Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop, 5–8 September 2003, Göteborg University, Sweden.
Robert Eklund (ed.), Gothenburg Papers in Theoretical Linguistics 90, ISSN 0349–1021, pp. 31–34.

Is a perceptual monitor needed to explain how speech errors are repaired?

Peter Howell

Department of Psychology, University College London, UK

Abstract

Kolk & Postma [6] proposed, following Dell & O'Seaghdha [1], that when a speaker chooses a word, phonologically-related words as well as the intended word are activated. Initially, the activations of all these words are similar, though eventually the intended word reaches a higher asymptotic value when activation is complete [1]. According to Kolk & Postma [6], if a response is made in the phase where activation is building up (rather than at full activation), there is a higher chance of the competing, rather than the intended, word being selected (i.e. an error). They propose that a speaker detects such errors when they are produced overtly using the perceptual system, and a monitor in the linguistic system responds by interrupting and initiating the correction [6].

Word repetition and hesitation (not errors in themselves) have been regarded as signifying underlying errors that are detected and interrupted before speech is output in a similar way to overt errors. An assumption in [6] is that activation for a word stops (or, if it continues, is ignored) immediately a candidate word is selected. The brain processes responsible for speech production have massive parallel capacity. Consequently, activation for all the candidates for a word slot could continue beyond the point where a word is selected in cases where a word is responded to prematurely. When the selected word reaches asymptote, the relative activations of this and the other candidate words indicate when an error has occurred (when the selected word has a lower activation than one of the competing words), and what correction is appropriate (the word with the highest activation). This provides the basis for error detection and correction without the need for a perceptual monitor. Continuing the buildup of activation after a word has been selected, implies that activation of nearby words in its phrase overlaps. It is shown, with some realistic assumptions about how activation builds up and decays across different words in a phrase, that this model predicts word repetition and hesitation and also part-word disfluencies (a characteristic of stuttering), again without the need for a perceptual monitor.

1. Investigation

Levelt's work has provided an enormous impetus to research on disfluencies in spontaneous speech. His 1989 model [7] has a very wide scope and has been the imprimatur of many other models. It has set the standard in the sense that an adequate model of disfluencies in spontaneous speech must aspire to explaining most, if not all, the phenomena Levelt has explained.

One feature, shared between Levelt's model and many others (including modular and interactive variants) is that generation of speech output is hierarchical, involving lexical and phonetic steps. Fluent speech control arises when all the steps are accurate. Conversely, disfluency occurs when any level in the hierarchical system malfunctions and gives rise to an error. Levelt's model uses mechanisms outside the production processes to recover after such errors. There are two connections to the outside processes in Levelt's model that allow this: 1) The last step in processing in Levelt's hierarchy is to place the results in a phonetic output buffer and the string is then sent to articulation processes. The articulation processes produce sound that is picked up by the auditory system that sends its information to the perception system (external loop). 2) Information about processing within linguistic planning is transmitted as it is generated to the speech perception system (internal loop). The information sent via the internal and external loops is deciphered by the speech perception system, and the results are sent to a monitor in the linguistic system that detects mismatches between the intended output and that achieved (i.e. whether an error has occurred). If an error has occurred, speech is interrupted and reinitiated. The problem with 1) is that it implies a particular model of the language-speech interface. This interface relies on auditory and speech perception mechanisms to detect whether one's own speech is accurate, which available data suggest may not be possible. The problem with 2) is that, if true, it operates in a way that makes the events that it detects (the errors) unobservable. Consequently, all the support for this process is indirect and questionable for this reason.

An important line of evidence that led Levelt to propose external and internal loops is his account of the pattern of recovery after a speaker has made an error (referred to as "repair"). An example of a repair is "in the back, in the front of the...". This utterance contains an overt error of lexical selection (back for front) that may have been detected over the external loop. According to [8], the monitor detects this error and interrupts speech (signified by the comma). Two words are repeated that occurred prior to the word in error ("in the") that are referred to as a retrace, and then the speaker makes the correction. An example like "in the, in the front" (according to [8]) is a covert repair, which might have occurred because the speaker made the front-back error but detected it over the internal loop, and interrupted the speech before it was output. Covert repairs were characterised by Levelt by interruption and retrace features: They consist of "either just an interruption plus editing term [words like "no" said after the pause], or the repeat of one or more lexical items" ([8], p.55). Subsequent authors, such as Hartsuiker & Kolk [3], have classified speech events with short overt errors as covert repairs with the short section of overt error being attributed to inertia in stopping on-going speech. This clearly does not fit with Levelt's definition and affects estimates of overt and covert events (some repairs Levelt would class as overt are reclassified as covert). Until the definition that allows overt errors to occur in covert repairs is defended, Hartsuiker and Kolk's simulations of the operation of the internal loop should be ignored (they could be simulations of the operation of the external loop).

Levelt's work has given license to certain terms in the area that are appropriate for those working within his own framework, but not for those taking different theoretical approaches. "Repair", "monitor" and "feedback" are three value-laden terms that connote a specific way of dealing with errors. If there is no observable error (as when there is only timing disruption), referring to these events as covert "repairs"

is inappropriate, and some neutral term for these events should be employed. If the identity of a putative error cannot be given, there is no way of specifying what the feedback is. The operation of a monitor cannot be specified if it is not known what "feedback" is telling this process.

Elsewhere I have examined the problems with a perceptual monitoring proposal [4, 5] that further underline my own reasons for dismissing the terms Levelt's work licences. These arguments will not be repeated here except to say they raise concerns about whether any external processes are coupled to the production system in the way Levelt describes. Instead, I want to examine the implications of one theory that derives from Levelt's work, Kolk & Postma's covert repair hypothesis (CRH) [6]. CRH is an account that has been applied to fluent speech control. CRH has also been applied to stuttering, a disorder where speakers have a high proportion of disfluencies. For this reason, stuttering is used as a test case for models of disfluency. Kolk & Postma [6] used Dell & O'Seaghdha's [1] spreading activation model to explain how a slow phonological system leads to speech errors. According to [1], when a speaker intends to say the word "cat" (the target unit), phonologically-related competing units are also activated. (e.g. "rat"). Dell & O'Seaghdha [1] have steps involving lexical activation and phonological encoding. Overall activation represents the interaction between these processes in the original model [1], but Kolk & Postma focus on how activation patterns in the model can lead to phonological errors after lexical selection has taken place (this seems reasonable as fluent speakers are accurate at lexical selection on 99.99% of occasions [2]). The buildup of activation for the target and competing units, follow similar trajectories in early epochs, but later in time they asymptote at different levels (see Figure 1). At asymptote, the target unit has the higher activation level, which generates the appropriate word as response (points to the right of "*S*" in Figure 1). Operating under time pressure (such as when speech has to be produced rapidly) requires a speaker to generate words in the period where activation is still building up, for example at points near "*S*-" in Figure 1. The word response at this point would still be the one with highest activation. However, as the target and competing options have similar activation-trajectories during build-up, by chance one of the competing options may have highest activation and be triggered (resulting in a speech error) if word selection is made in this time-region. Speakers who have slow phonological systems (as Kolk & Postma propose to be the case in speakers who stutter) will extend the amount of time in the build-up phase. A word response generated in the extended build-up phase, has a heightened chance of a speech error arising for the same reason as a speech produced under time pressure.

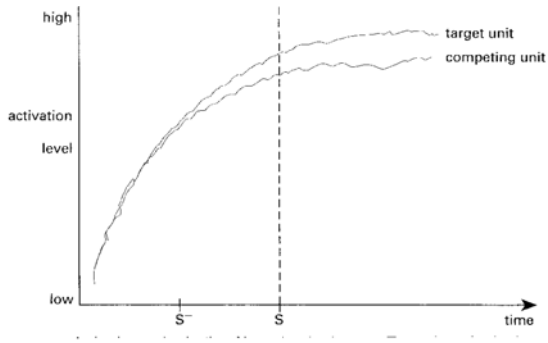

Figure 1: Activation versus selection. Two points of selection normal (*S*) and early (*S*-) are shown.

Kolk & Postma's account effectively involves imposing a decision rule for response selection (choose the candidate with the highest activation level at different imposed deadlines, empirically, at different points along the abscissa in Figure 1). The decision rule is arbitrary but defendable. The questionable issue, it seems to me, is why should activation stop building up at the point at which response selection is made when a response is made early? As Figure 1 shows, the trajectory of activation buildup beyond the deadline at which early response selection is made (points to the right of *S*), lead to the correct (target) unit having highest activation. Effectively truncating activation buildup at the deadline loses the information obtained up to this point, whereas, if processing continued for a short time, it would be clear that the word produced was in error. Continuing activation for this short time seems less costly than routing information through the perceptual system to the monitor that then interrupts and restarts speech as in CRH. Put simply, the monitoring system (internal and external loops, perceptual system and monitor) would not be required if activation was allowed to continue after the response was selected.

At asymptote, all candidate phones are fully activated. What would it mean, then, for a response to be initiated before activation level reaches asymptote? One way of looking at this issue is in terms of Levelt's phonetic output buffer [7]. A phonetic buffer with five slots for phones is shown in Figure 2. Activation is complete for the first two (shaded line), but not for the final three, though activation over all phones is above the minimum threshold that is required for activating production of a word. Buildup of activation will stop once all candidate phones are fully activated. Conversely, the plan is only partial when any of the slots is not fully activated.

Figure 1 shows that to ensure the plan is complete and guarantee no error, activation has to be at asymptote. If a word is selected and produced before asymptote, and activation of a competing word is higher at asymptote, the latter should have been the target (i.e. an error occurred), then that plan is available, can be substituted and yield the correct response immediately. A simple threshold process would automatically select words that need to be substituted ("repaired") because an error was made (only competing words that have activation levels that poke above the activation level of the word that was produced are in error and should be changed). In this way, error correction can occur without a perceptual monitor.

According to the proposed model, speakers have the wherewithal to detect and correct errors within the production system. Whilst this ability is built into the model, it should be noted that it is rarely called on (only 0.01% of words are in error [2]).

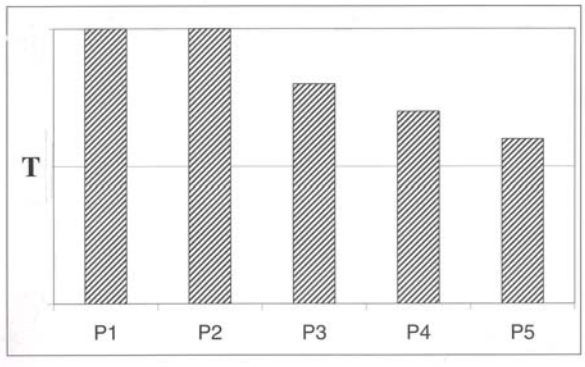

Figure 2: Activation states (shaded area) for five phones in a phonetic output buffer.

Extensions to the model are needed so that it can address the issues, a) how the features of covert repairs (hesitation and word repetition) arise, and b) how disfluencies on part of a complex word arise (these are features that are associated with, but not exclusive to, the speech of speakers who persist in their stuttering). The extensions are: 1) activation for words in a phrase takes place in parallel with the activation-onsets of words offset according to their order of appearance in the utterance, 2) activation builds up at different rates for words of different complexity, 3) activation begins to decay once a plan is completed, 4) (as a consequence of 3), when a word is initiated on the basis of a complete plan, some decay will occur after planning is complete during the time the word is being executed. When a word is initiated on the basis of an incomplete plan, activation will continue to build up after planning is complete during the time the word is being executed. In cases both where buildup for a word is or is not complete, activation for future words will be building up.

In the remainder of this paper, I will show how these four properties explain issues a) and b), for the phrase "in the spring". Using 1), planning involves activation building up for the words and these overlap in time, though the first word starts building up before the second and so on (i.e. the plans start at different offsets). So, activation starts to build up for "in" first, then "the" and finally "spring". Buildup of the phonetic output in a word also progresses left-to-right. Using 2), activation of words of different complexity builds up at different rates. This arises, to some extent, because of the complexity of the phonetic or phonological makeup of words. "in" and "the" build up rapidly as they have simple structure, whereas "spring" has a complex onset so its activation builds up more slowly (this most likely arises through phonological, as opposed to phonetic, influences).

The situation for fluent speech is considered now. The first word will start to decay when it starts to be executed (3 and 4), assuming its execution started with a complete plan. The build-up in activation continues for the subsequent words and, given the decay of the first word and offsets for activation of successive words (1), the next word in the sequence will be the one with maximum activation. The process continues, assuming the activation of successive words on completion of the current one is at or near that representing a complete plan (as would be the case in a well-configured biological system for fluent speech production).

The way disfluencies, that have been described as parts of covert repairs, arise is described next. The activation pattern at the time "the" has been spoken is shown in the centre panel of Figure 3. Activation patterns of the other two words over the same interval of time (*Tx* to *Ty*) are also shown. "in" (left panel) had built up to maximum previously, but activation by time *Ty* has dropped right off. "the" has also been at maximum and is showing decay during time for its execution (far less than for "in"). Rate of activation buildup for "in" and "the" was more rapid than for "spring" (right panel) which has the complex onset (property 2 above), and this is shown as having a gentler slope. At *Ty* (i.e. at the time "the" has been produced), the plan for "spring" is not complete, though some activation has occurred (this could be similar to that represented by the state of the phonetic buffer in Figure 2 with only the first phones complete). A threshold rule (produce the word whose activation is above *T* in Figure 3) would lead the speaker to repeat "the" in this case. A lower threshold that is still above that achieved at *Ty* by "spring" or a more rapid execution rate (that allows less time for decay of "in") could leave both "in" and "the" above threshold. In this situation, both words would be above threshold and the speaker would produce "in the, in the,". Activation for "spring" can continue during either of these examples of repetitions and can lead to enough time for the plan for "spring" to be completed [4, 5], its threshold to be above *T* and it would be produced. Essentially, the overlapping activation patterns permit word repetition when they precede a word with a complex onset (usually a content word in English). Pauses would arise when "in" and "the" have decayed below threshold (due to threshold and rate parameters again), and "spring" has not reached *T*. Such word repetition and hesitation, that Levelt and Kolk & Postma took as evidence for corrections to errors detected over the internal perceptual loop, arise in the proposed model from overlapping activation patterns and the decay and threshold parameters that apply to the activations in production.

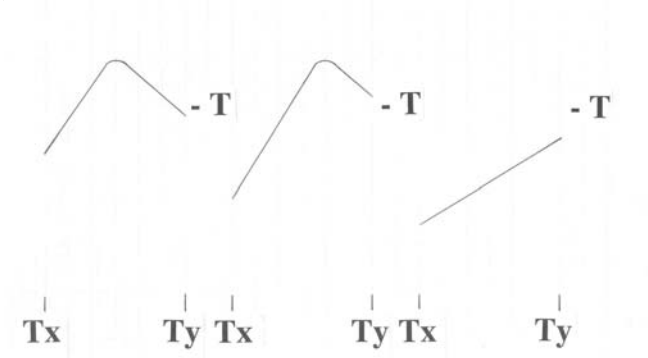

Figure 3: Activation patterns for the three words in the test utterance each shown for the interval of time *Tx* to *Ty*. *Ty* is after execution of the second word and represents a situation that will lead to word repetition.

The situation can arise, depending on threshold value, speech rate or rate at which activation builds up (phonological complexity), where the two initial words in the phrase have decayed to values lower than *T*, and the third word is at or above *T*, but its plan is not complete. Such a situation is shown in Figure 4. Execution of this word can commence at the requisite time. Some plan still needs to be completed (usually the later phones are the ones that will not be complete, as shown in Figure 2). The plans can be completed in the time taken to execute the first part (and this will usually be the correct word [2]). If the plan runs out, only the first part of these words can be produced (part-word disfluencies at onset). These are characteristics of persistent stuttering [4, 5].

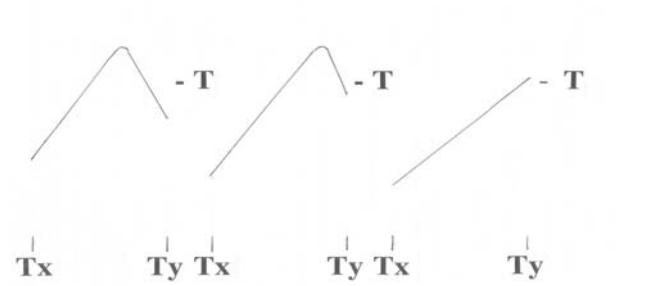

Figure 4. Activation patterns for the three words in the test utterance each shown for the interval of time *Tx* to *Ty*. *Ty* is after execution of the second word and represents a situation that will lead to part-word disfluency involving the onset of the third word.

The point of this exercise has been to show errors, word repetition and hesitation, and part-word disfluencies can arise in a spreading activation model without a perceptual monitor. The model is based on some reasonable assumptions about activation buildup and decay in phrases. As a perceptual

monitor has been discarded, this poses a challenge to CRH. Some differences relative to Dell & O'Seaghdha (the model on which [6] was based) should be noted. These authors include lexical activation that is set to zero immediately after lexical selection. The buildup patterns Kolk & Postma show are solely phonological. Dell & O'Seaghdha were modeling priming data and, therefore, they did not consider what happened to activation after response initiation. Although the current work relies on Kolk & Postma's phonological activation profiles, I want to emphasize that I do not want to commit myself either to a view that phonological activation is all that is important in leading to disfluency or to a different model in which phonological and lexical activation build up in the way Kolk & Postma propose which, for instance, differs from that in Dell & O'Seaghdha's model. My view at present is that anything that varies the time-course of activation patterns (e.g. syntax as well as lexical influences) needs to be taken into account in accounting for disfluencies.

Many of the ideas behind this exercise have been taken from the EXPLAN model of fluency control [4, 5] and applied and extended to CRH's representation of word activation. In EXPLAN, speech errors are ignored because they are rare, and fluency failures are focussed on as they are common. In EXPLAN, fluency failures arise because plans are not complete when the word needs to be executed. This leads either to word repetition or part-word disfluencies (the latter mainly in people who stutter). Part word disfluencies are considered problematic events that speakers should avoid. Consequently, a speaker needs to be aware of when this is happening and attempt to avoid it in the future. To achieve this, EXPLAN incorporates a model of the motor processes. Speech motor timing needs to be slowed when part-word disfluencies occur to avoid part-word disfluencies. (Slowing speech timing effectively allows more time for the part-plan to be completed, which is why disfluency is avoided.) How does the speaker become aware that speech timing needs to be altered? EXPLAN's answer is that all you need to do is to determine whether a complete plan was supplied at the point where execution commenced. This can be determined by subtracting the plan at the point in time execution commenced from the plan at the point in time execution is completed. If the whole plan was supplied, the two will be identical, they will cancel and speech will be fluent. If the speaker initiates speech prematurely, more of the plan will be generated in the time taken to execute the first part and the two will differ and speech needs to be slowed. The points in time that execution starts and execution is completed are landmarks in the account how errors arise and are corrected, presented above. Given that the location of these points is needed to account for errors and that the plan at these points in time is needed to determine whether slowing is necessary, the extra requirement in EXPLAN for determining whether slowing speech is needed can be efficiently dealt with by the minor modification of taking a copy of the speech plan at these landmark points.

Slowing is achieved in EXPLAN by sending the information after subtraction to an external timekeeping mechanism that regulates speech timing. This proposal about the speech-language interface fills a similar role in EXPLAN to the external loop in CRH. Howell [4] presents arguments in favor of the EXPLAN proposal about the connection between the speech planning and motor execution processes (as well as arguments against CRH's proposal).

There is a lot of work still to be done to link this work with that of Dell, CRH and EXPLAN. For instance, how can the fact that are children who stutter more likely to repeat function words whereas older speakers who stutter are likely to produce content word disfluencies, be explained in the current model? The answer could be either: 1) that function word activation decays more rapidly in older speakers who stutter than younger ones and fluent speakers, or 2) content word activation starts to build up at the same rate in adults who stutter as with older speakers, but plateaus for some reason (e.g. problems at the juncture between onset and rhyme).

In summary, the occurrence of errors (on the rare occasions they happen) have been explained, word repetition and hesitation (features of covert repairs) and aspects of stuttering accounted for after perceptual loops and a monitor have been discarded.

2. Acknowledgements

This work was supported by a grant from the Wellcome Trust.

3. References

[1] Dell, G. S. & P. O'Seaghdha. 1991. Mediated and convergent lexical priming in language production: A comment to Levelt et al. *Psychological Review*, vol. 98, pp. 604–614.

[2] Garnham, A., R. C. Shillcock, G. D. A. Brown, A. I. D. Mill & A. Cutler. 1981. Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics*, vol. 19, pp. 805–817.

[3] Hartsuiker, R. J., & H. H. J. Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. Cognitive Psychology, vol. 42, pp. 113–157.

[4] Howell, P. 2002. The EXPLAN theory of fluency control applied to the Treatment of Stuttering by Altered Feedback and Operant Procedures. In: E. Fava (ed.), *Current Issues in Linguistic Theory series: Pathology and therapy of speech disorders*. Amsterdam: John Benjamins, pp. 95–118.

[5] Howell, P. & J. Au-Yeung. 2002. The EXPLAN theory of fluency control and the diagnosis of stuttering. In: E. Fava (ed.), *Current Issues in Linguistic Theory series: Pathology and therapy of speech disorders*. Amsterdam: John Benjamins, pp. 75–94.

[6] Kolk, H. H. J. A. & Postma. 1997. Stuttering as a covert-repair phenomenon. In: R. F. Curlee & G. Siegel (eds.), *Nature and treatment of stuttering: New directions*. Boston: Allyn & Bacon, pp. 182–203.

[7] Levelt, W. J. M. 1989. *Speaking. From Intention to Articulation*. Cambridge, Massachusetts: MIT Press.

[8] Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition*, vol. 14, pp. 41–104.