seems a strange move in a strictly bottom-up recognition process. First, it seems to be a matter of arbitrary labels to call transitional phoneme probabilities "low," and lexical feedback "high." There is nothing inherently low or high in any of these kinds of information. Maybe one is precompiled, the other is computed online, but the principle is that in both cases information from a different source than the speech signal itself enters the recognition process. It is difficult, then, to understand on what principle the distinction is based: why is lexical information excluded, but not transitional probabilities?

Second, it seems at least debatable whether transitional phoneme probabilities will help phoneme recognition if, as argued before, phonemes are already recognized fast and accurately. Are phonemes recognized fast and accurately because the speech signal itself is processed efficiently, or because of the help of transitional probabilities? Third, how is the transitional knowledge about phonemes learned if not by some form of feedback to the phoneme recognition stage? Finally, instead of using phoneme-sized units, why not have higher-order recognition units like syllables that already incorporate the transitional phoneme information?