8
Self-Inhibition and the Recovery Cycle

> Organization . . . is just as important a property of behavior as it is of perceptions. The configurations of behavior, however, tend to be predominantly temporal – it is the *sequence* of motions that flows onward so smoothly as the creature runs, swims, flies, talks, or whatever. What we must provide, therefore, is some way to map the cognitive representation into the appropriate *pattern* of activity.
>
> (G. A. Miller, E. Gallanter, K. H. Pribram, 1960, p. 11)

The present chapter describes a central concept in the node structure theory: the process of self-inhibition. I first examine the likely mechanism of self-inhibition and some theoretical reasons for postulating a self-inhibitory process. I then outline some evidence bearing on the self-inhibition hypothesis, some predictions that follow from self-inhibition, and an application of the self-inhibition hypothesis to the phenomenon of pathological stuttering.

The Process of Self-Inhibition

Self-inhibition is the inhibitory process that terminates the self-sustained activation of mental nodes and temporarily reduces their priming level to below normal or resting state. Like neurons, nodes have an absolute refractory phase, but it is of such brief duration (less than 1 ms) as to be irrelevant. For all practical purposes, self-inhibition introduces only a relative refractory phase in the excitability of mental nodes. Although self-inhibition reduces priming to below-normal levels, the node can still become activated if priming from other (external) sources is strong enough to meet the most-primed-wins criterion. Although self-inhibition is a built-in characteristic of all nodes (see the following discussion), only nodes that become activated rather than just primed during the course of perception and action exhibit self-inhibition. For example, because of the principle of higher level activation, low-level nodes do not exhibit self-inhibition during everyday perception and neither do "uncommitted" nodes (until they finally

Self-inhibition, perceptual feedback and error detection. Ch 8-10 (pp. 141-193) in MacKay, D.G. (1987). *The organization of perception and action: A theory for language and other cognitive skills* (1-254). Berlin: Springer-Verlag.

become activated and committed), even though the self-inhibition mechanism for already-committed nodes plays a role in the node commitment process (D. G. MacKay, 1987).

The Rationale and Nature of Self-Inhibition

Three types of factors call for a self-inhibition mechanism in the node structure theory: empirical factors, specific theoretical factors, and general ("Gestalty") theoretical factors having to do with overall simplicity and elegance of the theory (Fodor, 1980). However, I have not been able to fully develop the general theoretical rationale in the present chapter or even the present book. Self-inhibition forms part of simple and elegant mechanisms for processing self-produced feedback (Chapter 9) and for forming connections between nodes (D. G. MacKay, 1987). In this chapter I examine only the empirical and specific theoretical rationales for building the self-inhibition mechanism into the node structure theory.

The Specific Theoretical Rationale for Self-Inhibition

The theoretical necessity of something like self-inhibition has long been suspected (Feldman & Ballard, 1982; D. G. MacKay, 1969b). Neural activity tends to persist in the absence of inhibition, and unless components in a behavior sequence become inhibited after serving their function, general convulsion could result (Neumann, 1984). The node structure theory illustrates the theoretical necessity of self-inhibition in much more specific terms. Self-inhibition is needed to prevent disruptive effects of internal and external feedback on mental nodes. Internal feedback refers to the bottom-up priming that is transmitted to a superordinate node as soon as one of its subordinate nodes becomes activated. When a mental node becomes activated during production, it primes its subordinate nodes via top-down connections, thereby enabling activation. Activating these subordinate nodes then sends immediate internal feedback (priming) back to the superordinate node via the bottom-up connection required for perception. External feedback is less immediate and results from sensory analysis of the auditory or other perceptual consequences of an action, which likewise returns priming to the nodes that originated the action.

The internal and external feedback (priming) that results from the two-way (bottom-up and top-down) connections between most mental nodes (Chapter 2) leads to the possibility of reverberatory effects at virtually every level in a system, as illustrated by means of the hierarchically connected, but otherwise arbitrary, mental nodes in Figure 8.1, 1 (superordinate) and 2 and 3 (subordinate). During production, 1 becomes activated, and primes 2 via the top-down connection. However, when 2 and 3 subsequently become activated, this could lead to a reactivation of 1 because of the bottom-up connections required for perception involving these same nodes.

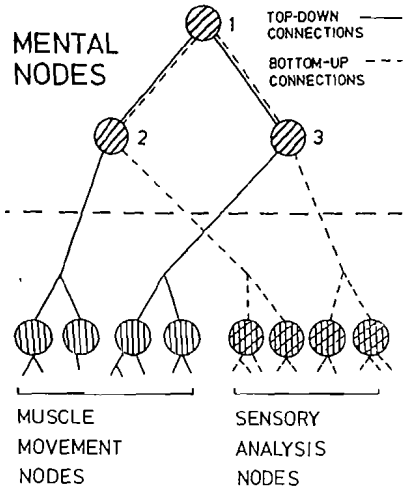

FIGURE 8.1. The relation between mental nodes, sensory analysis nodes, and muscle movement nodes.

Mental nodes must therefore become self-inhibited following activation to ensure that internal feedback (bottom-up priming) resulting from activation of subordinate nodes does not lead to repeated (reverberatory) reactivation of higher level nodes. If the self-inhibitory mechanism of a single mental node were to break down, an output resembling stuttering would result. If the self-inhibitory mechanism for large numbers of nodes were to break down simultaneously, the physiological and behavioral effects would resemble seizure or general convulsion.

Self-inhibition also helps to prevent similar and equally catastrophic effects of external feedback. For example, sensory analysis nodes automatically process the auditory feedback that arises during normal speech production and primes the same low-level mental nodes that generated the output in the first place, and reverberatory reactivation can only be prevented if the mental nodes receiving this returning, feedback-induced priming are still in their self-inhibitory phase following activation.

The Recovery Cycle

Self-inhibition is only the first stage in the cycle of recovery from activation. Two further stages are necessary to complete the cycle: hyperexcitability or post-inhibitory rebound – a period during which self-priming rises above resting level (for the generality of postinhibitory rebound, see Grossberg, 1982) – and the final return to resting level.

In all then, four phases of excitability (summarized in Figure 8.2) follow multiplication of priming: activation, self-inhibition (the relative refractory phase), hyperexcitability, and the return to resting level of priming. The activation phase begins at time t_0 in Figure 8.2 and ends at time t_1, when self-inhibition

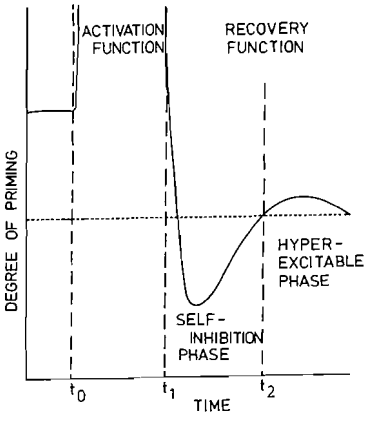

FIGURE 8.2. The activation and recovery phases for a single node. The activation function illustrates multiplication of priming beginning at time t_0, with self-sustained activation continuing until time t_1. The recovery function shows how priming first falls below resting level (self-inhibition) at termination of activation and then rebounds beginning at time t_2 (the hyperexcitability phase).

begins. This introduces the relative refractory phase, which lasts until t_2 in Figure 8.2 when priming first returns to resting level. The hyperexcitability or rebound phase follows, during which priming rebounds before returning to resting level again.

Temporal Characteristics of the Activation and Recovery Cycle

The duration of the entire activation and recovery cycle (from onset of activation until the final return to resting level) varies with the level of a node in the system. For the lowest level kinesthetic-muscle movement nodes, the entire cycle may last only a few milliseconds. However, for phonological nodes, the cycle can last up to 300 ms, with a relative refractory phase lasting as long as 100 ms, a hyperexcitability peak at 200 ms, and a return to resting level at 300 ms. For still higher level nodes, the activation and recovery cycle can take even longer.

Why must self-inhibition last so long in higher level (e.g., phonological) nodes, and why does the duration of self-inhibition vary with the level of a node in the system? One reason is that prolonged self-inhibition enables protection against external feedback, which takes a relatively long time to return in the case of mental nodes. A higher level node only begins to receive external feedback (priming) after three sets of events have taken place: (1) all of its subordinate nodes become activated, including its lowest level muscle movement nodes; (2) the environmental effects of the resulting action become relayed to the sensory receptors, enabling, for example, the airborne auditory feedback from speech to reach the ears; and (3) the sensory analysis and other subordinate nodes process the feedback and deliver priming to the higher level node(s) that originated the output. In order to prevent the reverberatory effects of external feedback, self-inhibition must therefore last as long as steps (1) to (3).

Why must the duration of self-inhibition vary with the level of the node in question? One obvious reason is that the time for steps (1) and (3) depend on the level of the node in an action hierarchy. The higher the node, the longer it will take for

external feedback to return to that node. Another reason concerns internal feedback. A higher level node continues to receive internal feedback over a longer period of time than does a lower level node, so that again, self-inhibition must last longer for higher level nodes in order to prevent catastrophic reactivation.

Neural Mechanisms and the Recovery Cycle

In general, the basic properties of nodes (priming and activation) mimic by design the basic properties of neurons (potentiation and spiking) except for time scale and underlying mechanism, and the same is true of self-inhibition. The self-inhibitory process that follows activation of isolated neurons resembles the self-inhibition of mental nodes, but differs greatly in time scale. For example, the relative refractory period lasts 1 to 2 ms in isolated peripheral neurons, as compared to over 150 ms in self-inhibited mental nodes. Also unlike neural recovery, time characteristics of self-inhibition vary with the level of the node in question. The higher the node in a perceptual hierarchy, the longer the duration of self-inhibition. The mechanism underlying self-inhibition versus neural refractoriness also differs. Neural refractoriness reflects a physiochemical recovery process, whereas self-inhibition of mental nodes reflects complex neuronal interactions and evolved for reasons other than recovery per se.

What are the neural mechanisms underlying self-inhibition? The simplest possibility is that mental nodes consist of two interacting components: a parent neuron and an isolated inhibitory collateral or satellite neuron with an inhibitory connection to its parent. Like other isolated neurons, the inhibitory satellite has an absolute rather than a most-primed-wins threshold. Once its (extremely high) threshold potential has been reached, the satellite becomes activated (generates spikes), and once its potential falls below threshold, it deactivates. Thus, when the parent neuron has generated spikes over some prolonged period of time, the potential of its connected inhibitory satellite summates up to threshold, whereupon the satellite becomes activated and inhibits its parent neuron. This self-inhibition deactivates the parent neuron, thereby reducing the excitatory potential of its inhibitory satellite to below-threshold levels. As a result, the inhibitory satellite itself deactivates, enabling recovery from self-inhibition to begin in the parent neuron.

What determines the temporal parameters of self-inhibition in mental nodes? Although inhibitory satellites are a built-in characteristic of all mental nodes, the history of activation of a node determines when its self-inhibitory phase will begin. The process is described in Chapter 1 (and in Eccles, 1972). That is, the built-in connection between a parent neuron and its inhibitory satellite is extremely weak in the case of uncommitted (never previously activated) nodes, so that the parent neuron must remain activated for a very long time in order to build up sufficient potential to activate its inhibitory satellite, and thereby terminate activation of the parent neuron. However, as repeated activation increases the linkage strength of the connection between parent and satellite, the time required before onset of self-inhibition will automatically decrease. Because

frequency of prior activation is directly related to the level of a node in the hierarchies for perception and action (Chapter 4; D. G. MacKay, 1982), the onset time of self-inhibition (i.e., the duration of activation) will therefore decrease with decreasing level of the node in question (all other factors being equal).

Empirical Evidence for Self-Inhibition and the Recovery Cycle

In what follows I discuss various sources of empirical evidence suggesting that mental nodes undergo a recovery cycle with the general characteristics previously discussed. The evidence ranges from neurophysiology, to errors in speech and typing, to repeated letter misspellings of dysgraphics, to the perception and recall of repeated letter misspellings by normal individuals, and even to the pattern of phoneme repetition in the structure of languages.

The Misspellings of Dysgraphics

A dysgraphic is someone who misspells common words with very high probability, due perhaps to cerebral injury, but not to lack of schooling or to a general inability to learn. The dysgraphic has a chronic spelling problem but sometimes spells one and the same word correctly on one occasion and incorrectly on others. And repeated letters often play a role in the misspellings. Lecours (1966) originally discovered the "repeated letter effect" in the misspellings of a dysgraphic (Lee Harvey Oswald) whose data were reanalyzed by D. G. MacKay (1969c) in a way that bears on the recovery cycle hypothesis. Oswald often dropped a repeated letter in a word, misspelling *elderly* as *eldery*, for example, but he sometimes added repeated letters of his own, as in *habitituated* and *Decemember*. D. G. MacKay (1969c) showed that deletions of repeated letters were significantly more common than additions and that deletions of the second of two repeated letters, as in *eldery* were significantly more common than deletions of the first, as in *ederly*, as would be expected if repeated letter deletions reflected self-inhibition of nodes that are to be activated in sequence.

However, D. G. MacKay's (1969c) most interesting data bearing on the recovery cycle hypothesis concerned the degree of separation of the repeated letters in Oswald's misspellings. Oswald frequently misspelled repeated letters that were close together, as in *anlyze*, but he rarely dropped repeated letters that were widely separated, as are the *i's* in *misspelling*. Figure 8.3 shows the probability that the dysgraphic *correctly* spelled words containing repeated letters as a function of the degree of separation of repeated letters in the sample at large. By way of illustration, the repeated *O's* in *cooperation* fall within the Separation 0 (zero) category because no letters separate the repeated *O's*. The repeated *A's* in *analyze* fall within the Separation 1 category because one letter separates the repeated *A's*, and so on, up to Separation 7.

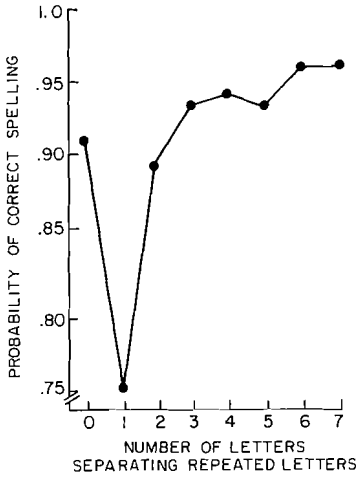

FIGURE 8.3. The probability of correct spelling (versus letter deletion) as a function of the number of letters separating a repeated letter in the diary of Lee Harvey Oswald. This probability (PC) was calculated for each degree of separation as PC = Fd/F, where Fd is the frequency of repeated letter deletions such as *eldery* and F is the frequency of correct spelling. (From "The Repeated Letter Effect in the Misspellings of Dysgraphics and Normals" by D. G. MacKay, 1969c, *Perception and Psychophysics*, *5*, pp. 103–104. Copyright 1969 by Psychonomics Journals Inc., Austin, Texas. Reprinted by permission.)

As can be seen in Figure 8.3, the probability of correctly spelling a repeated letter sequence was moderately high for immediately repeated letters (Separation 0), but dropped dramatically for repeated letters with Separation 1, and increased again to its highest level at Separation 6. The function in Figure 8.3 can be seen to resemble the recovery cycle function in Figure 8.2, except for the moderately high probability of correctly spelling immediately repeated letters. However, this special status of immediately repeated letters is to be expected. Unlike other repeated letters, immediately repeated letters generally are not pronounced in English, and do not represent separate phonemes. There are also special orthographic rules that apply only to immediately repeated letters and not to repeated letters with greater separation. Both of these factors suggest that data for repetitions with zero separation should be disregarded or treated separately in this and related functions.

The Perception and Recall of Misspellings

Another source of evidence for the recovery cycle hypothesis comes from a study of the detection of experimentally constructed misspellings. D. G. MacKay (1969c) examined how readily normal subjects could perceive and recall misspellings that resembled those produced by the dysgraphic discussed previously. The misspellings were planted in sentences that subjects read at a rate of about 77 ms per letter. The subjects then attempted to recall the sentence, writing it out exactly as it was spelled, guessing at misspellings if necessary. Following recall, each sentence was presented again, and subjects responded yes or no to each spelling error in turn, depending on whether they had noticed the error when reading the sentence.

The results showed that normal individuals experienced greatest difficulty in perceiving and recalling those experimentally constructed misspellings that resembled the ones produced most frequently by the dysgraphic. Repeated-letter misspellings such as *eldery* were harder to perceive than nonrepeated-letter misspellings such as *eldely*. Similarly, given that the subject claimed to perceive an error when reading the sentence, misspellings that involved repeated letters were more difficult to recall than those that did not.

As expected under the recovery cycle hypothesis, the functions relating degree of separation of the repeated letters to both perception and recall resembled the dysgraphic function in Figure 8.3. The perception function appears in Figure 8.4. The solid line in Figure 8.4 plots the probability of correctly perceiving nonrepeated-letter misspellings, while the broken line plots the probability of correctly perceiving repeated-letter misspellings, as a function of the degree of separation of the repeated letters. As can be seen in Figure 8.4, repeated-letter misspellings were only harder to detect than nonrepeated-letter misspellings when zero to three letters separated the repeated letters. With more than three intervening letters, repeated-letter errors were as easy or easier to detect than nonrepeated-letter errors.

The recall function appears in Figure 8.5, which plots the probability of correctly recalling a misspelling that was correctly perceived. The solid line in Figure 8.5 represents recall for nonrepeated-letter misspellings, while the broken line represents recall for repeated-letter misspellings, as a function of the degree of separation of the repeated letters. As can be seen in Figure 8.5, repeated-letter misspellings were only harder to recall than nonrepeated-letter misspellings when zero to two letters separated the repeated letters. With more than three intervening letters, repeated-letter errors were as easy or easier to recall than nonrepeated-letter errors.

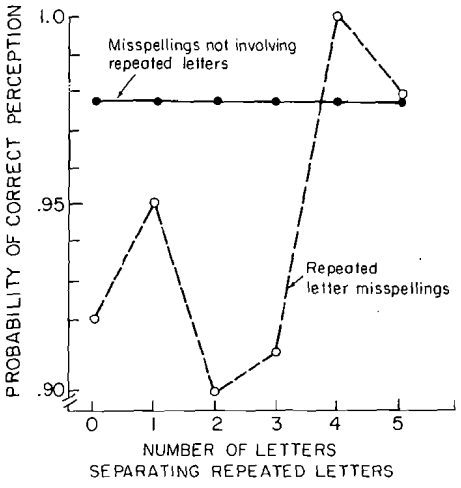

FIGURE 8.4. The probability of perceiving repeated-letter misspellings such as *eldery* (broken line) and nonrepeated-letter misspellings (solid line) such as *eldely*. (From "The Repeated Letter Effect in the Misspellings of Dysgraphics and Normals" by D. G. MacKay, 1969, *Perception and Psychophysics*, *5*, pp. 103–104. Copyright 1969 by Psychonomics Journals Inc., Austin, Texas. Reprinted by permission.)

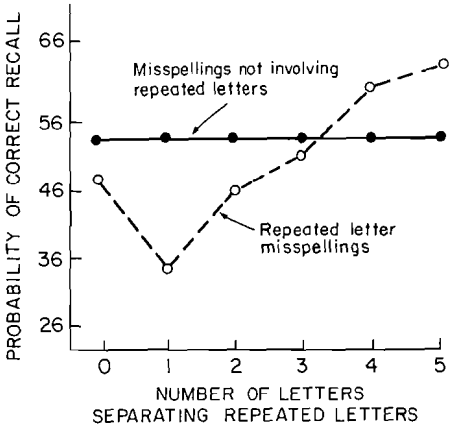

FIGURE 8.5. The probability of recalling repeated-letter misspellings (broken line) and nonrepeated-letter misspellings (solid line). (From "The Repeated Letter Effect in the Misspellings of Dysgraphics and Normals" by D. G. MacKay, 1969, *Perception and Psychophysics*, *5*, pp. 103–104. Copyright 1969 by Psychonomics Journals Inc., Austin, Texas. Reprinted by permission.)

An additional prediction from the recovery cycle hypothesis concerned the nature of repeated letter misspellings. If a repeated letter has been *added*, as in *elderdly*, then the added letter should be easier to perceive with high degrees of separation (4 to 5), than with low degrees of separation (0 to 2). However, if a repeated letter has been *deleted* as in *eldery*, then the error should be as easy to perceive with low degrees of separation (0 to 2) as with high (4 to 5), because of the hypernormal excitability of the node for the expected letter at the later point in time.

LIMITATIONS OF D. G. MACKAY (1969c)

Although D. G. MacKay's (1969c) results support the recovery cycle hypothesis, there are a number of ways to extend and refine these observations using computer-controlled stimulus presentation. Consider, for example, a task where subjects must detect a set of possible target letters in a rapidly presented sequence of letter displays, as in Shiffrin and Schneider (1977), except that target letters can be repeated in the sequence. The subject has three tasks. The top priority task is to respond as quickly as possible to the occurrence of any member of the set of targets, and the other, subsidiary tasks are to indicate (1) whether a second target followed the first and (2) the identity of the target(s). In the special case where the *same* target letter is repeated (in either identical or different locations in the display), the node structure theory predicts that detecting the first target presentation will (1) interfere with detection of the second when the intervening interval is short (up to 50 ms, say) but will (2) facilitate detection of the second when the intervening interval approximates some critical value corresponding to the hyperexcitability peak. Specifying this critical value in advance is difficult, because the recovery cycle varies with the level of a unit and its prior history of activation, but a value of 200 ms can be expected if letter-detection nodes resemble those for phonemes.

Repeated Phonemes in Fast Speech

When speaking rapidly, speakers tend to drop immediately repeated phonemes, as might be expected under the recovery cycle hypothesis. However, determining exactly which nodes are responsible for this phenomenon is difficult. For example, when speaking rapidly, speakers often drop one of the immediately repeated *k's* in *take care* (Heffner, 1964), but it is difficult to specify which *k* was dropped, or even to determine whether the omission was intentional. Moreover, such omissions cannot be attributed to self-inhibition at the segment level, because two different segment nodes control production of these *k's*, *k*(final consonant group) and *k*(initial consonant group), and because self-inhibition of nodes below the segment level could cause these omissions.

Errors in Speech and Typing

The fact that anticipations are much more common than perseverations in everyday speech errors (Cohen, 1967) bears directly on the self-inhibition hypothesis. So does Type I motor masking. Unlike the omissions that occur in rapid speech, the class of speech errors that D. G. MacKay (1969b) labeled Type I motor masking is undeniably inadvertent. The speaker unintentionally drops the second of two segments due to be repeated in close succession. An example from Meringer and Mayer's (1895) corpus of German errors is *der iese* instead of *der Riese*, where the immediately repeated /r/ has been dropped. Because Type I motor masking appears to be rare and requires that an observer make subtle perceptual discriminations under less than optimal observational conditions (Cutler, 1982), these errors only weakly support the self-inhibition hypothesis (D. G. MacKay, 1969b), and warrant further investigation using experimental error-induction procedures.

However, omission errors in skilled transcription typing exhibit a similar phenomenon that cannot be attributed to observer error. High-speed videotapes of the finger movements of skilled typists indicate that typestrokes exhibit three phases (Grudin, 1983): (1) a movement that positions the finger over the key, (2) a rapid downstroke to strike the key, (3) and an equally rapid rebound or lift-off from the key. During an omission error, the finger either fails to move toward the key, or fails to execute the downstroke. Like Type I masking errors in speech, typestroke omissions are usually preceded by an identical letter or identical finger movement, either in the same word or in the immediately preceding word. For example, Grudin (1983) had skilled typists transcribe a text as rapidly as possible and found that half of the typists had omitted the third *i* in the word *artificial*. Moreover, in Grudin's overall sample of omission errors, this left-to-right masking (omission) effect was strong enough to override a general tendency for word-initial letters to be correctly typed. Examples of such word-initial omissions are *the ntire* for *the entire*, and *keep utting* for *keep putting*. These typestroke repetition errors suggest that following activation, the mental nodes

controlling skilled typestrokes undergo a period of self-inhibition that interferes with the subsequent execution of an identical typestroke due to follow closely in time (see also Grudin, 1983). Omissions of immediately repeated movements required for different letters suggest a similar process for nodes that perceive and/or produce movement components, rather than whole letters.

On the surface, omission errors in typing, and the recovery cycle hypothesis in general, seem to conflict with the fact that people can repeat a finger movement more rapidly with a single finger than with two adjacent fingers in alternation. Why is the maximal rate of key pressing faster with one finger than with two fingers? One reason is that repeated finger movements can be executed entirely within the muscle movement system. Mental nodes are unnecessary for producing an up-and-down movement of the finger, so that the temporal parameters of self-inhibition in mental nodes are irrelevant to these movements. Another reason is that two-finger movements require a sequencing process that is not required for single-finger movements. The additional time required for sequencing two-finger movements (Chapter 3) could greatly reduce the maximum response rate.

Electrophysiological Evidence

Although the figure of 200 ms for the hyperexcitability peak of mental nodes is only approximate, it receives support from the electrophysiological literature. For example, postinhibitory rebounds have been observed that follow activation by as much as 200 ms in central neuronal aggregates (see Chang, 1959; Grossberg, 1982; Martin, 1985). These central neuronal aggregates therefore differ from isolated peripheral neurons, where the hyperexcitability peak arrives orders of magnitude sooner.

Electromyographic Evidence

Electromyographic potentials without full-blown muscle movements occur during mental practice and indicate in the theory (D. G. MacKay, 1981) that the relevant muscle movement nodes are being primed, and in some cases inadvertently activated. Because priming is assumed to peak spontaneously at about 200 ms following self-inhibition, the theory therefore predicts that the appropriate muscles will exhibit a peak in electromyographic potentials about 200 ms after production of a segment during speech production. Evidence for such a peak is found in a study by Ohala and Hirano (1967). They had subjects produce the syllable *pa* while recording electromyographic potentials in the obicularis oris muscles of the lips, and they observed two peaks of electromyographic activity. The first was causally related to contraction of the lip muscles, and the other, smaller peak about 200 ms later, suggests support for the recovery cycle hypothesis.

Phoneme Repetition in the Structure of Languages

Somewhat indirect but nevertheless interesting support for the recovery cycle hypothesis comes from a study of phoneme repetition in the structure of words (D. G. MacKay, 1970c). The rationale for the study was as follows. If a recovery cycle determines how easy it is to repeat an element in a word, then the structure of repeated elements in words should reflect that recovery cycle. Immediate repetition of an element should be rare, reflecting the self-inhibitory phase of the recovery cycle, but at some point following self-inhibition, repetition should become more likely than would be expected by chance, reflecting the hyperexcitability phase of the recovery cycle. Applying this hypothesis to phonemes, if a recovery cycle constraint makes immediate repetition of phonemes difficult, then the phonological structure of words in many languages should reflect this constraint, so that immediate repetitions of phonemes should be rare. Similarly, if phoneme repetition after some interval is extremely easy, then repetition of phonemes in the structure of words should be highly likely after a corresponding degree of separation. Implicit in this rationale is the (uncontroversial) assumption that speakers don't normally alter their rate of speech during ongoing production of words containing repeated phonemes.

To test this recovery-cycle prediction, D. G. MacKay (1970c) examined patterns of phoneme repetition in words of various lengths in two very different languages, Serbo-Croatian and Hawaiian. A very large and random sample of words was examined, and the probability of segment repetition at the various separations was corrected on the basis of word length, because long separations for short words are impossible.

Figure 8.6 shows the pattern of phoneme repetition for both languages, averaged over consonants and vowels. As can be seen, immediate repetition of phonemes was highly unlikely for both languages and significantly less likely than would be expected if phoneme repetition in a word were a random event (represented by the dashed lines in Figure 8.6). This finding is interesting because immediate repetition of phonemes is clearly physiologically possible. Sequences such as /b-b-b/ are a common occurrence in stuttering (Chapter 10).

With wider degrees of separation, probability of repetition rose sharply, peaking with three intervening elements in both Serbo-Croatian and Hawaiian (Figure 8.6). These peak repetition probabilities exceeded what would be expected if phoneme repetition in a word in either language were a random event. Following the peak, probability of repetition declined to approximately chance level. However, there was a difference between vowels and consonants (D. G. MacKay, 1970c), reflected in part by the two initial peaks in the functions of Figure 8.6. In both Serbo-Croatian and Hawaiian, probability of *vowel* repetition peaked with one intervening element, whereas probability of *consonant* repetition peaked with three intervening segments.

Differences between the two languages were superimposed on the similarities. As can be seen in Figure 8.6, the repetition pattern was more erratic for Hawaiian than for Serbo-Croatian. The reason is that virtually all of the Hawaiian syllables

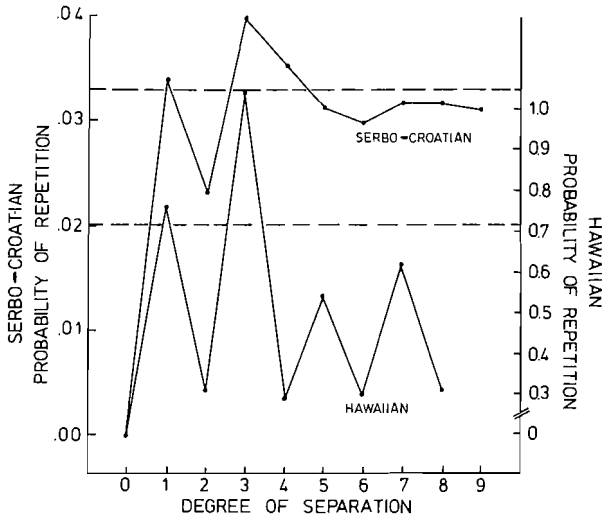

FIGURE 8.6. The probability of phoneme repetition in Serbo-Croation (left ordinate) and Hawaiian (right ordinate) as a function of degree of separation of the repeated phonemes. The horizontal lines represent probabilities under the hypothesis that phoneme repetition is a random event.

had a simple consonant-vowel structure, whereas Serbo-Croatian syllables were much more variable and complex in structure. Thus, because vowels and consonants alternate in Hawaiian, segment repetition at even numbered degrees of separation was of necessity infrequent. Given some way of factoring out syllable structure in computing the probabilities of repetition, the Hawaiian and Serbo-Croation functions would have resembled each other more closely.

As D. G. MacKay (1970c) points out, the peaks in Figure 8.6 suggest a "law of latent alliteration," and differences between Hawaiian versus Serbo-Croatian syllable structures serve to illustrate the statistical nature of that law. Many other factors play a role in determining the phonological structure of words and can thereby override the law of latent alliteration in any one instance. Another important factor, especially for complex or derived words, is the nature of affixes and stems that combined to make up the word. For example, adding a different prefix could change the structure of repeated phonemes in the word. It is therefore remarkable that a consistent repetition structure should appear at all and that it should be so similar for Hawaiian and Serbo-Croatian, languages that differ in type of segments, number of segments (13 versus 35), and average number of segments per syllable and per word. The data therefore suggest, but do not prove, that the pattern of segment repetition in Figure 8.6 is representative of all languages, as would be expected if general properties of nervous action were an underlying cause. In theory of course, a similar pattern should appear with only minor variations in all languages.

Phoneme Repetition and Language Evolution

Phoneme repetition is a factor in the phonological changes that words undergo over time, because repeated segments often become dropped as a language evolves. An example is the Latin word *stipipendium*, which changed to *stipendium*, dropping the repeated /p/ and repeated /i/ (Meringer & Mayer, 1895). Such changes fit the prediction that phoneme repetition represents a factor in the evolutionary processes that words undergo in the history of a language, but the extent to which the hypothesized recovery cycle plays a role in the initial invention of words, or in the evolutionary changes that occur over time, is currently unknown. A great deal of further research is required to fully specify how the pattern of segment repetition within words evolves.

Limits of the Recovery Cycle Hypothesis

I come now to some limits of the recovery cycle hypothesis for theories of speech and other behaviors. One limit is that the constraints on segment repetition discussed previously only apply to nodes that receive parallel top-down and bottom-up connections and therefore require protection from internal feedback. For example, the recovery cycle hypothesis does not apply to the lowest level mental nodes in an action hierarchy, which in the case of speech, are distinctive feature nodes. In theory, distinctive feature nodes are undisturbed by internal feedback, because they do not receive parallel top-down and bottom-up connections. It is therefore interesting that immediate repetition of distinctive features is common in many languages, and the period during which distinctive feature nodes remain activated generally exceeds that of segment nodes, spanning several adjacent segments, which could not occur if distinctive feature nodes undergo self-inhibition.

Another limit to the recovery cycle hypothesis concerns the potential or maximal rate of segment repetition. At best, data on the structure of segment repetition in a language only suggest a repetition pattern that may be *easily* produced at average rates for everyday speech. The data say nothing about the potential or maximal rate for producing syllables or for moving the speech muscles. Self-inhibition introduces a relative rather than an absolute refractory phase. By exerting greater effort, speakers can greatly exceed the natural repetition rates suggested in Figure 8.6.

Yet another limit to the recovery cycle hypothesis concerns the type of skill being investigated. Similar constraints on repetition rates can only be expected for skills that employ shared input–output nodes. Skills that do not engage mental nodes for production and perception will not display the same recovery cycle pattern.

Finally, the self-inhibition hypothesis applies only to highly practiced and rapidly activated mental nodes and not to seemingly similar phenomena that occur in unskilled behavior, such as psychological refractoriness and the Ranschburg (1902) effect, discussed in the following section.

Self-Inhibition and the Psychological Refractory Effect

The psychological refractory effect refers to the fact that responding to one signal can prolong the time required for responding to a subsequent signal. The basic empirical paradigm is as follows. A signal such as a tone is presented, to which the subject must rapidly respond with, say, a key press. Then, before the key press has been made, a second stimulus such as a light appears, to which the subject must make a different response, say, a vocal "yes." The remarkable and frequently replicated result is that responding to the first signal lengthens reaction time to the second. In the present example, the tone lengthens reaction time to the light, relative to when the light alone is presented. The difference in reaction time between these two conditions is known as the refractory delay and typically exceeds 200 ms. Refractory delays diminish over time, but often exceed 100 ms even when the response to the first signal has been completed before onset of the second signal.

On the surface, the psychological refractory effect seems to resemble effects of self-inhibition discussed above. Both effects have a central origin, are very general in nature, and reflect an inhibitory process. As Keele (1973) points out, psychological refractory effects occur with many input–output skills and reflect central rather than purely sensory or purely motor limitations. Refractory delays remain when different perceptual systems process the signals and when different output systems execute the responses, as in the example previously cited.

However, only some of the refractory delays in the literature can be attributed to self-inhibition of content nodes. When both stimuli and both responses differ in the psychological refractory paradigm (again as in the above example), different content nodes are becoming activated and self-inhibited, so that self-inhibition cannot explain the interference. However, the interfering stimuli and responses in the psychological refractory paradigm are almost invariably similar, so that their content nodes probably belong to the same domain. This suggests that psychological refractory effects may hinge on the reactivation of an activating mechanism or sequence node which has recently been quenched (Chapter 1). Under this hypothesis, refractory delays should disappear following extensive practice with independent input–output mappings. The practice will enable both mappings to engage nodes in different domains, activated via different sequence nodes (D. G. MacKay, 1987).

This view explains two otherwise puzzling sets of conflicting results in the psychological refractory literature. One concerns the effects of practice on psychological refractoriness. Some studies have shown that refractoriness disappears with practice (e.g., Greenwald & Schulman, 1973), while others have shown that it does not, even after 87 practice sessions (e.g., Gottsdanker & Stelmach, 1971). Under the theory, these conflicting outcomes reflect differences in compatibility of the input–output mappings that the tasks engage. Greenwald and Schulman's (1973) tasks involved highly compatible or practiced input–output mappings, specifically, moving a lever in the direction of an arrow and shadowing (naming auditorily presented letters). These highly practiced

mappings eliminated the psychological refractory effect. Responding to one stimulus had no effect on the rate of responding to the other stimulus. Under the node structure theory the reason for this absence of effect is that these different and highly practiced stimulus–response mappings engaged independent activating mechanisms, eliminating the possibility of cross-talk effects.

In Gottsdanker and Stelmach's (1971) tasks on the other hand, the input–output mappings were nonindependent, so that practice could not eliminate the possibility of cross-talk effects. That is, both stimuli were visual, and both responses were manual and engaged nodes in the same domains (perhaps even some of the same nodes in the same domains). As a result, identical activating mechanisms and content nodes were required for both responses, causing delays attributable to quenching, self-inhibition, and speed–accuracy trade-off.

The node structure account of psychological refractory effects also explains another set of conflicting results. When subjects are instructed to simply observe rather than respond to the first stimulus, some studies (e.g., Rubinstein, 1964) have reported full-blown refractory delays, whereas other studies (e.g., Borger, 1963) have reported no delays whatsoever. The difference is attributable to the fact that observation alone only causes refractory delays under the node structure theory when the principle of higher level activation must be abandoned and lower level nodes must become activated, that is, when the two signals are unfamiliar or highly similar and difficult to distinguish (D. G. MacKay, 1987). For example, Rubinstein (1964) observed full-blown refractory effects—even though the first signal required no response whatsoever—when the two stimuli were highly similar visual forms presented to the two eyes or were highly similar noises presented to the two ears. However, Rubinstein (1964) also showed that these refractory delays disappeared when the noise was presented to one ear as the first stimulus and the visual form was presented to the opposite eye as the second stimulus, or vice versa. In this condition, the stimuli occupy different domains, and the first stimulus neither activates nor primes nodes in the same domains as are required for producing the second response. As a result, the first stimulus interferes with neither perception of the second stimulus, nor production of the second response.

Finally, it should be noted that some refractory effects call for an explanation in terms of encoding processes, rather than either quenching or self-inhibition. For example, it is often the *first* response that is slowed down when the second signal arrives soon after the first. Moreover, with very short interstimulus intervals and prior experience with the stimuli, their order, and their interstimulus interval, the two responses sometimes become produced in quick succession and without refractory delays (Welford, 1968). As others have noted, the subjects can encode the stimuli as a pair and produce the responses as a group under these conditions, thereby avoiding refractory effects.

Self-Inhibition and the Ranschburg Phenomenon

The Ranschburg phenomenon refers to an inhibitory effect of repeated items on the immediate recall of short sequences of items. Sequences containing repeated

items are recalled more poorly than sequences containing nonrepeated items. In his original (1902) demonstration, Ranschburg presented a simultaneous string of six digits for about 333 ms, and found that a string containing repeated digits was more poorly perceived or recalled than a string containing nonrepeated digits. Moreover, repeated digits, and especially the second of two repeated digits, were the ones most poorly perceived or recalled. Ranschburg attributed these findings to a general inhibitory process for repeated items and predicted analogous effects in spelling and writing.

The Ranschburg effect has stimulated a large number of experiments over the last 85 years, and these more recent experiments suggest that poor performance on repeated items is attributable to a complex conspiracy of factors, rather than just inhibition per se (Jahnke, 1969). For example, encoding factors play a big role in the Ranschburg paradigm. Subjects sometimes recall that an item was repeated but can't recall which one and inadvertently repeat the wrong one. Even when they accurately recall which item was repeated, subjects often place the repeated items in the wrong positions in the overall sequence. Finally, *immediately* repeated elements generally are perceived as a group and are recalled *better* than nonrepeated items. All of these findings are attributable to encoding effects (the process of forming connections between nodes in the node structure theory) and are irrelevant to the activation and self-inhibition of already formed nodes and connections.

Moreover, some of the encoding processes in the Ranschburg (1902) paradigm are rather artificial and fundamentally unlike encoding processes for, say, natural speech production (Jahnke, 1969). For example, effects of spacing on recall of repeated elements in the Ranschburg paradigm sometimes resemble recovery cycle effects in speech, but only superficially. *Facilitatory* rather than inhibitory effects are observed in widely separated elements but these effects vary with position in the list, spacing per se, and ease of encoding (e.g., repeated items at the beginning and end of a list are especially easy to encode and recall; Jahnke, 1969). The Ranschburg paradigm also allows guessing strategies, which can account for the poorer recall of repeated items, because guesses have a lower probability of being correct when items from a limited set (such as the digits 0 to 9) are repeated than when they are not repeated (Hinrichs, Mewalt, & Redding, 1973). Together these factors make the Ranschburg paradigm a poor choice for testing the self-inhibition hypothesis.

Self-Inhibition and Pathological Stuttering

Having reviewed the evidence relating to self-inhibition and the recovery cycle, I now apply the recovery cycle hypothesis to the phenomenon of pathological stuttering. I begin with an overview of similarities and differences between different types of pathological stuttering, and then compare stuttering with other types of speech errors.

Types of Pathological Stuttering

Stuttering is a complex disorder that falls into two general categories: feedback-induced stuttering and intrinsic stuttering. The latter is not attributable to a disruptive effect of feedback but to the motor control process per se. I discuss the evidence for distinguishing between these two categories of stuttering in Chapter 10. Interestingly, feedback-induced and intrinsic stuttering have been responsible for a major division within the field itself. Two major approaches to the study of stuttering have gone their separate ways over the past several decades, one viewing stuttering as a disorder in motor control, and the other as a disorder in the processing of auditory feedback (Garber & Siegel, 1982).

I begin with a brief description of the general characteristics shared by both feedback-induced and intrinsic stuttering and then concentrate on intrinsic stuttering in the remainder of the chapter, leaving feedback-induced stuttering for Chapter 10.

General Characteristics of Pathological Stuttering

The motor control problem in pathological stuttering can be separated into an unlearned component and a learned component. All stutterers exhibit the unlearned component, the speech errors known as repetitions, prolongations, and blocks. In repetition errors usually only single consonants or consonant clusters are repeated, and only occasionally a syllable or monosyllabic word (Van Riper, 1982). Prolongations involve the unbroken lengthening of a (continuant) phoneme. For example, a prolongation within the word *practice* might lengthen the /r/ to more than 40 times its normal duration. Descriptively, it is as if the articulators become "locked" in position, resulting in prolonged production of a continuant sound. Blocks reflect an inability to utter any sound at all and occur most often at the beginning of words and utterances. As Van Riper (1982) pointed out, blocks can be considered a special type of prolongation where one or more articulators (e.g., the velum, lips, or glottis) become locked in an obstructive position, preventing speech by virtually eliminating airflow.

The other, learned component of stuttering reflects the stutterer's ability to recognize and anticipate the occurrence of repetitions, prolongations, or blocks and attempt to avoid them or shorten their duration by, for example, contortion of facial muscles, changes in breathing pattern, or altered choice of words. These behaviors are characterized as learned because they don't appear in children who stutter or in normal adults who occasionally repeat, prolong, or block on speech sounds (Meringer & Mayer, 1895). Children and normal adults haven't developed anticipatory fears of these errors and therefore don't exhibit the learned attempts to avoid them.

Stuttering and Other Speech Errors

The present book focuses on the unlearned component of stuttering: repetitions, prolongations, and blocks. Stutterers sometimes make other errors, and

nonstutterers sometimes stutter (Meringer & Mayer, 1895). The difference is that stutterers make repetitions, prolongations, and blocks orders of magnitude more frequently than do nonstutterers but make other errors with normal frequency (D. G. MacKay & Soderberg, 1972). In what follows, I compare stuttering with other speech errors in greater detail, my ultimate goal being to explain stuttering within the same framework as other speech errors.

Similarities Between Stuttering and Other Speech Errors

Stuttering resembles other speech errors in several respects. Variables such as ambiguity influence stuttering and other speech errors in similar ways (See D. G. MacKay, 1969a). Also, as in other errors, stuttering decreases with practice in producing a sentence (the so-called adaptation effect, Brenner, Perkins, & Soderberg, 1972) and with reduction in rate of speech (Perkins, Bell, Johnson, & Stocks, 1979). These parallel effects of practice and speech rate on stuttering and other speech errors are readily explained within the node structure theory (D. G. MacKay, & MacDonald, 1984).

Stuttering and other speech errors also occupy similar phonological loci. Like other errors such as anticipations, stuttering often occurs at the beginning of words and sentences (a phenomenon attributable to anticipatory priming) and involves stressed syllables more often than unstressed syllables (D. G. MacKay, 1970a). Finally, stuttering is not limited to people classified as stutterers. Meringer and Mayer (1895) showed that "nonstutterers" also produce repetitions, blocks, and prolongations, although subsequent collections of speech errors (e.g., Fromkin, 1973) have by and large excluded these everyday errors. The Epilogue following Chapter 10 examines the scope, rationale, and effects of this exclusionary strategy.

Differences Between Stuttering and Other Speech Errors

One of the main differences between stuttering and other speech errors, its individual-specific and error-specific nature, has already been noted. The same individuals produce repetitions, prolongations, and blocks with much higher frequency than normal speakers, but make other types of errors such as spoonerisms with about the same frequency as normal speakers. Another difference is that adult stutterers can anticipate when they are going to stutter, whereas nobody anticipates or struggles to avoid making other errors.

Stuttering also differs from other errors in disruptiveness or severity. Sometimes stuttering disrupts communication for sustained periods of time and cannot be voluntarily "corrected," unlike other errors, which either pass unnoticed, or are quickly and easily repaired. Also unlike other speech errors, stuttering is not always limited to the speech-motor system but manifests itself in other activities of stutterers. For example, stutterers have more difficulty than nonstutterers in reproducing the timing of a sequence of syllables or finger taps (M. H. Cooper & Allen, 1977), as if both speech and finger movement shared the same timing deficit. Stuttering also relates to the processing of sensory feedback in a way that

other errors do not (see Chapter 10). Finally, as discussed below, stuttering and other errors seem to originate at different levels of speech production.

The Level at Which Stuttering Originates

Although the level at which stuttering originates has been a source of controversy for the past half century (see D. G. MacKay & MacDonald, 1984), current evidence suggests that stuttering can be localized within the muscle movement system. I review some of these lines of evidence in the following discussion.

Fluency During Internal Speech

Stutterers do not report stuttering during internal speech, a phenomenon that is sometimes attributed to the possibility that speaking to one's self may provoke less anxiety and therefore less stuttering than speaking to others. However, reduced anxiety cannot fully account for the absence of stuttering during internal speech, because stutterers frequently report stuttering when speaking *aloud* to themselves (Van Riper, 1982).

By way of contrast, other speech errors occur as often during internal speech as during overt speech. Dell (1980) provided the crucial evidence. His (normal) subjects repeated tongue twisters (e.g., *Unique New York*) either aloud or to themselves at a fixed but rapid rate and reported the occurrence of errors. The results showed that other errors (e.g., transpositions, perseverations, and anticipations as in misproduction of *New York* as *You Nork*, *New Nork*, or *You York*) occurred with exactly the same frequency during both internal and overt speech. Under the node structure theory, this finding suggests that these other errors can be localized within the phonological system, which is shared by internal and overt speech, rather than within the muscle movement system, which is not. And, because stuttering occurs during overt, but not during internal speech, stuttering can be localized within the muscle movement system, the only additional system that becomes engaged during overt speech.

The Number of Muscle Movement Components

If stuttering represents a muscle movement disorder, then decreasing the number of muscle movement nodes that become activated during speech should decrease the probability of stuttering. A study by Brenner et al. (1972) strongly supported this prediction. They had stutterers produce sentences in three different ways: mouthing (moving only the lips), whispering, and full-fledged articulation. The results showed that severity of stuttering varied with the number of muscles involved. The stutterers stuttered least when moving only their lips, more when whispering, and most when engaging in full-fledged articulation.

Muscle Movement Problems

If stuttering reflects a muscle movement defect, then stutterers may exhibit this defect independently of the phonological processes that normally control

movements of the speech musculature. This is exactly what studies by McFarlane and Prins (1978) and Cross and Luper (1979) showed. They compared reaction times of stutterers and nonstutterers using speech muscles not to make speech sounds but simply to react as quickly as possible, for example, closing the lips in response to a pure tone stimulus. Stutterers were slower at initiating these movements than were nonstutterers, a finding that cannot be explained at the phonological level. A follow-up study by Reich, Till, and Goldsmith (1981) showed that these slow reaction times are confined to the muscles normally used for speech. Stutterers and nonstutterers exhibited equivalent reaction times when using nonspeech muscles, for example, when pressing a key in response to a tone.

Absence of Corresponding Perceptual Deficit

Stutterers show no corresponding deficits in the perception of speech sounds. They never misperceive someone to say "p-p-p-please," when in fact the person said "please." This lack of perceptual deficit fits the hypothesis that stuttering originates in muscle movement nodes (which are specific to production), rather than in sensory analysis nodes (which are specific to perception), or in phonological nodes (which play a role in both perception and production).

The Time Characteristics and Automaticity of Stuttering

Two salient characteristics distinguish muscle movement nodes from higher level nodes: automaticity and rapid activation. As D. G. MacKay (1982) points out, muscle movement nodes receive much more practice than higher level nodes and are much more likely to achieve automaticity, including independence from conscious control. Muscle movement nodes are also activated much more rapidly than higher level nodes. The lowest level muscle movement nodes become activated within a time frame of milliseconds, segments nodes within a time frame of tens of milliseconds, and lexical nodes within a time frame of hundreds of milliseconds.

Both of these characteristics (automaticity and rapid activation) are also characteristics of stuttering, as would be expected under the hypothesis that stuttering represents a muscle movement disorder. Stuttering is automatic and beyond conscious control. Stutterers are unable to alter the way they execute muscle movements so as to avoid stuttering, just as normal speakers are unable to select or alter their muscle movements so as to produce a contextually inappropriate allophone. Stuttering misarticulations also occupy a time frame characteristic of muscle movement nodes rather than phonological nodes. Like muscle movements, stuttering repetitions occur within milliseconds. Durations of 20 ms or less are not atypical for repeated lip movements in stuttering. And stutterers sometimes exhibit movement asynchronies (e.g., between lip and jaw movements) that are so brief as to be inaudible to the unaided ear (Zimmermann, 1980).

The Node Structure Theory of Intrinsic Stuttering

Can the node structure theory provide an integrated account of stuttering and other speech errors? I begin with a review of the basis for other errors in the node structure theory and then examine whether repetitions, prolongations, and blocks can be explained within the same framework.

Error-free output occurs under the theory when an "intended-to-be-activated" content node accumulates greater priming than any other node in its domain and becomes activated. The intended-to-be-activated node is the one that is receiving priming from a superordinate node in the output sequence, that is, the directly connected content node immediately higher in the hierarchy. This priming summates over time and eventually exceeds the priming of all other nodes in the domain, say, by time t_2 in Figure 8.7. At or after this point in time, the triggering mechanism will activate the intended-to-be-activated node under the most-primed-wins principle, and the output is error free.

Consider now substitution errors, where one component substitutes another, as in the anticipation error *coat cutting* instead of *throat cutting*. Substitution errors occur whenever an intended-to-be-activated node acquires less priming than some other node in its domain when the triggering mechanism is applied. Thus, the fundamental cause of substitution errors is that other, "extraneous" nodes in a domain are also receiving priming, and this extraneous priming or noise sometimes exceeds the systematically increasing priming for the intended-to-be-activated node at the time when the triggering mechanism is applied. As a

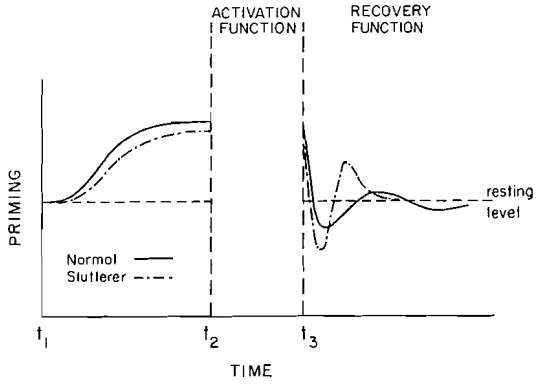

FIGURE 8.7. The priming function (relating degree of priming and time following onset of priming, t_1), activation function (following application of the triggering mechanism at t_2), and recovery function following onset of self-inhibition at t_3) of stutterers and nonstutterers. See text for explanation. (From "Stuttering as a Sequencing and Timing Disorder" by D. G. MacKay & M. MacDonald in *Nature and Treatment of Stuttering: New Directions* (p. 273) edited by W. H. Perkins & R. Curlee, 1984, San Diego: College-Hill. Copyright 1984 by College-Hill Press. Reprinted by permission.)

consequence, the wrong node becomes activated under the most-primed-wins principle, and an error occurs. Because of the way priming summates for every node in a hierarchy (Figure 8.7), substitution errors will be more likely at faster rates of speech. At faster rates the triggering mechanism is applied sooner following onset of priming (t_1), allowing less buildup of priming (D. G. MacKay, 1982).

Now consider repetitions, prolongations, and blocks within this framework for explaining substitution errors. During repetitions, a just-activated component becomes reactivated, and the issue is how the same sequence and content nodes can acquire enough priming following activation to become most primed in their domain and thereby reactivated. One possibility is that muscle movement nodes of stutterers may exhibit an abnormal priming and recovery cycle (illustrated in Figure 8.7), such that priming builds up abnormally slowly and rebounds abnormaly sharply following self-inhibition (as would occur if inhibitory satellites and parent neurons alike had abnormally high thresholds or insensitivity to cross-connection priming). As a consequence, just-activated sequence nodes have a high probability of again becoming most primed in their domain and reactivated with the next pulse from the timing node. Because a just-activated content node also rebounds sharply following self-inhibition, a repetition such a *p-practice* will result.

This reactivation cycle can of course reiterate, so that the same movement is repeated three or four times, but the reactivation cycle cannot go on indefinitely. Because satiation increases with repeated activation, a repeatedly reactivated node cannot rise to such high levels of priming on rebound from inhibition, and so eventually fails to achieve more priming than the next-to-be-activated node. The fact that priming for the next-to-be-activated nodes continues to summate during stuttering also mitigates against long strings of repetitions. Under the theory, the longer the period of stuttering, the greater the likelihood that the next pulse from the triggering mechanism will activate the correct node.

The same slow buildup of priming that contributes to repetitions under the theory can also cause prolongations and blocks. During a prolongation, the articulators become locked into a continuant (open) position, while during a block, they become locked into an obstruent (closed) position, preventing speech by prohibiting airflow. Again, the reason is that in stutterers, inhibitory satellites of muscle movement nodes accumulate priming abnormally slowly, causing delays in self-inhibition of the parent node. The muscle movement nodes of stutterers can therefore remain activated for longer than normal durations, as occurs during prolongations and blocks.

Major disfluencies occur under this account when several muscle movement nodes happen to malfunction as a group. However, minor disfluencies, inaudible to the human ear, may occur when only a few muscle movement nodes malfunction. This may explain why cineradiographic analyses of stutterers' utterances exhibit abnormal transitions between sounds and asynchronies between lip and jaw movements, even though the utterances sound fluent to the unaided ear (Zimmermann, 1980); some fraction of the underlying muscle movement nodes are being activated at the wrong times.

Stuttering, Sequencing, and Timing

As discussed so far, stuttering reflects a disruption in the ability to produce the next muscle movement component in sequence, a problem involving content and sequence nodes. Several sources of evidence support this postulated sequencing difficulty (Van Riper, 1982), but stutterers also have a timing problem that is relevant to the node structure theory. As already noted, stutterers exhibit asynchronies between the muscles for articulation, respiration, and voicing, not only during overt instances of stuttering but also during seemingly fluent speech (Zimmermann, 1980). Relative to nonstutterers, stutterers also have difficulty duplicating the timing of a sequence of finger taps (M. H. Cooper & Allen, 1977), as if the timing nodes governing nonspeech activities are also more variable and susceptible to mistiming. Perhaps timing nodes suffer the same slow buildup of priming and sharp rebound from self-inhibition as content and sequence nodes. The observed deviations from periodicity may therefore occur when the sharp rebound suffices to reactivate the timing node.

The possibility that stuttering reflects a periodic disruption in the timing of muscle movements may explain phenomena such as the "rhythm effect," where fluency is enhanced when a stutterer speaks in time with a metronome or any other rhythmic stimulus (visual, auditory, or tactile), provided that the rhythm is not abnormally fast (Van Riper, 1982). That is, an externally generated rhythm to a shared perception–production timing node (Chapter 5) may help facilitate the timing of motor patterns that are prone to asynchrony in stutterers. Timing factors may also explain why stutterers often become more fluent when singing (Van Riper, 1982). The externally provided rhythm of the notes may help establish periodicity for the timing nodes controlling syllable production (although speed–accuracy trade-off explanations remain to be ruled out).

Practical Implications

The fact that stuttering originates within the muscle movement system in the node structure theory suggests that manipulating muscle movement factors can directly ameliorate stuttering. However, muscle movement factors are difficult to manipulate. Processes within the muscle movement system are largely automatic, that is, overlearned, fast, unconscious, beyond voluntary control, and in general unmodifiable (D. G. MacKay, 1982; 1987).

Muscle movement processes are also extremely complex. For example, over 100 different muscles may become engaged in producing a single word, and each muscle must get its appropriate nervous impulses at the required moment and in the proper sequence if the word is to be spoken without disruption. Because we currently don't know how muscle movement nodes interconnect and interact in real time, let alone how to modify these interactions, fundamental solutions to the problem of stuttering are a long way off. Understanding stuttering and the structure and dynamics of the muscle movement system for speech will remain an important theoretical goal for many decades to come.

9
Perceptual Feedback in the Detection and Correction of Errors

> Self-repairs are . . . rather complex phenomena . . . they involve quite disparate phonetic processes, such as self-monitoring, the production and detection of phonetic, lexical and other types of speech errors, self-interruption, prosodic marking of the connection.
>
> (Levelt, 1984, p. 105)

Self-inhibition is a central concept in the node structure theory. In addition to explaining the many sources of evidence discussed in Chapter 8, self-inhibition is needed to explain how perceptual feedback is processed during an ongoing action and to explain how errors are detected and corrected. Self-inhibition also makes sense of the way speech production becomes disrupted when normal speakers hear the sound of their own voice amplified and delayed by about 0.2 s. Self-inhibition even contributes to "node commitment," as I call the process of forming new or functional connections between nodes. However, I have developed the last two topics elsewhere (Chapter 10, and MacKay, 1987). The present chapter discusses only feedback processing and the detection and correction of errors. I first examine how errors are detected and corrected and the constraints these phenomena place on theories of the relation between perception and action. I then take these constraints into account in developing a theory of mechanisms underlying error detection and correction and the processing of self-produced feedback in general.

Constraints on Theories of Error Detection

Perceptual feedback plays a major role in detecting self-produced errors, and theories of feedback processing must explain (at least) four general characteristics of error detection: the rapidity of error detection, the detection of internal errors, differences between detecting correct versus incorrect responses, and differences between detecting self-produced versus other-produced errors.

The Time Characteristics of Self-Interruption

Studies of spontaneous self-corrections, reviewed in Levelt (1984), indicate that errors can be detected very quickly, even before the error has been completed in the surface output. After making an error in word selection, subjects often interrupt themselves immediately, sometimes before they finish the word containing the error. For example, a subject in Levelt (1984) began to say the word *black* but stopped after the initial /b/ and then produced the intended word *white*, as in "*b- . . .* er, *white*." These within-word interruptions usually occur when the output is factually incorrect (as in this example) and not simply infelicitious or open to misinterpretation. Speakers not only detect and correct their errors very rapidly, but they apparently comprehend the factual versus infelicitious nature of their error before they interrupt and correct themselves.

Internal Error Detection

Although the evidence previously discussed suggests that a speaker can *sometimes* detect errors before their full-blown appearance in the surface output, stronger evidence is needed on this point. After all, corrections sometimes fail to occur, or they may follow a full-blown error with considerable lag. Experiments on the detection of errors in internal speech provide the additional support, indicating that under appropriate circumstances, *all* (phonological) errors can be detected internally, without ever appearing in the surface output. Dell (1980) showed that when subjects produce tongue twisters, either overtly or to themselves, they make as many errors and detect the occurrence of these errors as often during internal speech as during overt speech (Chapter 8). Theories of error detection must therefore explain how errors can be detected before their appearance in the surface output, and why *external* feedback is unnecessary for error detection.

Self-Produced Versus Other-Produced Error Detection

Error detection differs in interesting ways, depending on whether the error is self-produced or other-produced. Other-produced errors are more easily detected for units that are large and meaningful, such as words, rather than small and meaningless, such as phonemes and phoneme clusters (Tent & Clark, 1980). Self-produced errors, on the other hand, are detected and corrected about equally often, regardless of the size or meaningfulness of the units involved (Nooteboom, 1980). Apparently speakers respond to their own errors with equal sensitivity, regardless of error size or type, whereas listeners are especially sensitive to errors involving large and meaningful units (Cutler, 1982), a fundamental constraint on theories of error detection.

Detection of Correct Versus Incorrect Responses

Detecting that a response is correct takes more time than detecting that a response is in error. Rabbitt, Vyas, and Feamley (1975) examined how rapidly

subjects performing a two-choice reaction time task could indicate that a given response was either correct or incorrect, and they found that these "validation times" were consistently longer for correct responses than for incorrect responses.

Editor Theories of Error Detection

Current theories have encountered difficulties in explaining the nature of error detection. To illustrate some of these difficulties, I examine the process of error detection in the class of theories known as "editor theories." I then develop a new and somewhat simpler theory of error detection that seems to overcome these problems.

The defining characteristic of an editor theory is a mechanism that "listens to" the visual, auditory, kinesthetic, or other feedback that results from output; compares this feedback with the intended output; identifies errors; and then computes corrections using a duplicate copy of information originally available to the motor control system. Baars et al. (1975) and Keele (1986) provide recent examples that fall within this class, but the editor theory tradition extends back at least to Freud (1901/1914). Recent variants in the editor theory class have added the concept of efference-copy or "feedforward," a preliminary representation of an about-to-be-produced action that enables editing before the actual occurrence of the output (Keele, 1986).

Problems with even the latest editor theories are everywhere apparent. If an editor knows the correct output all along, one wonders why the correct output wasn't generated in the first place. One also wonders why errors sometimes pass uncorrected. According to data of Meringer (1908) and Meringer and Mayer (1895), speakers only correct about 60% of all word substitution errors, as in "*table*, er, I mean *chair*." Perhaps the editor is subject to the same laws of fatigue or haste as the output system and therefore tends to err at the very time that the output system makes an error. However, this explanation calls into question the purpose of an editor that tends to break down at the time when it is needed.

Another set of problems arises from the assumption, implicit in editor theories, that perception and production involve separate components at all levels. That is, editors use the same mechanisms for the perception of errors as for perception in general. Then, once an error has been detected, the editor computes corrections for execution by a production system that is logically distinct and separate from the perceptual system. However, if systems for perception and production are distinct and separate, then the many parallels and interactions between perception and production discussed throughout this book become difficult to explain.

Editor theories also have difficulty explaining asymmetries between detection of self-produced versus other-produced errors. If an editor uses general perceptual mechanisms to detect errors, then different types of errors should be equally easy to detect in self-produced and other-produced speech. As already noted, however, experimental data do not support this prediction. Lexical substitution

errors (e.g., *table* misproduced as *chair*) are detected more frequently than phonological substitution errors (e.g., *publication* misproduced as *tublication*) in other-produced speech (Tent & Clark, 1980) but equally frequently in self-produced speech (Nooteboom, 1980).

The fact that detecting correct responses takes more time than detecting incorrect responses (Rabbitt et al., 1975) presents another problem for editor theories. As Broadbent (1971) points out, if the output is continually monitored for correctness versus incorrectness, editor theories predict the opposite result. Because correct responses occur more often than errors, the editor should be more practiced, and therefore faster at monitoring correct responses. The failure of this and other predictions suggests that editor theories require modification along lines discussed in the following section.

The Node Structure Theory of Error Detection

The node structure theory can be considered to combine aspects of editor theories and corollary discharge theories (discussed later in this chapter) to provide a coherent account of mechanisms underlying feedback processing and the detection of everyday errors. I begin with an analysis of the relationship between top-down and bottom-up (feedback) processes involving mental nodes and then develop an account of the detection of errors in self-produced versus other-produced speech, the time characteristics of error detection, and the time required to detect correct versus incorrect responses.

The Role of Internal Feedback in Error Detection

Both internal and external feedback can play a role in the detection of errors under the node structure theory. Recall that internal feedback consists of bottom-up priming transmitted from a subordinate node to its superordinate node as soon as the subordinate node becomes activated during production. External feedback likewise primes a just-activated superordinate node bottom-up, but this priming arrives later, following sensory analysis of the auditory or other perceptual consequences of the action.

The role of internal feedback in the detection of self-produced errors is extremely simple in the node structure theory but depends on understanding how errors occur. As discussed in D. G. MacKay (1982; and Chapter 7), output errors occur when a primed-from-above node fails to achieve greatest priming in its domain, and some other node becomes activated. An example is the error "Put the bag in the . . . ." instead of "Put the box in the car" (from Goldstein, 1968). Here the intended or primed-from-above node is *box*(noun), but *bag*(noun) has acquired greatest priming in the (noun) domain and has become activated under the most-primed-wins principle, giving rise to the error *bag*, rather than the intended *box*.

Such an error can be perceived, although not necessarily prevented, even before it appears in the surface output. Because activation is an inherent aspect of both perception and production, and because perception and production use the same mental nodes, activating the wrong node in production can cause virtually immediate perception of the error. In the preceding example, erroneous activation of *bag*(noun) implies potentially immediate perception of the unintended concept, "bag." This explains how errors can be detected so quickly. Perception of an error begins in the node structure theory even before the error has been expressed. Of course, as in perception in general, activation is necessary, but not sufficient for *awareness* of an error. Moreover, error detection is necessary, but not sufficient for error *correction*. Because *listeners* often fail to detect errors (e.g., Marslen-Wilson & Tyler, 1980), *speakers* can adopt a fairly liberal criterion in deciding whether or not to correct their own errors.

This view predicts that awareness of a specific error should always be part of awareness that an error has occurred. Because *perceiving* an error is yoked to activating the mental node that *produces* the error in the first place, speakers will never perceive that they made some error, but not know what the error was. Perception of self-produced error will never become dissociated from the specific content of the error that has occurred. Interestingly, this will *not* be true for perception of "other-produced" errors under the theory. For other-produced errors, the perceiver *can* become aware that an error has occurred, but not know with certainty what the error is. In what follows, I discuss various other predictions derived from the node structure theory of error detection and show how the theory handles the constraints on theories of error detection discussed previously.

Detection of Internal Errors

Although external feedback can facilitate error detection under the node structure theory, internal feedback is sufficient for detecting errors involving mental nodes. This explains why phonological errors can be detected in internal speech (Dell, 1980) without actually occurring in the surface output. An error can be detected as soon as the wrong node in an output hierarchy has been activated, that is, before and independently from activation of the lower level nodes for expressing the error and for processing the external feedback that results.

Detection of Correct Versus Incorrect Responses

Under the node structure theory, detecting an error involves a simpler and more direct process than detecting a correct response. The reason is that internal feedback (bottom-up priming) can lead directly to perception in the case of errors, but internal feedback is canceled by self-inhibition in the case of correct responses.

Errors involve the activation of an extraneous node, which results in bottom-up convergent priming of higher level nodes. These higher level nodes are not undergoing self-inhibition and can therefore become activated to provide a basis for

detecting the error. Indeed, higher level nodes receiving bottom-up priming as a result of an error are extremely likely to become activated under the most-primed-wins principle, relative to the nodes for producing the remainder of the output sequence. These intended-to-be-activated nodes are only receiving top-down priming, which as discussed in Chapter 7, is nonconvergent and therefore weak, relative to the bottom-up priming from an error, which converges and summates and is therefore relatively strong.

The same bottom-up convergent priming also occurs when the appropriate nodes have been activated in producing the correct or intended output. The difference is that higher level nodes representing a correct output *are* undergoing self-inhibition and so cannot become reactivated and provide the basis for detecting that the output is in fact error free. In the case of a correct response, self-inhibition cancels the internal feedback, which provides the normal signal for error detection. As a result, verifying that a response is correct will take more time than detecting that a response is in error, as Rabbitt et al. (1975) observed.

Self-Produced Versus Other-Produced Error Detection

The node structure theory suggests an interesting explanation for differences between the perception of self-produced versus other-produced errors. An output error will occur when a single extraneous node in any domain in any system becomes activated and activation of a node enables perception, no matter what the size of the surface units involved (segment, segment clusters, syllables, words, or phrases). This means that speakers will perceive self-produced errors equally often for small versus large units (all other factors being equal), exactly as Nooteboom (1980) observed.

However, perception of *other-produced* errors is quite different. Because of the principle of higher level activation, other-produced errors involving (smaller) phonological or phonetic units are likely to pass undetected. Because listeners don't normally activate either phonetic or phonological nodes during everyday perception of other-produced speech, listeners are less likely to perceive phonological and phonetic errors, just as Tent and Clark (1980) observed. However, phonological errors become easier to perceive if these errors alter the higher level interpretation of an utterance, because the nodes representing higher level interpretations routinely become activated under the principle of higher level activation.

The node structure theory also predicts that for normal inputs occurring at normal rates, listeners will be able to detect other-produced errors either at the word level or at lower (phoneme or letter) levels but not at both levels simultaneously (D. G. MacKay, 1987). Although this prediction remains to be tested, the nature of proofreading provides tentative or preliminary support. Misspellings are notoriously difficult to detect when reading for higher level content, and understanding the content of a passage is notoriously difficult when proofreading for misspellings.

A Comparison of Node Structure and Editor Theories

The node structure theory of error detection introduces a series of refinements that overcome the problems of editor theories discussed earlier. In brief overview–summary, the node structure theory resembles editor theories in the use of identical mechanisms for perceiving self-produced and other-produced inputs. However, because lower level (e.g., phonological) nodes become activated and self-inhibited during *production*, but not during *perception*, the node structure theory predicts major differences between detection of self-produced versus other-produced phonological errors.

The error detection mechanism in the node structure theory consists of many distributed processors, instead of a single centralized processor (such as the editor in editor theories). The distributed processors in the node structure theory correspond to the nodes that become primed and activated when producing the original output. The node structure theory also detects errors automatically and directly instead of "indirectly" via a comparison or matching process, as in editor theories. On the other hand, the node structure theory cannot immediately and directly perceive the correctness of a correctly produced output (unlike editor theories) because of the self-inhibition that follows activation of a correct or primed-from-above mental node.

Other Aspects of Feedback Processing

Theories of feedback processing must explain more than just error detection, and in this section I examine two other constraints on a general account of feedback processing: the missing feedback effect and the role of feedback in learning.

The Missing Feedback Effect

Chapter 4 described the verbal transformation effect, the fact that perception changes when an acoustically presented word is repeated via tape loop for prolonged periods. After many repetitions of the word *pace*, for example, subjects reported hearing words such as *face*, *paste*, *base*, *taste*, or *case*, and the rate of perceptual change increased as a function of time or number of repetitions (Warren, 1968). What must be explained in theories of perceptual feedback is why similar perceptual changes *fail to occur* when subjects produce the repeated input *themselves*. Lackner (1974) showed that the auditory feedback that accompanies repeated *production* of a word fails to trigger verbal transformations. Lackner's subjects repeated a word every 500 ms for several minutes, and they later listened to a tape recording of their own output. The subjects experienced the usual transformations when *listening* to the tape repeated word, but for some reason, experienced no perceptual transformations when *producing* the repeated word. This missing feedback effect is interesting, because acoustic events at the ear are identical when hearing the input during versus after production.

Why does on-line auditory feedback fail to trigger verbal transformations? Lackner (1974) and Warren (1968) attributed the missing feedback effect to corollary discharge, an efference-copy that accompanies the motor command to produce a word. This corollary discharge cancels or inhibits the external (proprioceptive and auditory) feedback resulting from producing the word, so that the on-line auditory input fails to bring about the fatigue-induced perceptual changes that are typically observed in verbal transformation experiments.

Taken by itself, the corollary discharge hypothesis has difficulty explaining the many interactions between speech perception and production discussed throughout this book (especially the effects of delayed auditory feedback, Chapter 10). Corollary discharge also fails to explain an additional symmetry in Lackner's own (1974) data, namely that no *production errors* resembling the perceptual errors occurred while subjects actively repeated the words. And Bridgeman (1986), Steinbach (1985), and D. M. MacKay (1973) summarize several additional problems with corollary discharge as an explanation of other aspects of perception, such as the phenomenon of perceptual stability following voluntary eye movements and the occurrence of visual suppression during saccades (the fact that psychophysical thresholds increase by about a half-log unit when faint test flashes are presented during or just before a saccade).

Corollary Discharge as Self-Inhibition

Under the node structure theory, the self-inhibition that follows activation of mental nodes is responsible for many of the phenomena that have been attributed to corollary discharge, and the missing feedback effect provides a good illustration of this corollary discharge as self-inhibition hypothesis. If corollary discharge corresponds to the activation of inhibitory satellites, the missing feedback effect becomes immediately apparent. Mental nodes for producing and perceiving a word such as *police* are identical, so that when someone repeats a word such as *police*, auditory feedback returns as priming to the just-activated nodes that produced it, but has little effect, because these nodes are undergoing self-inhibition following production. Self-inhibition of lower level nodes also prevents returning external feedback from reaching higher level nodes, so that self-produced inputs fail to cause verbal transformations, which is the missing feedback effect.

The corollary discharge as self-inhibition hypothesis also explains the absence of output errors during repeated *production* of a word. For example, repeating the word *police* causes satiation of the corresponding mental nodes, but because top-down priming is unique (i.e., only a single node normally receives first-order priming in any given domain at any given time), only *police*(noun) and no other lexical node receives systematically increasing top-down priming and becomes activated under the most-primed-wins principle. The uniqueness of top-down priming during production therefore reduces the likelihood of *production* errors resembling those that occur in perception and explains this additional asymmetry in Lackner's (1974) data.

The node structure characterization of the missing feedback effect makes several interesting sets of predictions not made by other theories. One concerns the effects of production on perception. If subjects produce a word repeatedly for, say, 10 minutes, either *internally or overtly*, and then listen to the same word repeated via tape loop, the subsequent rate of perceptual change will be *at least as great* as if they had already heard the word repeated for 10 minutes. Another set of predictions concerns a *production* condition where subjects repeat a word (again either internally or overtly) while hearing a *phonologically and semantically different* input word repeated over earphones at the same time. After, say, 10 minutes, the subjects are signaled to stop production and report perceptual transformations of the input word as it continues to repeat for, say, 5 minutes. Under the node structure theory, perceptual transformations of the input word will be significantly slower in this *production* condition than in the usual *listening* condition where subjects attend to the repeating input word for the entire 15-minute period.

Feedback and the Formation of New Connections

Internal and external feedback have other functions in addition to error correction in the node structure theory. One concerns the commitment of new connections, a process discussed in detail elsewhere (D. G. MacKay, 1987). What follows is a severely truncated summary.

Feedback is required during early stages of acquiring a skill in order to ensure that an action is appropriate or has the desired effect. In learning to produce a German trilled /r/, for example, it is essential to hear the trilled /r/ that one produces in order to know whether one's muscle movements achieved the desired result. Without external feedback during this early phase of acquisition, practice can strengthen the wrong connections and lead to inappropriate actions (D. G. MacKay, 1981). During later stages of skill acquisition, the linkage strength of already formed connections can become further strengthened in the complete absence of feedback, and this explains how mental practice can improve skilled performance without the help of either sensory or experimenter-generated feedback (D. G. MacKay, 1981).

In short, external feedback is unnecessary for execution of behavior during later stages of skill acquisition, when connections for an output hierarchy have already been formed and strengthened. This explains why experimentally deafferented monkeys can accurately produce a previously learned and highly practiced response (Bizzi, Dev, Morasso, & Polit, 1978). It also explains why adults can in general produce intelligible speech when their auditory feedback has been masked, distorted, or eliminated altogether as a result of acquired deafness (Siegel & Pick, 1974). The effects of delayed and amplified auditory feedback (DAAF) on speech production provide the only alleged counterexample of feedback seeming to influence the control of highly skilled behavior. When auditory feedback is amplified and delayed by about 0.2 s, normal speech production

becomes severely disrupted. This remarkable and highly reliable phenomenon is discussed in detail in Chapter 10, and contrary to surface appearances, it does not support the feedback control hypothesis.

Constraints on Theories of Error Correction

Once an error has been detected, how is it corrected? Error correction does not proceed on a trial-and-error basis. Almost immediately after becoming aware of making an error, we also become aware of the exact locus and nature of the error, how to correct it, and how to signal the nature of the error to the listener. Theories of error correction must therefore explain three general classes of phenomena: the process of error signaling, the time course of error correction, and regularities in the way that errors are corrected, such as the structural and lexical identity effects discussed in a following section.

The Time Course of Error Correction

The process of error correction seems to begin almost immediately after an error has been committed. After making an error in a multiple choice reaction time task for example, subjects do not go through the entire process of deciding all over again what response to make. Rabbitt and Phillips (1967) showed that the interval between an error and its correction is much shorter than the average time required to produce the original response correctly. Moreover, the time required to detect and correct an error seems to remain constant when the original response time is extended by reducing stimulus–response compatibility, increasing the number of response alternatives, or increasing the number of stimuli appropriate to each response. As Broadbent (1971, p. 305) points out, the fact that task difficulty fails to influence the time to correct errors rules out the hypothesis that errors are corrected by reiterating the original response process, even if the correction process can begin before the occurrence of the error in the surface output. As Broadbent (1971) also points out, the rapidity of error detection and correction is surprising in view of the refractory effects that are generally observed when subjects must make two responses in rapid succession (Chapter 8). The (hypothesized) internally generated "error signal," which is used for correcting an error in editor theories, is clearly unlike an external signal, which follows soon after the first in the psychological refractory paradigm.

Error Signaling

When speakers stop to correct an error during speech production, they often introduce a term such as "uh," "er," or "I mean," which signals both the occurrence and nature of the error. These error signals fall into two classes: rejection error signals and supportive error signals. Speakers usually introduce rejection error signals when the error results in a *factually incorrect* statement, as in "feeding it

into the computer, *sorry* [*no, rather*], the printer." A change in prosody often accompanies these rejection signals, so that the repair receives, for example, greater stress than the trouble segment. However, when the error results in a *factually correct* but inadventitious statement, prosody remains normal, and a supportive error signal is introduced, as in, "Go into the room, *thus* [*that is*], the kitchen" (Levelt & Cutler, 1983). Theories of error correction must explain how speakers can so rapidly distinguish and signal these two types of errors.

The Structural and Lexical Identity Effects

Words used in correction usually maintain the same syntactic structure as the words they correct, as if the correction and corrected words represent conjoints in a coordinate structure (e.g., nouns joined by *and*). In "Is he seeing, er, interviewing Mary?" for example, the erroneous word, *seeing*, belongs to the same syntactic class as the correction, *interviewing* (Levelt, 1984).

Indeed, erroneous words and their corrections are often more than just structurally identical. Corrections that repeat ideas in the original utterance usually copy the original words verbatim, and this is especially true of factual errors, as in, "He talked frequently with his sister, uh, I mean, he talked frequently with his *mother* [emphasized]." Theories of error correction must therefore explain why corrections are usually structurally *and* (if possible) lexically identical to the originally intended words.

The Node Structure Theory of Error Correction

The efficiency of error correction is readily explained in the node structure theory. As previously discussed, error detection occurs almost as soon as the error itself in the theory, and error correction can begin shortly thereafter, when the correct or primed-from-above node becomes activated. By way of illustration, consider again the error "Put the *bag*, I mean, *box* in the car." Assume that *bag*(noun) has been activated, that the error has been detected, and that the flow of speech has been interrupted. During this time, the appropriate or primed-from-above node, *box*(noun), continues to accumulate priming from its immediately superordinate noun phrase node (until self-inhibition begins), as well as from other nodes representing, say, the situational context. This means that if the most recently activated sequence node (NOUN) is reactivated, the appropriate node, *box*(noun) in this example, will be activated, resulting in direct perception of the intended output. Perception of the intended output in turn enables immediate inferences as to the nature of the error and the type of error signal required.

Because the appropriate node has now been activated, the appropriate output can also be produced. Indeed, activating the appropriate or primed-from-above node is *part of* producing the correct response. No special mechanism and no extra time is required for computing which elements are in error or how to correct them. The correct response is immediately available under the node structure

theory. The correction is produced by activating the same sequence nodes and hierarchy of content nodes, except of course, for the one content node that was activated in error (plus all of its subordinate nodes). At the time when repair occurs, the appropriate node in the domain of nodes containing the error will have greatest priming and become activated as part of the repair.

Error-Correction Errors

Besides being consistent with current constraints on theories of error correction, the node structure theory makes some interesting predictions for future test. One concerns a class of "error-correction" errors. The node structure mechanism for error correction is itself susceptible to errors. The main prerequisite for error-correction errors is that during the normal course of producing the remainder of a sentence, the sequence node responsible for an error is due to become activated again soon after the error occurs.

By way of illustration, consider again the substitution error "Put the bag in the . . ." instead of "Put the box in the car." In producing the intended sentence, the NOUN sequence node must become activated in quick succession, first to produce *box* and then to produce *car*. Thus, if the rate of speech is sufficiently rapid when the substitution of *bag* for *box* occurs, the theory predicts a double error of the form, "Put the bag in the box." That is, the correct noun for the earlier slot will become activated (now erroneously) in the subsequent noun slot. The reason is this: After the inappropriate node *bag*(noun) has been activated, the originally appropriate node, *box*(noun), continues to accumulate priming, is likely to achieve greatest priming in the (*noun*) domain, and becomes activated automatically when its sequence node *NOUN* is applied to the (noun) domain soon after in the course of producing the remainder of the sentence. It is as if a noun has been corrected but the wrong noun!

Because error detection and self-interruption is so rapid, however, the likelihood of observing such error-correction errors is relatively low. Nevertheless, error-correction errors do occur (see Goldstein, 1968 for examples), and their occurrence can be viewed as support for the node structure theory of error correction. (Related processes are also required in explaining the subtly different "bumper-car errors" in Stemberger, 1985.)

The Structure of Error Correction

The node structure theory readily explains the structural and lexical identity effects, the fact that corrections copy the syntactic structure and, if necessary, as many correct words as possible from the original utterance. Under the node structure theory, corrections occur by simply reactiving the same sequence node(s) following detection of an error, and in the case of correct or "copied" words, this means activating the same content nodes as well. Only the content node that was activated in error will not be reactivated. This "erroneous" node will be recovering from activation, and the appropriate node will (because of

temporal summation) achieve greatest priming in its domain and become activated automatically as part of the repair. Under the theory, this appropriate node typically will constitute the only difference between the correction and the corrected output, error signals aside. Thus, even the prosody of the new or corrected utterance should correspond to that of the original or intended utterance, and this is exactly what Levelt (1984) found. Levelt (1984) recorded a large number of naturally occurring errors, erased any error signals or repeated words, and spliced the correction itself onto the original utterance. He then had subjects listen to the resulting utterances, and found that the spliced utterances sounded completely natural, even when up to 3 s of output had been deleted.

10
Disruptive Effects of Feedback

> All purposeful behavior may be considered to require negative feedback.
> (Rosenblueth, Wiener, & Bigelow, 1943, pp. 22).

> This is an exciting period for the study of action.
> (Gentner, 1985, p. 184).

Although feedback can help in detecting and correcting errors, and in learning new behaviors, feedback can also disrupt ongoing action. Many perceptual-motor systems exhibit feedback-induced disruptions, and speech production under conditions of "delayed auditory feedback" provides the most dramatic and carefully studied example. When auditory feedback from speech is recorded and then played back with amplification to the ears after a delay of about 0.2 s, speech becomes severely disrupted. Under these conditions, proficient speakers repeat, prolong, and substitute speech sounds, sometimes producing phonemes that are not part of any language familiar to them (B. S. Lee, 1950). The present chapter examines this and other feedback-induced disruptions and the constraints they impose on theoretical relationships between perception and action.

Experiments on delayed and amplified auditory feedback (DAAF) flourished in the 1950s and 1960s, because the phenomenon seemed compatible with the cybernetic or feedback control theory in vogue at the time (Neumann, 1984). However, the last 15 years have seen virtually no studies of DAAF, not because we now understand the phenomenon or because it proved unimportant or unreplicable; the effects of DAAF are highly reliable, and remain central to an understanding of the relation between perception and action. Rather, the problem was that the theory that originally stimulated interest in the phenomenon proved inadequate for explaining the new experiments conducted in its name. Only a new theory with predictions for future test can revive interest in DAAF, and my goal here is to develop such a theory. In the Epilogue, which follows this chapter, I examine some additional lessons of DAAF for the many other "abandoned phenomena" in psychology.

Constraints on Explanations of Feedback Disruption

Theories of feedback processing must account for five basic characteristics of DAAF disruption: the conditions for disruption, the automaticity and generality of the phenomenon, factors influencing the delay that produces maximal interference, and factors influencing the degree of disruption.

Basic Conditions for DAAF Disruption

The most basic condition for disruption is, of course, the fraction of a second delay in external feedback. Another basic condition, frequently overlooked in explanations of feedback disruption, is amplification of the returning feedback. Articulatory errors do not occur unless auditory feedback returns at louder than normal levels (see, for example, Fairbanks & Guttman, 1958). The importance of amplification prompted my use of the term *DAAF* (delayed and amplified auditory feedback), rather than the usual *DAF* (delayed auditory feedback).

At one time, amplification was assumed to have a masking function. Unless masked by *amplified* auditory input, bone-conducted feedback could be used for feedback control (because the skull transmits voice-correlated vibrations that cannot be delayed). However, masking can't be the sole effect of amplification (see also Howell & Powell, 1984). DAAF errors continue to occur when low-frequency auditory components are eliminated by means of a low-band pass filter (Hull, 1952), whereas under the masking hypothesis, low-band pass filters should fully restore fluency: lower level frequencies that bone-conduction primarily transmits will no longer be masked. Moreover, increasing feedback amplitude over a wide range continues to increase the probability of errors, long after air-conducted feedback could be assumed to have masked the bone-conducted feedback (Black, 1951). Some other factor is therefore needed to explain the effect of feedback amplification on DAAF errors (see also Howell & Archer, 1984).

Automaticity of the Phenomenon

As many investigators have noted, the automaticity of DAAF interference resembles pathological stuttering. Like pathological stuttering, DAAF errors cannot be voluntarily corrected or avoided if subjects are speaking at maximal rate and without pauses. Also like pathological stuttering, practice with disrupted output fails to eliminate DAAF interference (unless subjects are permitted to speak in short bursts, thereby desynchronizing their output and returning feedback; see Smith, 1962; see also D. G. MacKay & Bowman, 1969, for a transfer paradigm where practice *can* genuinely reduce DAAF interference).

These similarities aside, pathological and DAAF stuttering differ in some respects that also require theoretical explanation. One difference concerns the loudness of speech. Speakers hearing DAAF involuntarily speak with greater than normal amplitude and increase their output amplitude in proportion to the amplification of the returning auditory feedback (Black, 1951). This greater than normal vocal intensity is not characteristic of pathological stuttering and is theoretically puzzling in other ways (see the following discussion of feedback control theory).

Generality of the Phenomenon

The repetition errors that occur under DAAF are not limited to speech. For example, expert generators of Morse code also make repetition errors, producing n + 1 instead of n *dits* or *dahs* when they hear the click of their key amplified and delayed by about 0.2 s (Smith, 1962). DAAF also causes similar errors in whistling, singing, rhythmic hand clapping, and the playing of musical instruments (Kalmus, Denes, & Fry, 1955), indicating that many perceptual–motor systems are susceptible to repetition errors and that repetition errors may represent a general effect of DAAF.

Factors Related to the Critical Delay

The most interesting fact about DAAF has proved the most difficult to explain: the existence of a critical delay that produces maximal interference. Not all delays are equally effective in disrupting fluency. For adults, disruption increases as a function of feedback delay up to about 0.2 s and then decreases with longer delays but never disappears completely, even with delays as long as 0.8 s (Figure 10.1, D. G. MacKay, 1968). In this section I examine three factors that have been shown to influence or covary with the delay that produces maximal disruption of speech. I then discuss how these factors differ from those influencing overall disruption across all delays.

Age and the Critical Delay

D. G. MacKay (1968; see also Ratner, Gawronski, & Rice, 1964) showed that the delay that produces maximal disruption of speech varies as a function of age: 0.2 s for adults, 0.375 s for children aged 7 to 9, and approximately 0.725 s for children aged 4 to 6. Figure 10.1 illustrates the shape of the function for repetition errors, but similar functions are obtained for other error types (Fairbanks & Guttman, 1958) and for other measures of rate such as "correct syllable duration," where the effect of errors has been factored out (D. G. MacKay, 1968). Although the precise location of the peak interference delay for 4- to 6-year-old children is indeterminate in Figure 10.1, other data discussed in D. G. MacKay (1968) corroborate a hypothesized peak of about 0.70 to 0.75 s. An effect of age on overall degree of disruption can also be seen in Figure 10.1, but I will leave degree-of-disruption factors for later in the chapter.

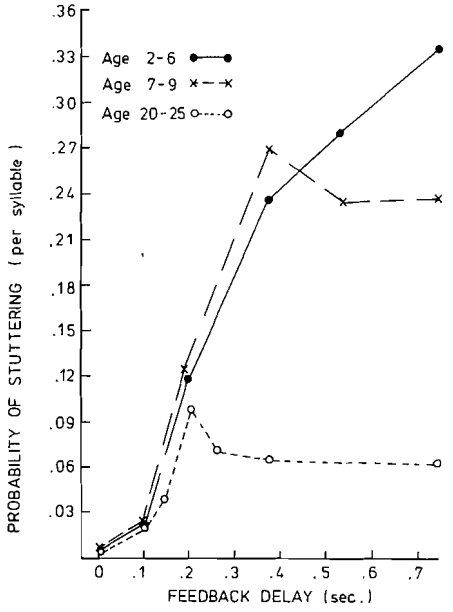

FIGURE 10.1. The probability of stuttering (per syllable) under delayed and amplified auditory feedback (DAAF) as a function of subject age and feedback delay. (Adapted from "Metamorphosis of a Critical Interval: Age-Linked Changes in the Delay in Auditory Feedback That Produces Maximal Disruption of Speech" by D. G. MacKay, 1968, *Journal of the Acoustical Society of America, 19,* pp. 816–818. Copyright 1968 by the Acoustical Society of America. Reprinted by permission.)

Potential Rate and the Critical Delay

Potential rate, operationally defined as a subject's maximum speech rate under ideal speaking conditions, is strongly correlated with the peak interference delay. D. G. MacKay (1968) first determined the potential rate for 32 subjects producing sentences with synchronous (undelayed) feedback, then had these subjects speak under DAAF, and determined the delay that produced maximal interference. Results for these two tasks were positively correlated (over $+0.5$, $p < .05$). That is, the slower a subject's potential rate (maximum speech rate with undelayed feedback), the longer the DAAF delay required to produce maximal interference in that subject. As D. G. MacKay (1968, pp. 818–819) points out, "Clearly, some factor or set of factors limiting a subject's maximum rate of speech must determine the (temporal) locus of delayed auditory feedback interference." I argue in the following discussion that the recovery cycle of mental nodes may be one of these factors.

Syllable Repetition and the Critical Delay

Syllable repetition seems to eliminate the peak interference effect. Data published for the first time in the following section indicate that manipulating feedback delay has relatively little effect on errors or speech rate under DAAF when subjects simply repeat a single syllable as opposed to producing a normal sentence.

D. G. MacKay and Burke (1972)

D. G. MacKay and Burke (1972) conducted a series of experiments systematically comparing the effects of DAAF delay on production of repeated versus nonrepeated syllables. Subjects (N = 40) either read a 20-syllable sentence or repeated a single syllable (*puh*, *buh*, *guh*, *nuh*, or *suh*) 20 times. In both conditions, subjects spoke at maximum rate until signaled to stop, about 5.0 s after speech onset. Sound pressure level was amplified to approximately 95 dB over earphones the subjects were wearing, and feedback was delayed by either 0, 50, 100, 150, 180, 200, 250, or 300 ms on any given trial. Order of presentation of the 8 delay conditions was randomized across both subjects and materials.

The results appear in Figure 10.2. As can be seen there, repetition errors and mean syllable duration varied in the characteristic way as a function of feedback delay during sentence production. Both measures of interference increased to a maximum with a delay of about 200 ms and decreased thereafter.

Syllable repetition gave a strikingly different pattern of results, however. No errors occurred for any delay during syllable repetition, and the time to repeat a syllable remained essentially constant over the 8 delay conditions (Figure 10.2). Figure 10.2 seems to suggest the possibility of a double peak in the syllable repetition function, one peak at 50 ms and another peak at 200 ms. However, we were unable to verify existence of these peaks statistically or to replicate them in subsequent experiments.

In either of these subsequent experiments, subjects (N = 20) repeated a digit (either *two*, *four*, or *eight*) while hearing their auditory feedback amplified and

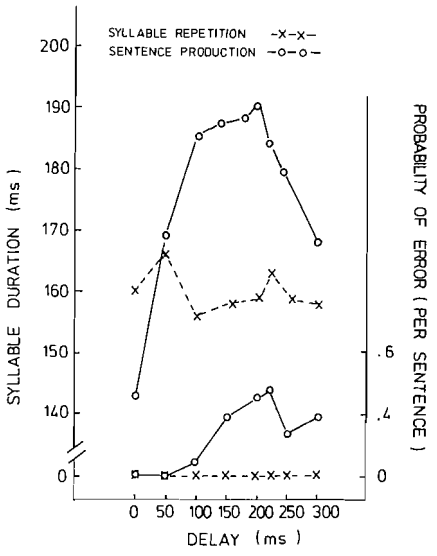

FIGURE 10.2. Syllable duration (left ordinate) and probability of error (right ordinate) for the production of sentences versus the repetition of syllables (D. G. MacKay & Burke, 1972).

delayed, and again, syllable duration remained relatively constant as a function of feedback delay. However, when the same subjects counted as rapidly as possible from 1 to 10, the 200-ms peak characteristic of connected sentences reappeared. This finding indicates that repetition per se was responsible for the flat feedback delay function in Figure 10.2 and not the nature of the syllables in the syllable repetition condition.

Another interesting difference between producing repeated versus connected syllables is the slower rate for syllable repetition in the 0-delay condition (Figure 10.2). This difference was especially marked toward the end of the 5-s production period. Syllable repetition became progressively slower over the 5 s and over the course of the experiment, as if a fatigue or satiation process (Chapter 1) were reducing the maximal rate of speech.

Factors Unrelated to the Critical Delay

Two factors *do not* influence the critical delay. These are the actual rate of speech and practice, and explaining why these factors are unrelated to the critical delay also presents a challenge for theories of DAAF.

Actual Rate of Speech

In contrast to *potential* rate, that is, a subject's maximal speech rate with synchronous feedback, *actual* (voluntarily determined) rate of speech under DAAF is unrelated to the delay that produces maximal interference. The critical delay remains unchanged when adults voluntarily speak more slowly. As can be seen in Figure 10.3 (D. G. MacKay, 1968), the frequency of repetition errors per syllable remains greatest at the 0.2-s delay for adults speaking at three different rates under two conditions of delay. This finding suggests that mechanisms determining the *actual* rate of speech must differ from those limiting the *maximal* rate (see following).

Practice (Language Familiarity)

Practice or language familiarity is another factor that seems unrelated to the peak interference delay. For example, when bilinguals produce their more and their less familiar language under DAAF, the delay that produces maximal interference with their speech remains constant (D. G. MacKay, 1970b; see also Figure 10.4).

Factors Influencing Degree of Disruption

Five factors have been shown to influence the overall degree of disruption independent of feedback delay. I have already mentioned one, level of amplification. Another factor influencing degree of disruption is actual speech rate. The law of speed–accuracy trade-off that characterizes other errors also characterizes errors under DAAF (Figures 10.3 and 10.4; see also Kodman, 1961; D. G.

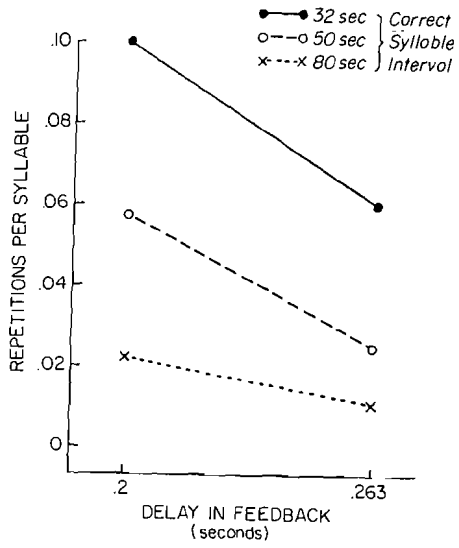

FIGURE 10.3. Repetition errors under two delay conditions for three rates of speech. (Adapted from "Metamorphosis of a Critical Interval: Age-Linked Changes in the Delay in Auditory Feedback That Produces Maximal Disruption of Speech" by D. G. MacKay, 1968, *Journal of the Acoustical Society of America, 19*, pp. 816–818. Copyright 1968 by the Acoustical Society of America. Reprinted by permission.)

MacKay, 1968; 1971). This finding indicates again that mechanisms for voluntarily speaking more slowly differ from mechanisms that limit maximal or potential speech rate. D. G. MacKay (1968) showed that errors under DAAF correlated *positively* rather than *negatively* with a subject's *potential* rate.

Another factor influencing degree of DAAF disruption is familiarity or practice with the materials being produced. For example, bilinguals stutter more when producing their less familiar language under DAAF (Figure 10.4), and

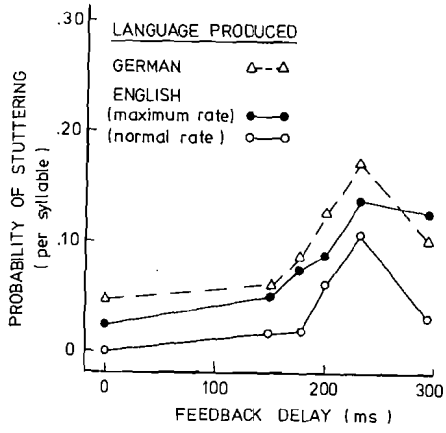

FIGURE 10.4. The probability of stuttering (per syllable) as a function of feedback delay for bilingual Americans speaking English (solid circles) and German (triangles) at their maximum rate and English at their normal rate (open circles). (Adapted from "How Does Language Familiarity Influence Stuttering Under Delayed Auditory Feedback?" by D. G. MacKay, 1970, *Perceptual and Motor Skills, 30*, p. 663. Copyright 1970 by Perceptual and Motor Skills. Reprinted by permission.)

practice in producing a sentence with synchronous feedback reduces the probability of stuttering when subjects subsequently produce the practiced sentence under DAAF (D. G. MacKay, 1970b; D. G. MacKay & Bowman, 1969). Overall disruption also diminishes with age, because practice, familiarity, or experience in producing speech increases as children grow older (Figure 10.1; D. G. MacKay, 1968).

Finally, altering either the normal muscle movements or the returning auditory feedback reduces overall interference under DAAF. D. G. MacKay (1969d) had subjects produce an "accent" by actively contracting velar muscles so as to nasalize all of their speech sounds. These unusual muscle contractions produced unaccustomed acoustic feedback and significantly reduced DAAF interference (with controls for speech rate and intensity level of the returning feedback). In a second experiment, D. G. MacKay (1969d) had subjects "passively" alter the nasal quality of their acoustic feedback by holding their nose while speaking, and this procedure also diminished the degree of DAAF interference. Hull (1952) and Roehrig (1965) used band pass filters to "passively" distort returning auditory feedback and likewise reported that DAAF disruption diminished with degree of distortion. The only experiment *not* reporting diminished DAAF disruption with feedback distortion is Howell and Archer (1984). However, their speech production task involved vowel repetition, and effects of DAAF are known to differ for repeated versus connected syllables (see preceding discussion; and Chase, 1958).

Feedback Control Theory and Feedback Disruption

I turn now to theoretical explanations of the effects of DAAF. Feedback control theories provided the initial framework and impetus for studies of DAAF but proved incapable of explaining either the overall degree of interference or the peak interference delay. This, along with some other problems discussed later, contributed to the current unpopularity of these theories in cognitive psychology (but see Holland et al., 1986; Rumelhart, Smolensky, McClelland, & Hinton, 1986, for a feedback control hypothesis involving internal feedback arising from thought or internally generated actions).

The main assumption of traditional feedback control theories is that external feedback from an ongoing action plays a direct role in controlling subsequent action, and this feedback control assumption has proven useful in describing innate regulatory behaviors such as the pupillary reflex. (See e.g., Oatley, 1978. Onset of a bright light causes reflex pupillary contraction, diminishing pupil diameter, and thereby reducing the amount of light falling on the retina, a feedback effect that in turn causes diminished contraction, until some goal or control point is reached.)

Feedback control theories have proven less useful in describing learned and highly skilled behaviors such as speech. Several different feedback control theories have been advanced for speech production (D. G. MacKay, 1969d), and all predict that distorting or eliminating feedback should cause interference. Of

course, masking or distorting auditory feedback has no such effect, and feedback control theories usually assume that speakers switch to bone-conducted feedback for controlling speech under these unusual circumstances. However, speech production also remains unimpaired for many months after an injury that causes complete and total deafness (Siegel & Pick, 1974), a finding that makes it difficult to imagine that articulation requires auditory control of any kind. Articulatory adjustments for producing speech sounds have a time span of milliseconds not months.

All versions of feedback control theory also have difficulties with both the detailed and the more general effects of DAAF. For example, why is there a delay that produces maximal disruption of speech? Under feedback control theory, disruption should either remain constant or increase monotonically as a function of delay, rather than *decreasing* after some critical delay (D. G. MacKay, 1969d). Why is it necessary to *amplify* the returning feedback in order to bring about articulatory errors? Why do subjects speak *louder* when their amplified auditory feedback is delayed? After all, *undelayed* amplified feedback causes people to speak *softer* (as expected under feedback control theory; see Siegel & Pick, 1974). Why does distortion of returning auditory feedback *reduce* DAAF interference? Under feedback control theory, distortion should *increase* the difficulty of feedback control. Why does practice or familiarity with a sentence reduce DAAF interference? Under the feedback control theory of Adams (1976), practice strengthens an internal trace of the expected feedback, and successive movements are driven by the discrepancy between the ongoing feedback and the expected feedback or feedback trace. This means that practice should *increase* rather than decrease the probability of errors for sentences produced under DAAF. These and other discrepancies suggest that articulation is not under direct feedback control and that a new explanation for disruptive effects of feedback is needed (see also Siegel & Pick, 1974). The remainder of this chapter develops such an explanation.

Feedback Disruption and the Recovery Cycle

In order to underscore the intimacy of the relation between self-inhibition and the processing of feedback in the node structure theory, I will first review the potentially disruptive effects of bottom-up internal feedback, noted briefly in Chapter 8. My thesis is that preventing these potentially disruptive effects of external and internal feedback is one of the main functions of self-inhibition. However, DAAF happens to bypass this and other defense mechanisms that have evolved to prevent these disruptive effects.

Potentially Disruptive Effects of Internal Feedback

In the case of internal feedback, the problem is this. Because mental nodes receive both bottom-up and top-down connections, internal feedback can poten-

tially cause reverberatory activation of mental nodes at every level in the system. During production, a superordinate node becomes activated and primes its subordinate nodes via top-down connections. However, the subordinate nodes become activated soon thereafter, priming their superordinate node via the bottom-up connections required for perception using exactly the same nodes. Self-inhibition is therefore needed at every level to ensure that bottom-up priming from subordinate nodes does not lead to the reactivation of just-activated nodes.

Disruptive Effects of External Feedback

Self-inhibition is also required to prevent similar effects resulting from external feedback. Like internal feedback, the external feedback that arises automatically from sensory analysis of auditory or other perceptual consequences of an action introduces bottom-up priming, which could potentially cause repeated reactivation of the lowest level mental nodes that gave rise to the feedback in the first place. To prevent such reactivations, self-inhibition must continue for a relatively long time, that is, at least as long as it takes for the lower level muscle movement nodes to become activated—for the, say, airborne feedback to reach the ears—and for the sensory analysis nodes to process the feedback and deliver priming to the mental nodes that originated the output.

By way of illustration, consider node *Z*, a low-level mental node that becomes activated and primes the muscle movement nodes that eventually give rise to auditory output. When this auditory output arrives at the ears, sensory analysis nodes automatically process this feedback and eventually prime *Z*, the node responsible for generating the output in the first place. This feedback-induced priming could result in the reactivation of *Z*, except that it arrives during *Z*'s period of self-inhibition. As a consequence, *Z* cannot accumulate enough priming to become most primed in its domain and cannot become reactivated when the triggering mechanism is applied to its domain during ongoing production of the remainder of the word or sentence.

Other Defenses Against Feedback-Induced Disruption

Self-inhibition is not the only defense mechanism that has evolved to prevent disruptive effects of self-produced external feedback. Stapedial attenuation provides a similar sort of defense. The stapedius muscle in our middle ear contracts just before we begin to speak and remains contracted throughout the period of speech, thereby attenuating the amplitude of eardrum vibration in response to hearing one's own voice.

Stapedial attenuation is sometimes viewed as a feedback-induced reflex for preventing damage to the eardrum that might otherwise arise from prolonged screaming. However, preventing injury cannot be the sole purpose of stapedial attenuation (see also Simmons, 1964). For example, stapedial activity also accompanies everyday speech and whispering at levels that cannot possibly cause

peripheral damage. Nor is stapedial attenuation an externally triggered reflex. Borg and Zakrisson (1975) were able to actually see, as well as electromyographically record, stapedial attenuation in otherwise normal speakers with a perforated eardrum, and they found that stapedial contraction preceded vocalization by about 75 ms (even when subjects spoke with very low acoustic intensity). Stapedial attenuation must therefore arise from a central command that precedes vocalization, rather than from a peripheral reflex triggered by vocal feedback. An interesting possibility under this "central command hypothesis" is that stapedial attenuation will precede and accompany production of internal speech in the complete absence of auditory feedback from the voice (Chapters 8 and 9).

Stapedial attenuation has other consequences under the node structure theory in addition to providing defense against disruptive effects of external feedback. For example, stapedial attenuation adds to the list of differences between perception of self-produced versus other-produced speech and represents another factor in the conspiracy of factors (together with self-inhibition and bone-conducted feedback) that contribute to the fact that one's own voice sounds differently during ongoing speech production than during a subsequent replay of the same sounds via tape recorder.

The Recovery Cycle Explanation of DAAF Interference

Under the recovery cycle hypothesis, feedback plays no direct role in controlling the form of a highly skilled behavior, and DAAF has its effects by overcoming the defenses against feedback-induced disruption of action. Under this view, the exact duration of the feedback delay is critical. DAAF must arrive at the point in time when mental nodes originally responsible for the output and feedback are especially susceptible to reactivation. Feedback arriving at the time when these just-activated mental nodes are self-inhibited can have no effect. To cause significant interference, DAAF must arrive slightly later, during the hyperexcitable phase of these just-activated nodes, when their level of priming is already greater than normal.

As noted in Chapter 8, the pattern of segment repetition in the structure of words suggests that hyperexcitability peaks at about 200 ms after a phonological segment node has been activated and returns to normal or spontaneous level by about 300 ms following activation.

The maximal influence of DAAF with a delay of about 0.2 s therefore reflects an effect of feedback-induced priming arriving with sufficient strength at the critical 0.2-s period in the recovery cycle of just-activated nodes. Amplification of the returning feedback adds further to the priming of the just-activated nodes, and when these combined sources of priming exceed the top-down priming for the appropriate nodes in the same domain, these just-activated nodes are automatically reactivated under the most-primed-wins principle, so that the output resembles stuttering (D. G. MacKay & MacDonald, 1984). In order to simplify exposition, I have, of course, ignored the time it takes for muscles to move following activation of a phonological node, the time it takes airborne auditory

feedback to arrive at the ears, and the time it takes sensory analysis nodes to process the feedback and deliver bottom-up priming to the just-activated phonological node.

Under this account, bottom-up priming arising from DAAF causes repetition errors directly and causes substitution errors indirectly via summation with top-down priming being transmitted to the intended or appropriate phonological nodes. If this summated top-down and bottom-up priming happens to make a new combination of distinctive feature nodes most primed in their domains when their activating mechanisms are applied, speakers could produce a phoneme that is not part of any language familiar to them, just as B. S. Lee (1950) observed.

The fact that subjects speak louder at the most disruptive delays (Black, 1951) may reflect an attempt to augment top-down priming and thereby enable the appropriate nodes to dominate in the most-primed-wins competition with bottom-up priming from DAAF. Transfer effects of practice or language familiarity (D. G. MacKay, 1982) are explained in a similar way and follow directly from the linkage strength assumption of the node structure theory. That is, practice without DAAF reduces interference when a sentence is later produced under DAAF because linkage strength of top-down connections will increase, thereby enabling top-down priming to compete more effectively against bottom-up priming from DAAF.

Consider now the factors influencing the delay that produces maximal interference. Under the recovery cycle hypothesis, two underlying processes determine the peak interference delay. The main one concerns the temporal characteristics of rebound hyperexcitability in sequence and content nodes. This rebound from self-inhibition is automatic but individual specific, varying with individual-specific factors such as age and experience (see following discussion). Rebound from self-inhibition is unrelated to processes governing the *actual* (voluntarily specified) rate of speech, determined by the oscillation rate of timing nodes.

However, temporal characteristics of recovery from self-inhibition *do* influence a subject's *potential* rate (maximal speech rate with synchronous feedback). Potential rate is determined in part by the rate at which low-level nodes can be reactivated (for a given error criterion, see D. G. MacKay, 1982), and this reactivation rate is influenced in turn by the time characteristics of recovery from self-inhibition. This means that potential speech rate and time of peak hyperexcitability will be positively correlated, explaining D. G. MacKay's (1968) observation that the faster a subject's potential rate, the shorter the critical delay that produces maximal interference under DAAF. However, because factors influencing the time course of recovery from self-inhibition are central, automatic, and beyond voluntary control (see following), one would not expect the delay that produces maximal interference to shift when subjects *voluntarily* speak slower or prolong speech sounds, and no such shift is found (D. G. MacKay, 1968; 1970b).

The rate at which a node can be reactivated is also related to the level of priming that the node achieves, relative to all other nodes in its domain, and degree of satiation of the node directly influences this level of priming. Satiation

therefore explains the slower maximal speech rate for producing repeated versus connected syllables, as well as the decrease in repetition rate with increasing numbers of repetitions (see preceding discussion of D. G. MacKay & Burke, 1972).

By reducing the overall level of priming of a node, satiation also increases the overall probability of errors, without influencing either the hyperexcitability peak, or the delay that produces maximal interference. Some other factor therefore is needed to explain the fact (discussed earlier) that repeating a syllable under DAAF proceeds without errors and at a rate that does not vary as a function of delay.

Practice and the Peak Interference Delay

The recovery cycle hypothesis predicts a positive correlation between prior practice and the delay that produces maximal interference under DAAF. Recall from Chapter 8 that practice speeds up the activation and recovery cycle of individual nodes by strengthening the connection between the parent node and its inhibitory satellite. Onset of hyperexcitability will therefore vary with a node's history of prior activation, and hyperexcitability peaks much longer than 200 ms can be expected for unpracticed phonological nodes.

Practice-induced speed-ups in the activation and recovery cycle readily account for the longer delays required to produce peak interference in 2- to 6-year-old children (Figure 10.1) but seem to fly in the face of my own data for bilinguals (Figure 10.4). D. G. MacKay (1970b) reported that peak interference delay remained constant when German–English bilinguals produced their more and their less familiar language under DAAF. However, all of the native German speakers in this experiment had *at least* 5 years of intensive prior practice with English, and some had more than 20 years of prior practice. And because phonological nodes receive so much practice so quickly (D. G. MacKay, 1982), phonological nodes for these subjects must have already reached asymptotic levels of practice for both languages (especially in view of the extensive overlap between phonological components for German and English). And because phonological nodes provide the "first line of defense" against disruptive effects of feedback, no differences in peak interference delay for the more versus less familiar language could be expected in D. G. MacKay (1970b). However, the relationship between prior practice and peak interference delay warrants further test. The node structure theory predicts peak interference delays much longer than 200 ms for English speakers learning, say, Swahili, a language with phonological units very different from English.

Feedback-Induced Stuttering and the Recovery Cycle Hypothesis

What follows is another in a long history of attempts to provide an account of feedback-induced stuttering that integrates stutterers and nonstutterers. Under the recovery cycle hypothesis, the muscle movement nodes of stutterers display

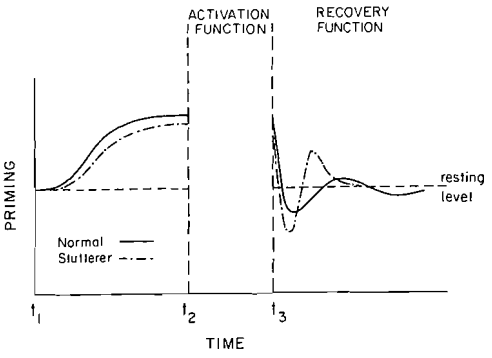

FIGURE 10.5. The priming function (relating degree of priming and time following onset of priming at t_1), activation function (following application of the triggering mechanism at t_2), and recovery function (following activation offset, t_3) for stutterers and nonstutterers. (Adapted from "Stuttering as a Sequencing and Timing Disorder" by D. G. MacKay and M. MacDonald in *Nature and Treatment of Stuttering: New Directions* (p. 273) edited by W. H. Perkins and R. Curlee, 1984, San Diego: College-Hill. Copyright 1984 by College Hill Press. Reprinted by permission.)

an abnormal recovery cycle, as illustrated in Figure 10.5. That is, rebound from self-inhibition comes earlier than normal and rises to a higher level of priming. As a result, just-activated nodes in their hyperexcitability phase may have greatest priming in their domain at the time when the next node is to be activated.

Heightened hyperexcitability alone could cause repetition errors, but amplifying the auditory feedback further increases the probability of stuttering. The delayed or attenuated stapedial contraction that has been observed in (feedback-induced) stutterers (Hall & Jerger, 1978; Horovitz, Johnson, & Pearlman, 1978) will likewise provide less of a defense against feedback-induced stuttering. Unattenuated external feedback will return with greater than normal amplitude to the just-activated mental nodes and further increase the probability of reactivation.

Mental nodes and the recovery cycle hypothesis explain why masking the returning auditory feedback reduces the probability of stuttering (see Baar & Carmel, 1970; Cherry & Sayers, 1956; and findings discussed below), and amplifying it has the opposite effect in feedback-induced stutterers (findings discussed below). Mental nodes also make sense of auditory induced fluency, the fact that appropriate auditory input can guide fluent production of a word, causing release from blocks and prolongations. Blocks are overcome when someone else utters the word on which a stutterer is blocking, because this input primes the appropriate or next-to-be-activated nodes, enabling these nodes to achieve greatest priming in their domain and become activated. Shadowing and choral rehearsal likewise prevent stuttering, because these auditory inputs augment priming for the appropriate or next-to-be-activated nodes.

One interesting prediction from the recovery cycle hypothesis is that the delay that produces maximal disruption with speech under DAAF will be related to potential rate, that is, to a stutterer's maximal articulatory rate without DAAF. The reasoning is already familiar. Potential rate is determined in part by the maximal rate at which nodes can be reactivated, which by hypothesis is

determined by the speed of recovery following activation of these nodes. Thus, a stutterer whose maximal rate of speech is relatively fast can be expected to display a shorter period of self-inhibition, with faster rebound and shorter peak interference delay under DAAF than a stutterer whose maximal rate of speech is relatively slow.

Another interesting prediction is that feedback-induced stutterers will react differently from intrinsic stutterers and normals to DAAF. Specifically, the delay that produces maximal interference with speech will be shorter for feedback-induced stutterers than for intrinsic stutterers and normals under the recovery cycle hypothesis. Further research is needed to test this prediction, because systematic comparisons of the effects of different delays on speech of normals, intrinsic stutterers; and feedback-induced stutterers' have never been undertaken. Previous DAAF studies have lumped intrinsic and feedback-induced stutterers together and either have used only a single feedback delay, have failed to amplify returning feedback, have omitted the normal control group, or have neglected basic controls for speech rate, distraction effects, order of the delays, and possible practice effects over repeated readings of the same materials (D. G. MacKay & MacDonald, 1984).

A Test of the Recovery Cycle Hypothesis

An experiment by D. G. MacKay and Birnbaum, reported for the first time here, incorporated all of the necessary control procedures just noted (for details, see D. G. MacKay, 1970b). The experiment examined effects of DAAF on three groups of subjects of about the same average age: feedback-induced stutterers who read a set of sentences more fluently when hearing white noise than when not hearing white noise (see Table 10.1), intrinsic stutterers who read as fluently or less fluently when hearing white noise than when not hearing white noise (see Table 10.1), and nonstutterers who, of course, read fluently under either noise or no-noise conditions. In the main experiment, the subjects read sentences as quickly as possible with auditory feedback amplified to about 95 dB under 10

TABLE 10.1. The time (in seconds) to read sentences (10 ± 1 syllables in length) under two conditions of undelayed feedback and two conditions of amplification by two groups of stutterers (see text for explanation).

Feedback conditions by stutterer group	White noise	No white noise
Feedback-induced stutterers		
Loud feedback	2.44	3.88
Soft feedback	2.43	3.01
Intrinsic stutterers		
Loud feedback	3.92	3.58
Soft feedback	3.20	3.24

different conditions of feedback delay (0.0 s, 0.01 s, 0.025 s, 0.075 s, 0.150 s, 0.200 s, 0.250 s, 0.400 s, 0.550 s, and 0.800 s). Effects of these delays were as follows. Peak interference delay (measured via either errors or syllable duration) always exceeded 0.01 s and differed among the three groups. The delay that produced maximal disruption was 0.2 s for the nonstutterers (as usual) and was consistently longer than for the *feedback-induced* stutterers (median 0.150 s) with no overlap in the distributions. This finding strongly supports the prediction of the recovery cycle hypothesis that peak interference delay is shorter for feedback-induced stutterers than for nonstutterers.

Results for the *intrinsic* stutterers were strikingly different. For this group, the delay that produced maximal interference was *longer* than for nonstutterers, that is 0.400 s for intrinsic stutterers versus 0.200 s for nonstutterers. This difference further corroborates the distinction between intrinsic versus feedback-induced stuttering and is reminiscent of the effects of DAAF in children. The peak interference delay is 0.4 s in 7- to 9-year-old children, and longer than in adults (0.2 s) (D. G. MacKay, 1968). One therefore wonders whether a development deficit may play a role in intrinsic stuttering, such that the age-linked shift in peak interference delay (and by hypothesis the faster activation and recovery cycle of underlying nodes) has failed to occur.