FOCUS OF ATTENTION IN THE HEARSAY-II SPEECH UNDERSTANDING SYSTEM[1]

Frederick Hayes-Roth
The RAND Corporation

Victor R. Lesser
University of Massachusetts

ABSTRACT

Using the concepts of stimulus and response frames of scheduled knowledge source instantiations, competition among alternative responses, goals, and the desirability of a knowledge source instantiation, a general attentional control mechanism is developed. This general focusing mechanism facilitates the experimental evaluation of a variety of specific attentional control policies (such as best-first, bottom-up, and top-down search strategies) and allows the modular addition of specialized heuristics for the speech understanding task. Empirical results demonstrate the effectiveness of the focusing principles, and possible directions for future research are considered.

INTRODUCTION

The Hearsay-II (HSII) speech understanding system (Lesser, et al., 1974; Erman & Lesser, 1975; Lesser & Erman, 1977) is a complex, distributed-logic processing system. Inputs to the system are temporal sequences of sets of acoustic segments and associated hypothesized phonetic labels. Diverse sorts of speech understanding knowledge are encoded in several (11, currently) independent knowledge source modules (KSs), including one or more KSs specific to each of the following knowledge domains: acoustic-phonetic mappings, phone expectation-realization relationships, syllable recognition, word hypothesization and verification, and syntax and semantics. The state of processing at any point in time is represented by a global data base (the blackboard) which holds in an integrated manner all of the current hypothesized elements, including alternative guesses, at the various levels of interpretation (e.g., segmental, syllabic, lexical, and phrasal). In addition, any inferred implicative or confirmatory relationships among various hypotheses are represented on the blackboard by weighted, directed links between associated hypotheses. The weight and direction of a link reflect the degree to which the hypothesis at the tail of the link supports (or confirms) the hypothesis at the head. The blackboard may be viewed as a two-dimensional problem space, where the time and information level of a blackboard hypothesis serve as its coordinates. Such a view permits consideration of specific "areas" of the problem space and enables us to speak meaningfully of hypotheses in the "vicinity" of a specific data pattern.

Processing in the system consists of additions, alterations, or deletions made to data on the blackboard by the various KSs. Each KS is data-directed, i.e., it monitors the blackboard for arrival of data matching its precondition pattern, a configuration of hypotheses and links with specific attribute values. Whenever its precondition is matched, a copy of the KS is instantiated (invoked) to operate separately on each satisfying data pattern. Finally, when the KS is executed, its (arbitrarily complex) logic is evaluated to determine how to modify the data base in the vicinity of the precondition pattern that triggered the invocation. The data pattern matching the precondition of a KS is called the stimulus frame (SF) of the invocation, and the changes it makes to the data base are referred to as its response frame (RF). Each KS may be schematized as a production rule of the form [precondition => response]. Each instantiation is then schematized [SF => RF], reflecting the fact that the RF data pattern is produced in response to the determination that the SF matches the rule's precondition. Because of the complexity of KS processing, a precise definition of the RF cannot be calculated directly from the SF without actually executing the KS. However, a non-procedural abstraction of each KS is used to estimate the RF directly from the SF. This abstraction specifies the type of changes that may be made (e.g., the addition of a new hypothesis or new link, the modification of a hypothesis's validity) and their location relative to the SF (i.e., time interval and level of interpretation). Subsequent discussions of RFs refer to the approximations derived from such abstractions.

As is well known in speech understanding research (Reddy, 1976), each KS is imperfect. At any level of analysis, a very large number of errors may be introduced, including misclassifications, failures to recognize, and inappropriate "don't care" responses to truly significant portions of the utterance. The common approach in speech understanding research is to construct systems that can recognize utterances in spite of such errors by evaluating simultaneously many weakly supported alternative interpretations of the speech. A practical consequence of this parallel evaluation of numerous alternatives is that, at any point in time, a great number of KS applications are warranted by the existence of hypothesized interpretations matching the various KS preconditions. One objective of attentional control is to schedule the numerous potential activities of the KSs to prevent the intractable combinatorial explosion that would inevitably result from an unconstrained application of KSs. More specifically, the focus of attention problem is to minimize the total number of KS executions (or total processing time) necessary to achieve an arbitrarily low rate of error in the semantic interpretation of utterances.

We believe that most of the issues relating to attentional control in Hearsay-II will also arise in other large-scale knowledge-based systems operating in errorful domains (where errors arise from imperfect KSs or incomplete or inaccurate data). Only the dimensions of the problem space (level of linguistic abstraction and time, in the speech domain) are likely to vary across problem domains. Many poorly understood problems apparently require the use of multiple, diverse sources of knowledge that can cooperate or compete in attempts to achieve a solution. As a result, the problems and approaches considered in this paper should be relevant to a wide variety of complex systems.

[1] This work was performed while both authors were employed by the Department of Computer Science, Carnegie-Mellon University. This research was supported by the Defense Advanced Research Projects Agency under contract no. F44620-73-C-0074 and monitored by the Air Force Office of Scientific Research.

The approach taken here is different in two ways from approachs taken in two other speech understanding systems (Woods, 1977; Paxton & Robinson, 1975). The first difference is that our approach does not rely upon explicit (pre-compiled) information about the type and characteristics of KSs currently contained in the system, and thus is more general; the second difference is that our approach does not guarantee that the first solution reached is optimal as in Woods, 1977. We did not choose an approach that would have this property because we feel that inorder to guarantee this property there is a severe cost in terms of time spent searching, and there is a restriction on the types of KS interaction permitted (i.e., in the speech-understanding domain, it requires the completion of all bottom-up processing to the word level before the search can begin).

FUNDAMENTAL PRINCIPLES AND MECHANISMS

One can view the focusing problem as a complex resource allocation problem. For example, consider the expenditure of money on alternative types of information useful in locating oil. The alternative information sources including seismologists, geologists, drilling teams, and satellite reconnaissance, are the KSs for the task. Each produces its response data only with significant cost and with a substantial probability of error, and there are sequencing constraints requiring some KSs to delay their processing until other KSs terminate and particular findings are obtained. How should one invest in their potential contributions? Five fundamental principles have been identified for the control of processing in such tasks, and these are listed below. Each of these principles is used to define a separate measure for evaluating the importance that should be attached to each KS invocation not yet executed. These measures associated with each KS invocation are not necessarily constant for the lifetime of the invocation but may need to be recalculated dynamically as the state of the blackboard changes in the general vicinity of KS's stimulus and response frames. A function based on these measures is then used to assign a priority to each KS instantiation.

(1) The competition principle: the best of several local alternatives should be performed first. This principle governs the ordering of several behavioral options which are competitive in the sense that some definite outcome of one obviates the others. For example, consider the problem of determining whether oil exists at site A and suppose that the functions of a geologist and seismologist are substitutable vis-a-vis this objective. If either the seismologist or geologist has already performed and positively indicated the presence or absence of oil, that result obviates employing the other scientist to perform an equivalent function. In this sense, it can be said that the previous result competes with the yet-to-be-performed alternative; that is, the former response is at a higher level of analysis in the same area of the problem space as is the alternative pending action. However, if oil on site B can be determined only by seismological techniques, hiring a geologist for site A does not compete with hiring a seismologist for site B, according to this principle. In the context of speech understanding, competition is exemplified by two alternative hypotheses at the same level of interpretation (e.g., two different word hypotheses) spanning overlapping time intervals. Two KSs proposed to operate upon these hypotheses are (locally) competitive.

(2) The validity principle: KSs operating on the most valid data should be executed first. This principle says that, everything else constant, one KS invocation should be preferred to another if the former is working on more credible data. Where the previous principle could be interpreted as a local best-first search strategy, this principle is tantamount to a global best-first execution of alternatives. For example, consider the case where preliminary seismological readings were taken at two different sites, and on the basis of these readings one of the sites was much more likely to contain oil. The validity principle dictates that further exploration should occur first at the site which is more promising based on the preliminary indications. Similarly, in the speech domain, various KSs will be instantiated to contribute to the interpretation of specific data patterns on the blackboard. Each hypothesis in a SF will contain a validity rating derived from the validities and implications of hypotheses linked to it. Thus, this principle implies that KSs invoked to work on the most valid SFs are most preferred. Once these KSs have performed, the hypotheses in their responses will also be rated for validity and will, in general, derive their validity directly from the hypotheses in the SF. By preferring KS invocations with the most credible SFs, the system maximizes the expected validity of its responses.

(3) The significance principle: those KSs whose RFs are most important should be executed first. This principle aims at insuring that when a variety of behaviors can be performed, the most important are done first. For example, while filing a claim on land and drilling are both necessary prerequisites for successful completion of an oil hunt, at the outset of prospecting the former is the more important and should be done first. As an example in the speech domain, a situation might arise where a sequence of phones could be either recognized as a word or subjected to analysis for coarticulation effects. The first of these two actions is more important and, on a priori grounds, should be performed first. One heuristic in the speech understanding domain for defining significance is to give preference to KS invocations operating at the highest levels of analysis within any portion of the utterance (closest to a complete interpretation). A more general statement of this heuristic is that preference should be given to the KS invocation whose RF can potentially produce a result which is closest (in terms of level of interpretation) to the overall goal of the problem solver.

(4) The efficiency principle: those KSs which perform most reliably and inexpensively should be executed first. Obviously, if one geologist is more reliable than another and the two charge the same for their services, the former should be preferred. Conversely, of two equally reliable geologists, one should prefer the less expensive. Similarly, in the speech domain, some KS applications are more efficient than others and should be preferred. As an example, a bottom-up word hypothesizer is found to be more accurate than a syntactic top-down hypothesizer at generating word hypotheses at the initial and final positions of the utterance. Everything else equal, when both of these KSs have been invoked to generate new word hypotheses, the bottom-up hypothesizer should be executed first.

(5) The goal satisfaction principle: those KSs whose responses are most likely to satisfy processing goals should

be executed first.[2] The oil hunt managers might establish a goal of determining the depth of water at site A. This would induce additional preference for those agents (e.g., the seismologists and drillers) whose ordinary activities could concomitantly satisfy this additional goal. In the speech domain, similar circumstances arise. As an example, a goal might be established to generate new word hypotheses in a particular time region of the utterance. A KS whose RF satisfies this goal is preferred. The desire for a specific type of processing is specified in HSII by establishing a goal on the blackboard representing the time and level of the desired hypotheses. KS instantiations whose RFs match the objectives specified in the goal are made more desirable. More generally, KS invocations may be evaluated as more or less likely to help satisfy each specific goal. The higher the probability that a KS invocation will contribute to the satisfaction of a goal and the greater the utility of the goal, the more desirable its execution becomes. Through this mechanism of adding goals to the blackboard, a focusing policy KS can introduce dynamic, task specific focusing rules into the scheduling algorithm. Because KS activity is data-directed, this KS (as all others) executes only when the data patterns indicating the need for a specific focusing action are detected.

In order to evaluate the preferability of one KS invocation vis-a-vis the others, the five control principles require a number of ordering relationships to hold. In overview, the major operational principle for focusing is to schedule for earliest execution the most desirable KS invocation according to the five rules provided. The focusing mechanism first evaluates the desirability of each KS invocation as a measure of the degree to which it satisfies the various objectives of the system and then executes the most desirable first (with an appropriate generalization for executing several KSs simultaneously in a multiprocessing system). Thus, the major subproblem in the construction of a focuser is the estimation of a KS invocation's desirability. How this desirability is computed will now be described.

Each KS invocation is characterized by a number of attributes. Its SF has a credibility value (between -100 and +100) estimating the likelihood that the detected pattern of hypotheses and links is valid and satisfies the KS's precondition (negative values imply evidence against this possibility). The credibility value of a SF is determined as a function of the validity ratings on each of the hypotheses in the SF. As previously indicated, these ratings themselves are determined from the strengths of implications on links, the original probabilities assigned to each of the acoustic segment labels provided as input (i.e., the lowest-level hypotheses on the blackboard), and the derived validity ratings of intermediate level hypotheses. In our current implementation, the credibility of the SF of a KS instantiation is taken to be the average of the validity ratings (ranging from -100 to +100) of the hypotheses in the SF weighted by the number of syllables in each. (The number of syllables seems to be a good measure of the information content of speech hypotheses).[3]

Each KS invocation can be viewed as a transformation of the SF into the RF. Associated with the KS invocation then is the estimated level(s) (e.g., phonetic, lexical, phrasal), the estimated validity, and the estimated time (i.e., location and duration) of the potential RF hypotheses. Each of these estimated values contributes to an appraisal of the significance and probable correctness of the RF to be produced by the KS.

The objectives of the validity, significance and efficiency principles can be achieved if the desirability of a KS invocation is computed by an increasing function of the credibility of its SF, the estimated level, duration, and validity of RF hypotheses, and the estimated reliability of the KS (to produce correct RFs of the form it anticipates). The objective of the validity principle, to operate on the most valid data first, is accomplished by making desirability an increasing function of the credibility of the SF. The objective of the significance principle, to perform the most significant behaviors first, is achieved by making desirability an increasing function of the level and duration of RF hypotheses. Since hypotheses closest to complete interpretations will be at the highest level and span the entire duration of the utterance, actions expected to produce or support such hypotheses will be most preferred. The objective of the efficiency principle, to prefer KSs which perform best, is achieved by making desirability an increasing function of the amount of uncertainty a KS reduces per unit "cost" (execution time).

These objectives are implemented by computing the expected value of the RF as follows: Each KS module provides the system with non-procedural specifications of its expected cost-effectiveness (goodness) and a priori significance (priority) as well as the temporal vicinity, duration, and level of its RF relative to the associated SF. Given these specifications, the expected.RF.value for a KS instantiation is determined by the following formula.

$$\text{expected.RF.value} = \text{SF.validity} * ((1 + \text{duration} * \text{duration.weight})/(1+\text{duration.weight})) * \text{goodness} * (\text{level}/\text{max.level})$$

where:

SF.validity is the credibility of the SF (the average validity of the SF hypotheses weighted by their durations in syllables);

duration is the expected duration in syllables of the RF, computed from the SF and the non-procedural specifications;

goodness is the estimated cost-effectiveness of the KS (between 0 and 1);

level is the expected level of the RF computed from the SF and the non-procedural action specifications (where the segmental level, the lowest, is 1 and the phrasal level, the highest, is 10);

max.level is the highest *possible level (10)*;

and duration.weight is an empirically tuned control parameter.[4]

[2] At first glance, the significance and goal satisfaction principles appear similar. However, the key difference between them is that the former reflects a static (a priori) view of importance while the latter reflects a dynamic view based on the current state of processing.

[3] The scheduling of precondition processes is integrated into the same framework encompassing KSs. The credibility of a precondition is taken to be the maximum validity rating of all hypotheses belonging to its monitored set of relevant new or changed hypotheses.

[4] As explained below, two distinct conceptions of local

To understand how the other objectives, deferring obviated behaviors (competition) and goal-directed scheduling, are achieved in the system, it is necessary to introduce a number of additional concepts. The mechanisms required to operationalize the desired effects of competition will be considered first.

The first objective of the focuser is to insure that the understanding system moves quickly to a complete interpretation of a spoken utterance and, in particular, avoids apparently unnecessary computation. Specifically, if any KS invocation is expected to produce a RF in the same time interval as an existing, higher-level, longer (in duration), and more credible hypothesis, its activity is potentially useless. It is therefore less preferred than the action of a KS expected to produce higher-level, more extensive, and more credible interpretations of the utterance than currently exist. In order to compute this preference, HSII uses a statistic called the state of the blackboard; this is a single-valued function of each time value, from the beginning of an utterance to its end. The state S(t) for some point (time) t in the utterance is, roughly, the maximum of the values V(h) of all hypotheses representing interpretations containing the point t. The value V(h) of an hypothesis h is an increasing function of its level, duration, and validity and, in particular, is computed by the same formula used for the value of an RF. Thus, the highest possible value for an hypothesis would be that associated with the hypothesis representing a complete parse of the entire utterance with a validity rating of +100 (the maximum). To the extent that the utterance is partially interpreted in some interval [t1, t2], the state S(t) will be high in this region. Thus, S(t) provides a single metric for evaluating the current success of the understanding process over each area of the utterance. From a more general viewpoint, the metric V(h) indicates how close a hypothesis h is to the desired overall goal state, the metric S measures both what aspect of the overall goal has been solved (e.g., in the case of speech, what time interval) and how good the solution is (e.g., in the case of speech, the validity of the hypothesis and how close in terms of level of abstraction it is to the sentential phrase).

S(t) can be employed to decide whether a prospective action is likely to improve the current state of understanding. If the estimated value V(h) of a RF hypothesis h exceeds S(t) anywhere in the corresponding interval, the KS invocation should be considered very desirable; otherwise it should be inhibited by the existing, more valuable, competitive hypotheses. This is how the objective of the competition principle is accomplished. In addition to its dependence upon the variables already considered, the desirability of a KS invocation is made to be an increasing function of the ratio of the estimated value of the RF to the current state S(t) (where S(t) is taken to be the minimum over the interval corresponding to the time location of the RF). In this way, preference is given to KS invocations expected to improve the current state of understanding.

One can think of S(t) as defining a surface whose height reflects the degree of problem solution in each area. In this conception, operations yielding results below the surface are undesirable (unnecessary) and those raising the surface most are preferred. The relative desirabilities of various actions are illustrated in Figure 1.

Two distinct notions of competition have been explored in our research. They correspond to simple implementation variants of the current state function S(t), but accomplish somewhat different objectives. In phrase-specific competition, S(t) represents the maximum value of all phrasal hypotheses spanning time t. Because each phrase hypothesis corresponds to a grammatical sequence of words and is at a higher level than any single word, the highest valued phrase is likely to represent the best hypothesis yet generated. Thus, inhibiting KS instantiations whose RFs are lower-valued than the best phrase hypothesis in their areas is tantamount to pursuing a depth-first search for an interpretation. That is, once any sequence of words is found that is valued significantly higher than all current phrase hypotheses, the focus of attention algorithm will prefer to attempt to extend this phrase than to extend other shorter, lower-rated alternatives in the same area. As long as word hypotheses can be generated to extend this hypothesis, the value of the new and longer sequences will tend to increase, and lower-valued pending KS instantiations will be increasingly inhibited. Thus, phrase-specific competition encourages parallel searches only in non-overlapping time regions. As a consequence, phrase-specific competition should be expected to produce some correct interpretations very quickly but, on the other hand, it may cause many correct sentence interpretations to be missed before the available time or space is exhausted. How desirable such depth-first processing is will depend completely on the performance characteristics of the KSs available. Related empirical results are reported below.

The second type of competition that has been explored is word-specific. In this case, S(t) represents the highest value of any word hypothesis (spanning time t) that is incorporated into any grammatical sequence. Thus, an expected RF is better (or worse) than S(t) to the extent that it incorporates words that are better (or worse) than those already supporting some phrasal hypotheses. The result of this implementation of S(t) is to explore more combinations of words into phrases, unless some phrase can be found that essentially incorporates all of the best available word hypotheses. Because the relative validity ratings of words are far from perfect (e.g., a correct hypothesis receives, on average, the fifth best rating in a 1000-word vocabulary), word-specific state values should be more conservative and robust. Usually, this implementation will produce a breadth-first search of alternatives. Only when many of the correct words are rated highest does it produce the fast and direct search behavior that phrase-specific competition attempts.

The last objective to be operationalized is that of the goal satisfaction principle. In general, a goal may specify that particular types of hypotheses are to be created (e.g., create word hypotheses between times t_0 and t_1) or existing hypotheses modified in desired ways (e.g., attempt to reject the hypothesized word "no" between t_3 and t_4 by establishing disconfirming relationships between it and the acoustic data). Two types of adjustments are made to the desirability ratings of KS invocations based on their relationships to such goals. The first case arises when there is direct goal satisfaction, meaning that a KS invocation is a possible candidate for solving a goal because its RF matches the desired attributes of the goal. In this case, the

competition were explored in our work, word-specific and phrase-specific competition. The respective values of the control parameter for each of these cases will be detailed in the next section when the complete formula for KS desirability is specified.

desirability of the KS invocation is increased by an amount proportional to the utility of the goal (the importance associated with the goal when it is created).

The second type of effect is the result of indirect goal satisfaction. In this case, a KS invocation does not directly satisfy a goal but apparently increases the probability that it will be satisfied by producing some partial result useful for the achievement of the main goal. Two types of indirect goal-satisfying actions can be identified. First, there is goal reduction: a KS invocation generates subgoals whose solution(s) will entail satisfaction of the original goal. For example, as the result of recognizing the sequence "The (gap) dog," the system might establish a goal for the recognition of an adjective between the two recognized words to replace the gap in understanding. Subsequently, some KS might establish several disjunctive subgoals related to this one, such as goals for recognizing the words "shaggy," "cute," "sleepy," etc. Because the satisfaction of any one of these would constitute satisfaction of the original objective, the KS invocation indirectly satisfies the original goal. Its desirability is less than that of a KS invocation directly satisfying the same goal but may be more than that of other KSs.

The second type of indirect goal satisfaction occurs when a KS invocation approaches a goal by producing a RF which is close to the goal but does not quite satisfy it. For example, in the context of the preceding "adjective" goal, a general increase in the activity of knowledge sources which generate and improve phone hypotheses, syllable hypotheses, and phrasal hypotheses in the area of interest will be more or less proximate to the desired response. Since each KS is schematized as a rule of the form [precondition => response], a means-ends analysis can be performed to estimate the probability that some KS invocation will produce a response contributing to the ultimate solution of a goal. The more closely its RF approaches the desired goal, the higher is the probability that execution of a KS invocation will contribute to the goal's ultimate satisfaction and the greater the desirability of the KS instantiation.[5]

In summary, the desirability of a KS invocation is defined to be an increasing function of the following variables: the estimated value of its RF (an increasing function of the reliability of the KS and the estimated level, duration, and credibility of the hypotheses to be created or supported); the ratio of the estimated RF value to the minimum current state in the time region of the RF; and the probability that the KS invocation will directly satisfy or indirectly contribute to the satisfaction of a goal and the utility of the potentially satisfied goal. Scheduling KS invocations according to their desirabilities therefore accomplishes the objectives established by the preceding five basic principles. However, there are some inadequacies of such a basic attentional control mechanism; these are considered in the next section. Subsequently, the complete desirability formula is presented.

ADDITIONAL MECHANISMS FOR PRECISE FOCUSING

Basically, while the five fundamental principles appear correct and universally applicable, they do not provide the precise focus control necessary to handle certain issues. Three additional issues are now introduced, and the control mechanisms used to handle these are discussed. The topics considered include dynamically modifiable recognition and output generation thresholds on KS logic; an implicit goal state (approximately the inverse of the current state S(t)) that can be used to determine the desired balance between depth-first and breadth-first approaches to the understanding problem; and methods for avoiding "false peaks" or "cognitive fixedness" in the recognition process.

Nearly all KS behavior can be separated into two components: a pattern recognition component and an output generation component. For example, a word hypothesizer may look for patterns of phones (pattern recognition) in order to produce a new word hypothesis (output generation). Both components operate in fuzzy, errorful ways. In the pattern recognition component, the KS must accept partial (incomplete) and inexact matches of its templates, because that is the nature of speech recognition. Conversely, the word hypotheses it generates are necessarily probabilistic. The probable correctness of its hypotheses are reflected by validity ratings or implication weights on its outputs. Thresholding occurs in such processes in two ways. First, the degree of fuzziness tolerated in pattern matching is arbitrarily set to some moderate criterion to prevent an intractably large number of apparent matches. Second, the strengths of the output responses are measured against some threshold to insure that only sufficiently credible responses are produced. The credibility of the response may depend not only upon the credibility of the stimulus frame but also upon the type of inference method used to generate a response. For example, the word recognizer might employ a distance metric for recognition and classification, in which case the credibility of the output word is a decreasing function of the distance between the stimulus phones and the phones of the most similar word template. Responses which are too weak vis-a-vis this second threshold are held in abeyance rather than being produced or forgotten.

Now the general scheme of the robust overall policy that is employed can be sketched. At the beginning of an analysis, relatively high thresholds are specified for pattern matching goodness and output goodness. Processing continues based on the other scheduling principles until thresholds are changed (discussed below). When a threshold change occurs, it may be specific to certain levels or time regions of RFs or to the types of KSs used to produce them. As an example, if all of the utterance were correctly interpreted except the first word, we would set very low thresholds for behavior for all KSs in the beginning portion of the utterance. Our current policy, in specific, successively lowers bottom-up word hypothesization thresholds in areas where only poorly rated words have been hypothesized until either "enough good" hypotheses or "too many" hypotheses are generated in each time interval. Specifically, thresholds are lowered in each area until either all words in the fourth-highest equivalently validity-rated set of words have been hypothesized or more than 20 words are generated bottom-up. No other dynamic thresholding is currently performed.

Where dynamically modifiable pattern match and output

[5] This design for the goal satisfaction principle was not completely implemented, and the empirical results to be presented later are not based on a system that used explicit goals as a focusing mechanism. Instead a different mechanism, called threshold control was used to implement speech-specific control strategies. This threshold control mechanism will be discussed in the next section.

goodness thresholds are not used, the KSs necessarily embody numerous parameters whose values are determined at the outset for all problem tasks. For this reason, these KSs are probably very sensitive to the particular values chosen. Unfortunately, as a practical issue, not all KSs were developed incorporating the logical decomposition necessary to accomplish dynamic thresholding. As a general approach, however, dynamic thresholding insures that each of the KSs can be encouraged to perform more work in any area of the blackboard by simply lowering two general sorts of control variables. This is seen as a fundamentally important control principle relating to the controllability of the generative aspect of KSs *per se* rather than to their comparative expected responses.

The second additional concept embodied in the focuser is the *implicit goal state* I(t). It is only a slight oversimplification to think of I(t) as the arithmetic inverse of the current state S(t). To the extent that S(t) is large (representing that the portion of the utterance adjacent to t has been highly successfully analyzed), I(t) will be small. A small I(t) value means that there is little to be gained by trying to improve the understanding around t. Conversely, a large I(t) means that the portion of the utterance in the neighborhood of t greatly needs additional analysis. As a result, one might suppose that KSs operating in that region should be conceived as satisfying an implicit goal of raising the level of understanding (the surface of the current state S(t)) wherever it is lowest. In fact, the best role for the implicit goal state is probably as a weak contributor to the desirability of a KS invocation. It remains an empirical question whether it is better to work in the regions of the highest peaks in S(t) (*depth-first*) or more evenly throughout the entire utterance (*breadth-first*). Although an optimal strategy is not known, it is clear that in computing the desirability of a KS invocation, the estimated value of the RF and the ratio of the RF value to the minimum of S(t) in the same region are two contributing factors whose relative weightings can be experimentally manipulated to achieve exactly the desired balance between depth-first and breadth-first.

As is well known in problem solving and search paradigms, there is a constant danger of getting trapped on "false peaks," as when one bases actions on the apparent correctness of highly rated but ultimately incorrect interpretations. A number of the preceding focusing principles have been formulated to insure that processing in the region of highly valued hypotheses is facilitated at the expense of other potential actions. A consequence of this paradigm is that the focuser must take precautions to prevent the "cognitive fixedness" resulting from a failure to abandon dead end paths. Such precaution is achieved in a simple manner. The highest peak in understanding at any point t in the utterance corresponds to the highest-valued hypothesis in that region, and its value is just S(t). Thus, stagnation of the understanding process in a region can be detected whenever S(t) fails to increase for a prolonged time. This is implemented by periodically updating I(t) whenever the highest-valued hypothesis supporting S(t) has not been superseded by a higher-valued hypothesis (in both word- and phrase-specific cases) or, in the case of word-specific competition only, whenever the highest-rated word hypothesis has not been incorporated into a new higher-valued phrase. While preference should still be given to the execution of KS invocations working on the surface of S(t) and promising to increase its value, the focuser must conclude that other KS invocations should now become more desirable than they previously seemed, because they at least may improve the analysis in the stagnant area. This is accomplished by increasing the implicit goal state I(t) whenever S(t) is stagnant for a specified length of time. As a result of increasing I(t), KS invocations operating near the surface of S(t) and previously viewed as marginally desirable become sufficiently desirable to be executed. If any one of them succeeds in increasing S(t), I(t) is promptly reset to be the inverse of S(t). However, each time S(t) stagnates for the specified duration, I(t) is again increased. The effect of stagnation of S(t) over the execution of n instantiations is to increase I(t) by $4 * n * (n-1)$. This formula is designed to penalize S(t) geometrically as stagnation persists. Thus, false peaks are avoided by actually recognizing the behavioral characteristics of cognitive fixedness: as long as the degree of its understanding remains stagnant, the desirability of the competing KS alternatives, which previously appeared to be suboptimal in the area of stagnation, is continually increased. This modification of I(t) due to stagnation continues until a complete sentence is found that spans the entire utterance and has a minimally acceptable credibility. After this event, I(t) is not permitted to exceed its current value. If a new, complete, more highly rated parse is found, the current I(t) becomes the new upper bound.

It is now possible to specify the complete desirability calculation used by the scheduler. The desirability of a KS instantiation is calculated in terms of (1) instantiation characteristics (i.e., its invocation priority, SF.validity, and expected.RF.value), (2) the current state of processing (i.e., S(t) and I(t)), and (3) empirically tuned parameters (i.e., RF.mult, RF.to.S.mult, goal.reduction.mult, and goal.mult). These parameters control the relative priorities given to the validity, significance, competition and goal satisfaction factors. In order to improve comprehensibility, some boundary-value conditions requiring special treatment have been omitted from the formulae:

```
    desirability  ←  RF.factor  +  ratio.RF.to.S  +
          amount.goal.reduction  +  goal.factor  +
          invocation.priority;
where:
    RF.validity ← (if S(t) is word-specific then SF.validity
          else expected.RF.value);
    current.state ← min value of S(t), appropriately
          reduced by stagnation, over time interval of
          RF;
    distance.to.goal ← max possible value of S(t) -
          current.state;
    implicit.goal ← max value of I(t) over time interval of
          RF;
    RF.factor ← expected.RF.value[6] * RF.mult / 100;
```

[6] The control parameter duration.weight used to compute the expected.RF.value is set to 1 and 0.5 in the word-specific and phrase-specific schemes, respectively. These settings are somewhat suprising since the word-specific scheme, which supposedly is a more breadth-first strategy than the phrase-specific scheme, uses a higher weighting for duration. However, in the calculation of desirability in the word-specific case the only use of duration is in the expected.RF.value, because RF.validity and S(t) are not functions of duration in this case. Thus, the higher

```
ratio.RF.to.S ← (RF.validity / current.state) *
        RF.to.S.mult;

amount.goal.reduction ← ((RF.validity - current.state) /
        distance.to.goal) * goal.reduction.mult;

goal.factor ← implicit.goal * goal.mult;
```

The empirically tuned multipliers RF.mult, RF.to.S.mult, goal.reduction.mult, and goal.mult are set to 4, 10, 10, 5 and 4, 3, 15, 2, respectively, in the word-specific and phrase-specific schemes. The different settings for these multipliers help to normalize the values of RF.validity and S(t) which in the word-specific case are generally much larger and at the same time closer together than in the phrase-specific case. In both schemes, the setting are designed to give most importance to the RF.factor which is a measure of significance, less importance to the ratio.RF.to.S which is a measure of competition, and lowest and approximately equal importance to the amount.goal.reduction and goal.factor which are measures of goal satisfaction.

ALTERNATIVE POLICIES FOR FOCUS OF ATTENTION

Up to this point, general principles for focusing and mechanisms to achieve the realization of these principles have been described. However, a wide variety of policies can be superimposed upon these mechanisms to effect various, specific, global search strategies for speech understanding. This flexibility is considered one of the outstanding virtues of the focuser design since it affords the possibility for empirical evaluation of alternative focus of attention policies. Each policy can be implemented by one or more specific policy modules, a KS-like program that is activated whenever specific conditions of interest are detected. For example, consider the policy dictating that, whenever possible, understanding is to proceed bottom-up, from the acoustic segments to the phrasal level. Such a policy would be effected as follows: at the outset, the policy module would set a goal with infinite positive utility for RFs at the lowest level and a goal with infinite negative utility for RFs at higher levels. When the system became quiescent[7], the policy module would be reinvoked by the system. Its response would be to modify the goals so that processing at the two lowest levels would be facilitated and all others inhibited. This process would continue until the highest level was facilitated. Similarily, a purely top-down analysis could be controlled in the same way, substituting "highest" for "lowest," etc.

In summary, it is suggested that the principles and mechanisms described in the preceding sections provide a parameterized framework for the elaboration of numerous alternative "macroscopic" policies for attentional control in the speech understanding problem. Each of the typical sorts of heuristic problem solving policies can be realized by simple policy modules that manipulate goal utilities and thresholds and respond to quiescence and stagnation in policy-specific ways. Another means of flexible scheduling, which has not been explored, is to make the parameters of the priority evaluation function (which is used by the main scheduler) modifiable by policy KSs. Thus, a policy KS could influence strategy by adjusting the importance attached to particular kinds of KS actions (e.g., creating new word hypotheses bottom-up or top-down) -- this would affect the order of execution of invoked KSs, in addition to the influence that policy KSs now exert by controlling the invocation of KSs.

THE HEARSAY-II FOCUSING POLICY AND RESULTS

The empirical results presented later in this section were obtained while Hearsay-II operated under a specific focusing policy, defined as follows[8]. First, all segmental hypotheses are generated from the parametric representation of the acoustic signal. Next all grammatically feasible sentence-initial and sentence-final words are predicted top-down, and possible interior position words are predicted bottom-up based on stressed syllable hypotheses constructed from segmental information. These predicted words are then rated, and the most likely words in each time interval are placed on the blackboard. Up to this point of processing, control of activity is implemented using thresholds; further processing, however, is strictly based on the desirability calculations of the scheduler. Next, a heuristic word-sequence hypothesizer attempts to identify the most probable sequences of word hypotheses (consisting of successive language-adjacent word pairs). Because this KS exploits statistical methods to improve credibility, the initial word sequence hypotheses are much more accurate than are hypotheses based on single words. Subsequently, KSs are invoked to attempt to parse the hypothesized word sequences to determine if they are grammatical, to predict possible time-adjacent grammatical word extensions, to hypothesize and verify new words satisfying these goals, to concatenate grammatical and time-adjacent word sequences, to reject phrases and words, and to generate new word sequence hypotheses.

Termination of processing is accomplished by elimination of all pending KS instantiations or by exceeding a fixed amount of processing time or space (Hayes-Roth, Lesser, Mostow & Erman, 1976). Once any sentence is recognized that completely spans the utterance, hypotheses that are apparently erroneous are rejected or deactivated. An hypothesis is rejected if its validity is so low that any phrase that could incorporate it and would span the entire utterance by combining the best available word hypotheses at all other points in time would yield a lower value than the best available spanning sentence hypothesis. All KS instantiations whose SFs contain a rejected hypothesis are deleted from the system. An hypothesis is deactivated unless its validity is greater than the validity of the corresponding

weighting for duration is just a way of slightly biasing the desirability calculation for a longer duration RF whereas in the phrase-specific case duration plays an important role in each of the factors contributing to the desirability calculation.

[7] One of the primitive events that can be used to trigger the execution of a precondition process is quiescence, defined as the state in which the desirabilities of all KS instantiations fall below a specified threshold. Another type of event that is monitored for is stagnation, which occurs when the state function S(t) has not been modified during a specified number of KS executions.

[8] See Lesser & Erman 1977 for a more detailed description of the Hearsay-II knowledge source configuration and appropriate references.

temporal interval of the best available spanning sentence hypothesis. Any pending KS instantiation whose SF contains only deactivated hypotheses is also eliminated. Whenever a more valid overall sentence hypothesis is generated, hypotheses rendered weakly (locally) improbable are deactivated, strongly (globally) improbable hypotheses are rejected, and associated pending actions are eliminated.

A significant amount of tuning of the focussing parameters has been attempted. Nevertheless, the current parameter values are probably not optimal, and it seems clearly impossible to determine what the optimal values are. In addition, owing to the interesting relationships between the desirability of breadth- or depth-first searches and the specific performance characteristics of the particular KSs used in the system, no absolute conclusions are warranted. Only the general focusing problem and our suggested general approaches appear universally valid; statements regarding the validity of particular parameter settings must await major breakthroughs in the development of our mathematical models and analytical techniques.

Our focus of attention mechanisms were evaluated by executing Hearsay-II with three different scheduling algorithms over the same 61 test utterances spoken by one male speaker[9]. The three algorithms used are listed below:

A (alternating): All instantiated preconditions are executed, then all pending KS instantiations are executed in decreasing order of desirability, and this program is repeated until the termination condition is satisfied.

W (word-specific competition): Preconditions and KS instantiations are all scheduled according to desirability, the most desirable being executed first. The desirability of a precondition is taken to be the maximum expected desirability of all of its potential KS instantiations. Competition is based on a current state function representing the highest-valued word hypotheses that support any phrasal hypotheses. Stagnation in a time region reflects the time elapsed since the word supporting the state function in that region was last incorporated into a longer and more valid phrasal hypothesis. This best-first scheduling algorithm continues until the termination condition is satisfied.

P (phrase-specific competition): This algorithm is identical to algorithm W except that competition is based on a current state function representing the highest-valued phrase hypotheses spanning each time area. Stagnation in a time region reflects the time elapsed since the state function value last increased in that region.

The termination conditions were identical for all conditions. Whenever a more valid complete sentence was understood, improbable hypotheses and pending instantiations were deactivated or rejected. The system was halted in any case whenever the time limit of 200 million instructions (including output generation) or the storage limit of 200k (about 40k for dynamic storage of hypotheses, links, and arrays) on the PDP-KL10 was exceeded. When the system quiesced or was halted, the highest-rated hypothesized sentence was chosen as the utterance interpretation. Thus, three outcomes could occur: (1) every word in the sentence was correctly identified; (2) although some words were incorrectly recognized, (e.g., article was recognized as articles), the semantic interpretation of the entire sentence was correct; or (3) the sentence was incorrectly understood.

While algorithm A can be viewed as a control condition against which to measure the effects of focusing, it will perform significantly better than a completely unfocused system just because whenever a correct sentence is found, it is usually produced by the first executed KS instantiation in the last complete cycle. At that point, most of the other pending KS instantiations are eliminated as a result of hypothesis deactivation and rejection. For our purposes, the interesting comparisons are between algorithms W and P, both of which are expected to produce significantly better performance than A. As previously explained, we expected P to produce more of a depth-first search of the solution space. Thus, to the extent that bottom-up word hypotheses are valid and reliably ranked (correct hypotheses rated higher than incorrect ones), P should rapidly home-in on correct interpretations and suppress competing actions. Conversely, to the extent that these conditions do not arise, the depth-first nature of P should entail many long and fruitless searches being performed before the system (essentially backs up and) continues execution of the stagnating but correct KS instantiations. Thus, in comparison to W, P is fast but risky. While the conservative competition policy W should produce consistently moderate search times, P is expected to produce a more bi-modal distribution resulting from many fast and many very slow searches.

The results from the 61 test sentences were as follows. First, as expected, A resulted in a significantly lower rate of correct understanding than either W or P. While the overall rates of correct semantic interpretation for A, W, and P were .62, .77, and .75, it is interesting to compare the three algorithms where their differences are most detectable. The corresponding rates for exact recognition were .52, .70, and .66. Considering only those (48) sentences not immediately recognized bottom-up (i.e., excluding all sentences where focusing strategy played no part), the rates of exact (word-for-word correct) recognition were .52, .71, and 69. The differences between A and W or P are statistically reliable (one-tailed Binomial homogeneity tests, $p < .05$), but the difference between W and P is negligible. In addition to its poor recognition performance, the control focusing algorithm A ran considerably slower than (more than twice the time of) both W and P. As a result, algorithm A will not be discussed further.

The remainder of our analyses are based on the 33 sentences that were not recognized immediately bottom-up but were correctly understood using either W or P. Six different but necessarily correlated measures of the relative effectiveness of the algorithms were considered and are reported in Table 1. All measures show a statistically reliable superiority of W over P. The average magnitude of this superiority ((P-W)/W) is about 20 percent, and the largest effect is shown in the average CPU time to hypothesize the correct sentence.

[9] The grammar used in this evaluation was more complex than that usually used for system testing. By thus increasing the size of the search space, we hoped to differentiate the alternative strategies as much as possible.

In sum, competition is an effective scheduling principle. Furthermore, the effectiveness of the competition principle is very sensitive to two varying realizations based on the alternatives of small grain size (competing words in the same region) or large grain size (competing phrases in the same area) in the computation of the current state function. It appears that the efficacy of the smaller grain size is due to the pressure it exerts for a conservative, limited breadth-first search. Alternatively, competition based on larger units of evaluation is undesirable because it too often results in long depth-first searches that, while persistently promising success, are fruitless.

CONCLUSIONS

By schematizing knowledge sources as [precondition => response] rules, each potential behavior of the Hearsay-II system is viewed as an instantiation of such a form. These KS instantiations are seen to be [stimulus frame => response frame] action descriptions. The desirability of an instantiation is then computable from several characteristics of the stimulus and response frames. Based on these principles for attentional control, a desirability measure is produced that accomplishes most of the focusing objectives. Several elaborations of this simple strategy are desirable. For more precise control, computations are made of the current state of the analysis, the implicit goal state of the system, and the relative degree of goal satisfaction of each KS invocation. Once the desirability of each KS invocation is computed, the execution of the most desirable first serves to accomplish an apparently optimal allocation of computing resources. In addition, our framework provides an excellent environment for empirically evaluating the utility of various global focusing strategies. Each of these can be expressed in terms of particular weightings of the contributions of various terms to the desirability of a KS invocation or by simple modules to create, modify, and monitor goals controlling the direction of analysis. The relatively small grain size of knowledge representation and fine identification of the type and location of knowledge source contributions apparently affords great advantages in experimenting with mechanisms to control a large, distributed, knowledge-based understanding system.

Analysis of our results indicates that large cost reductions can be obtained by straightforward realization of the proposed focusing principles, particularly if a moderate grain size (the level of word hypotheses) is chosen as a basis for implementing the notions of current state, competition, and stagnation.

ACKNOWLEDGMENTS

We would like to acknowledge the help of the following people in the design and implementation of these ideas in the HSII system: Jack Mostow, Craig Everhart, Don Kosy and David McKeown.

REFERENCES

Erman, L. D. and Lesser, V. R. A multi-level organization for problem solving using many diverse cooperating sources of knowledge. Proc. 4th Inter. Joint Conf. on Artificial Intelligence, Tbilisi, USSR, 1975, 483-490.

Hayes-Roth, F., Lesser, V. R., Mostow, D. E. & Erman, L. D. Policies for rating hypotheses, halting and selecting a solution. In Speech Understanding Systems: summary of results of the five-year research effort. Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pa., 1976.

Hayes-Roth, F., & Lesser, V. R. Focus of attention in a distributed-logic speech understanding system. Proc. 1976 IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, Philadelphia, Pa., 1976.

Lesser, V. R., Fennel, R. D., Erman, L. D., & Reddy, D. R. Organization of the HEARSAY II speech understanding system. IEEE Trans. on Acoustics, Speech, and Signal Processing, 1975, ASSP-23, 11-23.

Lesser, V. R. and Erman, L. D. A retrospective view of the Hearsay-II architecture. Proc. 5th Inter. Joint Conf. on Artificial Intelligence, Boston, 1977.

Paxton, W. H., & Robinson, A. E. System integration and control in a speech understanding system. A. I. Center, Tech. Note 111, SRI, Menlo Park, Ca. 1975.

Reddy, R. Speech recognition by machine: A review. Proc. of the IEEE vol. 64, no. 4, April 1976.

Woods, W.A. Shortfall and density scoring strategies for speech understanding control. Proc. 5th Inter. Joint Conf. on Artificial Intelligence, Boston, 1977.

FIG. I. Response Frame Desirability as a Function of Current State

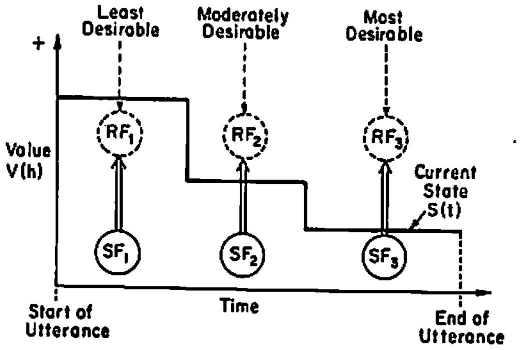

Table I. Comparison of Costs under Algorithms P and W

Cost Measure	When[1] measured	W	P	P-W	$\frac{P-W}{W}$	$t_{(32)}(P-W)$[2]
Total phrases hypothesized	H	55.48	64.27	8.79	.16	2.10***
	T	69.52	80.85	11.33	.16	2.24***
Total KS Instantiations	H	63.15	76.21	13.06	.21	1.67*
	T	105.88	127.21	21.33	.20	2.21***
Total CPU Time (Secs.)	H	45.09	58.03	12.94	.29	1.97**
	T	87.06	103.30	16.24	.19	2.01**

Notes:

1. The costs were measured either as soon as the correct sentence phrase was hypothesized (H) or when processing was terminated (T).

2. One-tailed matched-pairs t-tests, 32 degrees of freedom, with significance levels indicated as follows:

   * $p < .1$
   ** $p < .05$
   *** $p < .025$