Is Comprehension Necessary for Error Detection?
A Conflict-based Account of Monitoring in Speech Production

Nazbanou Nozari

Ph.D. University of Illinois at Urbana-Champaign, May 2011

Humans function efficiently despite being error-prone, because they can monitor their behavior, detect errors and apply corrections online. Sophisticated cognitive models of action monitoring have been proposed as the result of studying error detection and post-detection correction (e.g. Botvinick, Braver, Carter, Barch, & Cohen, 2001). However, the data for these models come from laboratory tasks with a limited number of pre-determined button-push responses, and little effort has been made to generalize these mechanisms to natural cognitive tasks such as speaking. Moreover, cognitive monitoring has made little contact with monitoring motor movements.

This lack of communication has serious drawbacks. A good example is that the most widely accepted theory of monitoring in speech production (the *perceptual loop*; Levelt, 1983, 1989) is purely perceptual, which makes it fundamentally different from models of monitoring based on forced-choice cognitive tasks or models proposed in motor control (e.g. *forward models*). This is cause for concern, since the evidence strongly points to commonalities between monitoring in different systems. For one thing, behavioral studies have demonstrated that the timeline of detection and correction is incompatible with monitoring through external perceptual systems for both limb movements and language production. Moreover, event-related brain potential (ERP) studies have discovered a waveform that appears to be correlated with the detection of performance errors; this component is similar regardless of the system committing the error (e.g., motor movement, visual processing, language production), and also independent of the modality that commits the error (button-push, eye-movement, verbal). These domain-general effects require a domain-general theory of monitoring.

This dissertation was motivated primarily by the fact that the current theory of speech monitoring fails to explain important facts about error detection. The research presented here argues that although the evidence against the model comes from the domain of language, the model's failure is a domain-general problem. It further shows how implementing a simple, but crucial, principle borrowed from models of action monitoring can solve the problem. Specifically, instead of relying mainly on perception for error detection, the newly-proposed model uses the information generated by the production system (as do the forward models of limb movement monitoring). A specific form of a non-perceptual monitor (Conflict Detector, Botvinick et al., 2001) is selected among the cognitive models of forced-choice error detection, because of the support it receives from the neurophysiological studies (ERP) and the similarities of those

findings in linguistic and non-linguistic tasks. The new theory of conflict-based error detection in speech production is then tested computationally, in a well-established connectionist model of word production (Dell & O'Seaghdha, 1991). Once the model's predictions about error detection in normal speakers are verified, the model is further tested on the data obtained from a group of chronic aphasic patients with left hemisphere damage due to stroke. The resulting model is the first model of speech monitoring that explains error detection behavior in both healthy speakers and brain-damaged patients, and is at the same time, compatible with the evidence of domain-general monitoring. More generally, this research is a successful example of developing a new theory, by integrating information from multiple domains of neuropsychology, neurophysiology, cognitive and motor control, and computational modeling.

The dissertation has 5 sections. First the current theory of speech monitoring (the perceptual-loop model) is reviewed, along with its shortcomings. Next a model of error detection in forced-choice tasks is discussed, in which detection of errors is carried out by monitoring the amount of conflict between response alternatives. Third, a formal of model of error detection in language production is proposed, which uses conflict detection, but tailors it to the properties of the language production system. This section contains two computational simulations testing the principles of the model. Fourth, a coding scheme is developed for coding natural error detection in aphasic patients, by using transcriptions of the responses of 63 patients in a picture naming task. The data of 29 of those patients' –who matched the inclusion criteria– are then analyzed to test the model's predictions derived from the computational simulations. In the final section, the main properties of monitoring models in different fields are discussed. It is then argued that although the new conflict-based model has module-specific features that make it suitable for detecting errors in language production, it shares the core features of the models created for other systems, and is thus compatible with the evidence of domain-general monitoring. In this précis, I follow the same organization, and provide a brief summary of each section, leaving out the details that can be consulted in the main body of the dissertation attached.

Monitoring in language production

Detecting errors in another person's speech, by default, requires listening, comprehending and analyzing their speech for discrepancies or unexpected utterances (e.g. upon hearing "I read a dog" the semantic incongruence of the sentence raises the possibility that the speaker has committed an error). It is tempting to extrapolate this mechanism to monitoring one's own speech, and claim that speakers listen to their own utterance and analyze its content for discrepancies with the intended message. Only, they do not have to wait for the utterance to become overt, since they have access to the constructed utterance in the form of "inner speech" (i.e. speech before it is articulated). According to this view, monitoring one's own speech

is carried out by the comprehension system via two routes, an internal channel, which monitors speech before it is spoken and an external channel which monitors speech after it is spoken, pretty much in the same manner as it does others' speech. This model, called the *perceptual loop* (Levelt, 1983, 1989) has been the dominant model of monitoring speech production for nearly 30 years. In spite of its success in explaining many aspects of error detection, the model faces a big challenge: There are several reports of aphasic patients with near-perfect comprehension who fail to detect their own –but not others'– speech errors (e.g. J. Marshall et al., 1998) as well as patients who do detect their speech errors in spite of impaired comprehension (e.g. R. Marshall, Rappaport and Garcia-Bunuel, 1985). This double-dissociation between comprehension and error detection is a serious problem for a comprehension-based model of error detection.

An alternative would be a model that instead of relying on comprehension for error detection would rely on the production system itself. Examples of such production-based monitoring models have been proposed, but were either never implemented (e.g. De Smedt & Kempen, 1987; Laver, 1980; Postma & Kolk, 1993) or their implementation has been unsuccessful in accounting for some of the fundamental empirical data (e.g. MacKay's Node Structure Theory, 1987, 1992), and have thus failed to supplant the perceptual loop model.

Conflict model of error detection in forced-choice tasks

Dissociation of comprehension from error detection also has support in the ERP data. The neural signature of error commission is a negativity with frontocentral distribution, called the Error Related Negativity (ERN; e.g. Gehring, Goss, Coles, Meyer, & Donchin, 1993). Crucially, the emergence of the ERN has been shown to be independent of conscious awareness of the error, meaning that ERNs were detected after both errors that participants reported and the ones that they did not (Nieuwenhuis, Ridderinkhof, Blow, Band, & Kok, 2001). In addition, the ERN originated from the same brain region (Anterior Cingulate Cortex or the ACC) regardless of the system in which the error was committed (motor, language, visual), and also regardless of the response modality (hands vs feet; e.g. Holroyd, Dien, & Coles, 1998). These findings suggest that commission of an error in a system sends a signal to a central brain region (most likely the ACC), which generates a domain-general error signal (the ERN), regardless of the system from which the error has originated.

An example of such a monitoring mechanism was proposed by Botvinick et al. (2001), for forced-choice button-push tasks. According to their theory, the conflict between the activation of response alternatives is monitored by the ACC, and larger values of conflict are associated with higher error probabilities. Hence, detection of an error is blind to the actual "correct" response, and does not require comprehension of the

executed response, but instead relies on the amount of computed conflict over a number of response alternatives.

The Conflict-based model of error detection in language production

The conflict-based model described by Botvinick et al. (2001) has all the right properties for a production-based model of error detection in language production. As noted above, the model acts independently of comprehension and relies on the conflict of activation of more than one response. Dell and O'Seaghdha's (1991; See Figure 1 in the dissertation) word production model provides a good scaffold for implementing such a conflict-based monitor. The model has three layers (semantics, lexical nodes and phonemes) and names an object in two steps. In the first step, the semantic features of the object become activated, the activation spreads throughout the network and the most highly activated node in the lexical layer is selected (selection point 1). In the second step, the selected node receives a jolt of activation and, after spreading the activation throughout the network, the most highly activated nodes are selected at the phoneme layer (selection point 2).

An advantage of the model is that it has been successfully used to model aphasic language production (e.g. Nozari, Kittredge, Dell, & Schwartz, 2010), by lesioning the weights of the connections between the three layers. Specifically, the weight of the connections between the semantic and lexical nodes (s-weight) determines the probability of making a semantic error (e.g. "dog" instead of the target "cat") and the weight of the connections between the lexical and the phonological nodes (p-weight) determines the probability of a nonword error (e.g. "zat" for "cat"). Therefore, given the response pattern of each patient on a picture naming task (i.e. the proportions of various kinds of errors), the model assigns unique s and p values to them. This is important because the ultimate test of the conflict-based model of monitoring has to come from the aphasic patients, who may or may not show a dissociation between comprehension and error detection abilities. Dell and O'Seaghdha's model provides a way to quantify production ability, which can then be objectively compared with comprehension scores for predicting patient's error detection abilities.

Although the idea of conflict detection was borrowed from Botvinick et al. (2001), the implemented model differs from theirs in important ways, and cannot be viewed as a simple implementation of that model in language production. For one thing, detection of conflict is tied to the selection points (see above), such that conflict detection is viewed not as an extra mechanism useful only for error detection, but an integral part of the process of selection for production. Thus, conflict is measured once at the level of lexical and once at the level of phonological nodes at the points of selection.

Moreover, the measure of conflict is also changed to reflect this theoretical choice. Instead of Hopfield energy (which requires tracking changes of the conflict measure throughout each trial), two new measures are used: standard deviation of the activation of all response alternatives (to reflect competition from ALL responses) and the difference between the two most highly activated nodes (to denote competition from the strongest competitor). Both measures yielded a similar pattern, but the latter produced stronger signals and is thus a better predictor of the occurrence of an error.

Successful detection by conflict in the model requires that three principles hold:

(1) Detection sensitivity. The amount of conflict must be predictive of the probability of error occurrence.
(2) Layer specificity. Conflict at each layer of the system should specifically predict the error type arising from that step. Therefore, conflict at selection point 1 (lexical layer) must be predictive of the occurrence of semantic errors and conflict at selection point 2 (phonological layer) must be predictive of the occurrence of nonword errors.

(3) Integrity contingency. The main objection to the comprehension-based monitor was that monitoring failure did not always parallel failure of comprehension. Generally, a monitor is expected to fail when the system underlying its operation is severely impaired. The principle of integrity contingency predicts that, regardless of the status of the comprehension system, when the production system is broken down, monitoring must also be impaired.

The first two principles were tested using computational simulations of a normal speaker. In addition, I showed that the model achieves hit rates comparable to the values reported in the literature for error detection in everyday speech, with a reasonable false alarm rate. In the published version of this dissertation, my co-authors and I augmented the model with a precise theory of how the model learns to set its conflict criterion level for detecting an error (Nozari, Dell, & Schwartz, 2011). In order to verify the third principle –and recheck the first two- I simulated five aphasic patients with varying degrees of impairment in the s and p weights (Table 1 in the dissertation), and showed that the conflict-based model predicted a correlation between the strength of the s and p weights and the efficiency of error detection for semantic and nonword errors respectively. I also verified that if the production weights are too low (i.e. extensive damage to the production system) the arbitrariness of the conflict signal limits its usefulness for error detection, and in such cases monitoring fails.

In summary, the simulations showed that monitoring conflict at the layers of the production system is a good predictor of error occurrence. Similar to the other conflict-based models, the conflict signal can then

be sent to the error detection center (e.g. the ACC) to produce the ERN observed at the time of error commission. Next section tests the predictions of the model using real-life data.

Error detection in a sample of aphasic patients

Aphasic patients provide a better test for the model than healthy speakers, because (1) they make many errors even when naming pictures of individual objects, and (2) they commit varying numbers of both types of errors (semantic and nonword errors) necessary to test the model's *layer specificity* principle, while healthy speakers rarely make nonword errors. In short, patients vary a great deal in their production abilities and disabilities, thus providing variance on the key independent variables. The model makes a specific prediction: conflict at the lexical layer should be predictive of the occurrence of semantic errors, and conflict at the phonological layer should predict the occurrence of nonword errors. Since the amount of conflict at the lexical and phonological layers is correlated with the strength of the s and p weights respectively, the theory predicts a correlation between the strength of the s weights and detection of semantic errors, and the strength of the p weights and detection of phonological errors.

The first step was to develop guidelines for coding error detection to be applied across patients. Transcripts from 63 aphasic patients' picture naming attempts were used to develop a coding scheme, which can be found as an attachment to the dissertation. To my knowledge, this coding scheme is the first extensive and demonstrably reliable coding scheme for aphasic error detection. Once this scheme was refined to the degree that each response could be coded unambiguously by two independent coders, the data from twenty-nine of the 63 patients, who passed the inclusion/exclusion criteria, were coded using its final version. The key model-derived predictions were then tested. As predicted, detection of semantic errors correlated significantly with the strength of the s weights ($r = .59$, $p = .001$), and the detection of phonological (nonword) errors with the strength of the p weights ($r = .43$, $p = .02$). Importantly, I found little evidence of correlation between error detection and measures of comprehension using four standard comprehension tests. These tests were chosen to reflect comprehension at the different levels of the system: comprehension at the semantic level was measured using Pyramids & Palm Trees, comprehension at the lexical level was measured using the Synonym Judgment-Noun and PPVT-III tests, and finally comprehension at the phonological level was assessed using the Phoneme Discrimination task. The correlation between detecting semantic and phonological errors and none of these test scores exceeded 0.24. I checked these results using a hierarchical mixed model with random effects, which drew similar conclusions; production weights were predictive of error detection in a layer-specific fashion (s-weights predicting semantic error detection; p-weights predicting nonword error detection), while comprehension scores were not.

Next I showed the double dissociation between comprehension and error detection, and the association between each production weight and detection of a specific error type at the level of individual patients, by analyzing in detail four individual patients who had a sufficient number of errors to allow for consideration of them as case studies. In all of these patients, error detection performance was unexpected (too high, or too low) for their comprehension scores, but was well predicted by the strength of their production weights, as the conflict-based model would predict.

Domain-general principles of monitoring for error detection

Errors in a cognitive, perceptual, or motor system reflect specific properties of that system. For example, the architecture of the language production system predicts semantic and phonological errors, while an equivalent may not be found in, say, arm movements. In this sense, monitors are domain-specific to the degree that their implementation depends on the properties of the system being monitored. However, this domain-specificity should not be taken too far. A fundamental principle of monitoring is that perceptual processes cannot be the primary mechanism of monitoring (see behavioral and neurophysiological evidence mentioned earlier). This principle is implemented in both forwards models and the conflict-based models, which instead use the information within the error-generating system to monitor the same system. This means that there is no standard of correctness provided by an external module to be compared to the output of the system under monitoring, but the comparison takes places between two or more alternatives that are activated by the production system, even though the details of the comparison processes differ.

The perceptual loop model of monitoring in language production does not operate according to this principle, and the model's failure in accounting for the dissociation between error detection and comprehension is a reflection of this. Therefore, a natural first step in developing the new model was to ensure its compliance with this domain-general principle. The model was then detailed to reflect specific properties of the language production system. The resulting conflict-based model of speech error detection is the first to explain error detection performance where the old theory fails, and at the same time, be compatible with a variety of findings from the behavioral and ERP literature in linguistic and non-linguistic error detection tasks.

References

Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. D. (2001). Evaluating the demand for control: Anterior cingulate cortex and crosstalk monitoring. *Psychological Review, 108*, 624–652.

Dell, G.S., & O'Seaghdha, P.G. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1991). *Psychological Review, 98,* 604-614.

De Smedt, K., & Kempen, G. (1987). Incremental sentence production, self-correction, and coordination. In G. Kempen (Ed.), *Natural language generation: recent advances in artificial intelligence, psychology, and linguistics* (pp. 365-376). Dordrecht: Martinus Nijhoff Publishers.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science, 4,* 385–390.

Holroyd, C. B., Dien, J., & Coles, M. G. H. (1998). Error-related scalp potentials elicited by hand and foot movements: Evidence for an output independent error-processing system in humans. *Neuroscience Letters, 242,* 65–68.

Laver, J. D. M. (1980). Monitoring systems in the neurolinguistic control of speech production. In V. A. Fromkin (Ed.), *Errors in linguistic performance: slips of the tongue, ear, pen,and hand* (pp. 287-305). New York: Academic Press.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14,* 41–104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

MacKay, D. G. (1987). *The organization of perception and action: a theory for language and other cognitive skills*. New York: Springer-Verlag.

MacKay, D. G. (1992). Awareness and error detection: New theories and research paradigms. *Consciousness & Cognition: An International Journal, 1*(3), 199-225.

Marshall, J., Robson, J., Pring, T., & Chiat, S. (1998). Why does monitoring fail in jargon aphasia? Comprehension, judgment, and therapy evidence. *Brain & Language, 63*(1), 79- 107.

Marshall, R. C., Rappaport, B. Z., & Garcia-Bunuel, L. (1985). Self-monitoring behavior in a case of severe auditory agnosia with aphasia. *Brain & Language, 24,* 297-313.

Nieuwenhuis, S., Ridderinkhof, K. R., Blow, J., Band, G. P. H., Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology, 38 (5),* 752–760.

Nozari, N., Dell, G.S., & Schwartz, M.F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology, 63(1),* 1-33.

Nozari, N., Kittredge, A.K., Dell, G.S., & Schwartz, M.F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language, 63,* 541–559.

Postma, A., & Kolk, H. H. J. (1993). The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. *Journal of Speech & Hearing Research, 36*(3), 472-487.