10.

ALTERNATIVE CONTROL STRUCTURES FOR SPEECH UNDERSTANDING SYSTEMS

Gary Goodman
Raj Reddy
Carnegie-Mellon University
Pittsburgh, PA 15213

10-1. ABSTRACT

Control Structures are an essential part of any speech recognition system. They are the devices by which passive knowledge about the task and language is transformed into active and effective processes. In this chapter we define and discuss three areas of control structures: knowledge source interaction, knowledge source activation, and knowledge source focusing. Discussion relates the concepts presented to systems developed during the five-year ARPA speech understanding project.

10-2. INTRODUCTION

Speech understanding systems are characterized by high data rates, diverse sources of knowledge representing large numbers of rules and facts, incomplete and inaccurate knowledge, and error and uncertainty in individual decisions. We know that all the available sources of knowledge must communicate and cooperate in the presence of error and uncertainty. We do not know how to do it effectively or efficiently. The problem of control in a speech understanding system refers to how knowledge is organized, activated, and focused to constrain the search. In this chapter we show how error leads to search, how knowledge constrains search, and how decisions about activation and focusing of knowledge affects the computational complexity of the recognition process. The control strategy, i.e. activation and focusing of knowledge, used by a speech understanding system thus seriously affects the speed and accuracy of the recognition process.

In Section 10-3 we discuss the problem and paradigms of knowledge source interaction, i.e., how knowledge sources communicate with each other. In Section 10-4 we show how the problem of error can be viewed as a search problem and discuss various search techniques useful to speech understanding systems. In Section 10-5 we show how various systems cope with the problem of focusing, that is, how they decide which of the many competing requests for knowledge source activation should be satisfied.

10-3. KNOWLEDGE SOURCE INTERACTION

Various forms of knowledge must be applied if a speech understanding system is to be effective in deducing the intended message from the speech signal. Further, these sources of knowledge must be able to cooperate with one another. The kinds of knowledge employed and the ways in which they interact are part of the many design decisions affecting the structure and control of speech understanding systems.

10-3.1 Knowledge Sources

The distinctive characteristic of speech understanding systems is the active use of knowledge of the language, the environment and the context in understanding an utterance. These sources of knowledge (KSs) include the characteristics of speech sounds (phonetics), variability in pronunciations (phonology), the stress and intonation patterns of speech (prosodics), the sound patterns of words (lexicon), the grammatical structure of language (syntax), the meaning of words and sentences (semantics), and the context of the conversation (pragmatics). Part II of this volume covers the definition and use of these types of knowledge. The following discussion refers to the levels of representation shown in Figure 10-1. These correspond, roughly, to a hierarchical structure of the forms of knowledge.

10-3.2 Models for Knowledge Source Interaction

A model of knowledge source interaction presumes some information on which the knowledge is to act. Knowledge sources speak "different languages" in the sense that they deal with diverse areas of knowledge. However, some common representation is necessary if knowledge sources are to interact cooperatively. The application of knowledge may affect this representation in essentially two different ways. Knowledge may be used to alter the representation within a level. This may happen at the lexical level when the occurrence of a word (e.g., author's first name) is used to predict an adjacent word (e.g., author's last name). This use of intra-level knowledge usually occurs at the phonetic level because of coarticulation. Another form of application occurs when the knowledge, as it relates to some level, functions to alter or infer the representation at another level. An example of this is when syntactic knowledge is used to infer <author> at the phrasal level after identifying the sequence of first-name and last-name at the lexical level. The direction of knowledge flow is not necessarily from a lower level to a higher level. The opposite case arises when syntactic knowledge is able to "strongly" predict a missing word from an otherwise complete sentence.

The types of knowledge employed, the communication paths, and their directions form the primitives for models of knowledge source interaction. The remainder of this section describes several general models in this framework. Interaction models specific to the ARPA recognition systems may be found in the excellent review by Klatt (1977).

Hierarchical model: The model, shown in Fig. 10-2, is the most straightforward model of interaction. It is completely data-driven with all communication paths going from level to level, bottom to top. Interpretations of the data at any level are available only to the next higher level. The direct, one-way KS interconnections greatly simplify control for this model. The model is limited, however, because an error in interpretation at some level propagates to the next level resulting in compounding of errors and an error of omission is difficult, if not impossible, to correct. Consider, for instance, one word of an otherwise complete sentence missing at the lexical level. Syntactic or other

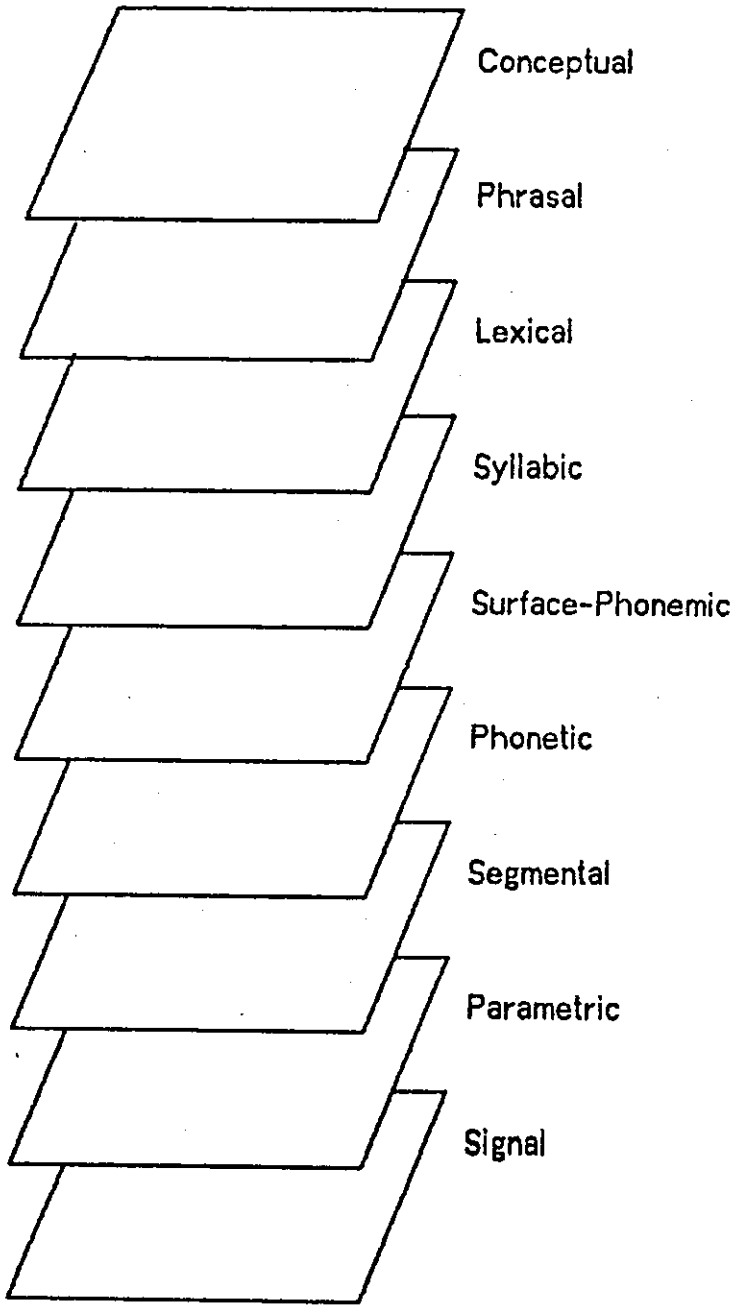

Figure 10-1

Levels of Representation for Speech Understanding Systems

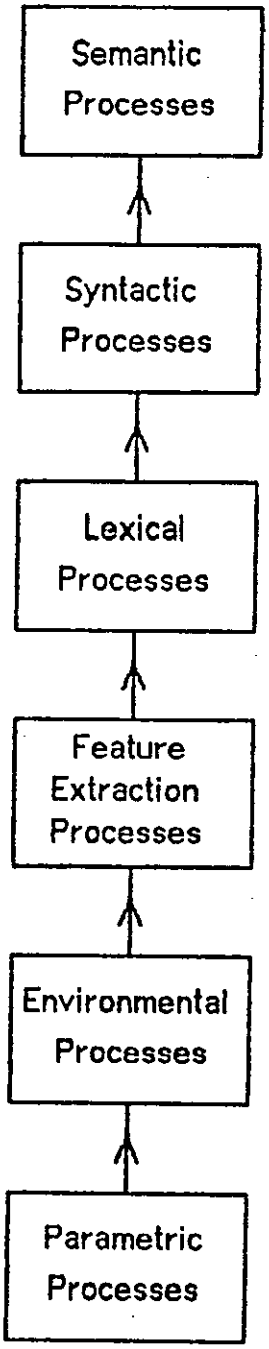

Figure 10-2

The Hierarchical Model

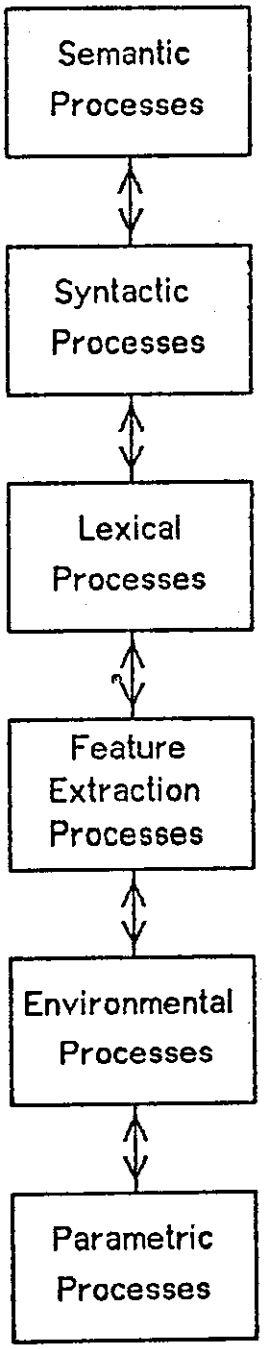

Figure 10-3
The Goal-Directed Model

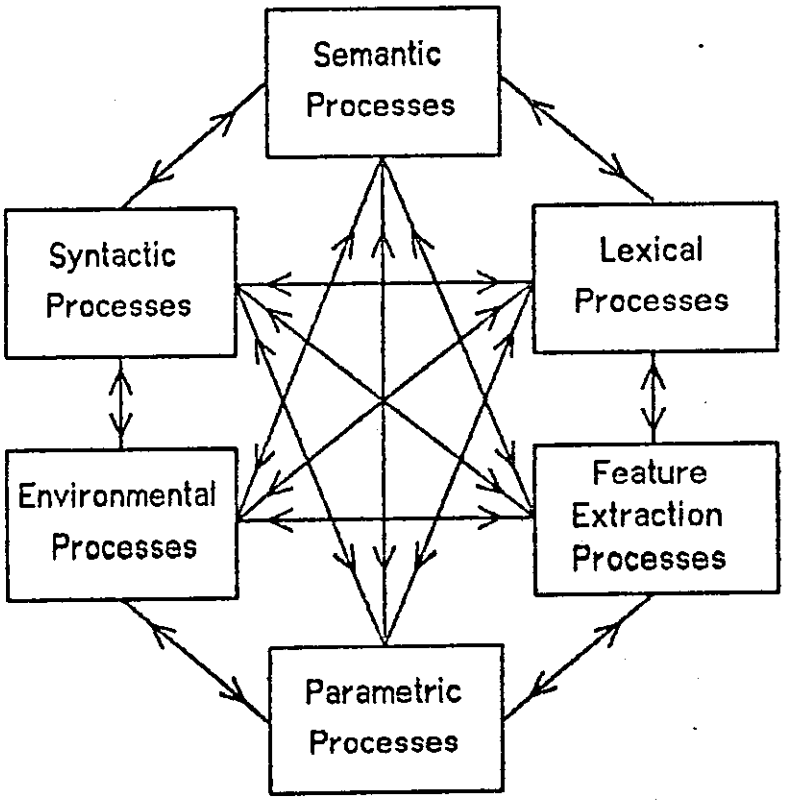

Figure 10-4
The Heterarchical Model

knowledge at the phrasal level must blindly assume the missing word was spoken in order for the sentence to have an interpretation at the sentential level. A KS has no opportunity, in this model, to request other knowledge sources to verify the existence of that word. This model, because of its simplicity, would be the correct one to use if no uncertainties existed in the interpretation (i.e., no errors) or if all uncertainty could be eliminated by application of knowledge. Since this is not the case in current speech recognition systems, this model is not used in its pure form. Some systems, however, do use this model for their front-end, transforming the signal, hierarchically, to some intermediate level. Uncertainties remaining at that level must be disambiguated by higher-level knowledge.

Goal-directed model: This model, shown in Fig. 10-3, is also known as the top-down, generative, or predictive model. Communication begins at the highest level with each knowledge source predicting at the next lower level. This continues until the signal level is reached, at which time the prediction may be either confirmed, denied, or more typically, given some score representing an estimate of the credibility that the prediction is true. Although the primary mode of interaction is higher level to lower level, the decisions, or scores, are reported back to the higher level. This method suffers from the fact that the search space starting at the top level is quite large for any reasonably sized, habitable grammar. The simple nature of the interaction, in the presence of error, limits the effectiveness of these first two models.

Heterarchical model: The heterarchical model, shown in Fig. 10-4, abandons simple interaction by allowing any knowledge source to interact with any other. One pays a price for this in the increased complexity of the representation and search. If there are k knowledge sources, then $k*(k-1)/2$ separate data paths exist where k-1 previously existed. Each new data path requires another common representation and increases the complexity of the knowledge source. Also, the search becomes more complicated to control since it may proceed in many different ways.

Blackboard model: Control of the search process may be simplified by having each knowledge source communicate through a central data base having one form of representation. This is known as the blackboard model, Fig. 10-5. Each knowledge source is an independent entity which examines the data base and after doing so may evaluate hypotheses created by other knowledge sources or create its own hypotheses. Even with this simplification, control of search is difficult for this model. This model was used successfully in the Hearsay-II (HS-II) system developed at Carnegie-Mellon University. More detail concerning this system may be found in Erman & Lesser (this volume).

Locus model: In the locus model (Fig. 10-6), all syntactic, lexical, and word juncture knowledge has been precompiled into an integrated network representing a complete description of every pronunciation of every possible sentence. The input signal is hierarchically transformed into a segmented and phonetically labeled form which is matched against the network to yield an optimal network path. The locus model of search uses a graph-searching technique in which all except a "beam" of near-miss alternatives around the best path are pruned from the search tree at each segmental decision point, thus containing the exponential growth without requiring backtracking. Control is greatly simplified because of a single uniform representation of all the different sources of knowledge.

The Harpy speech understanding system (Lowerre, 1976) is an example of the locus model. It was the first connected speech system to satisfy the original specifications given in the Newell report (Newell, et al., 1971)

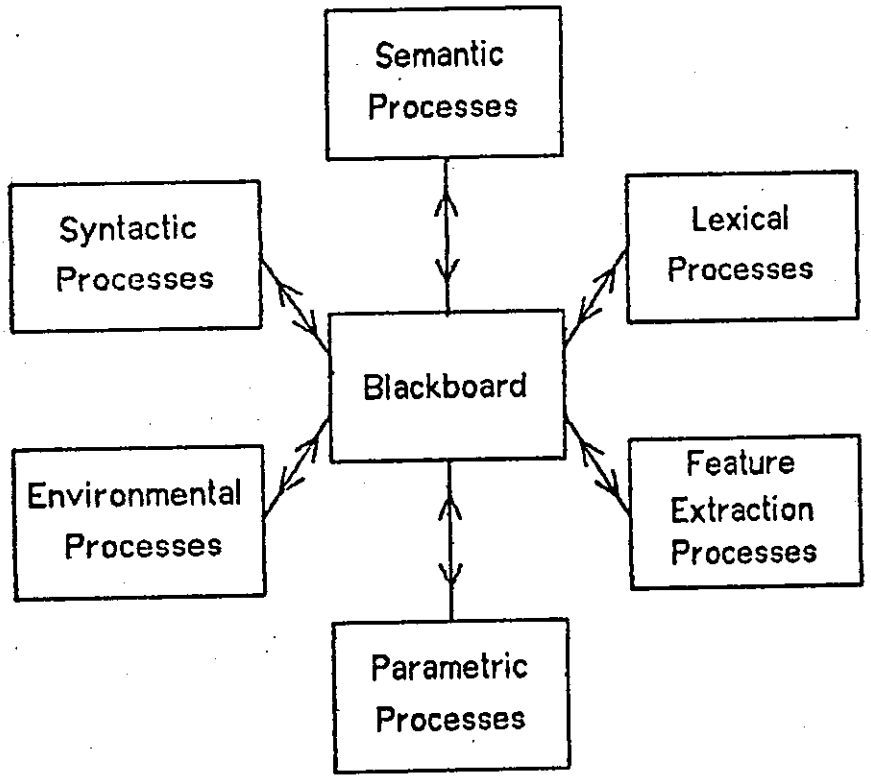

Figure 10-5
The Blackboard Model

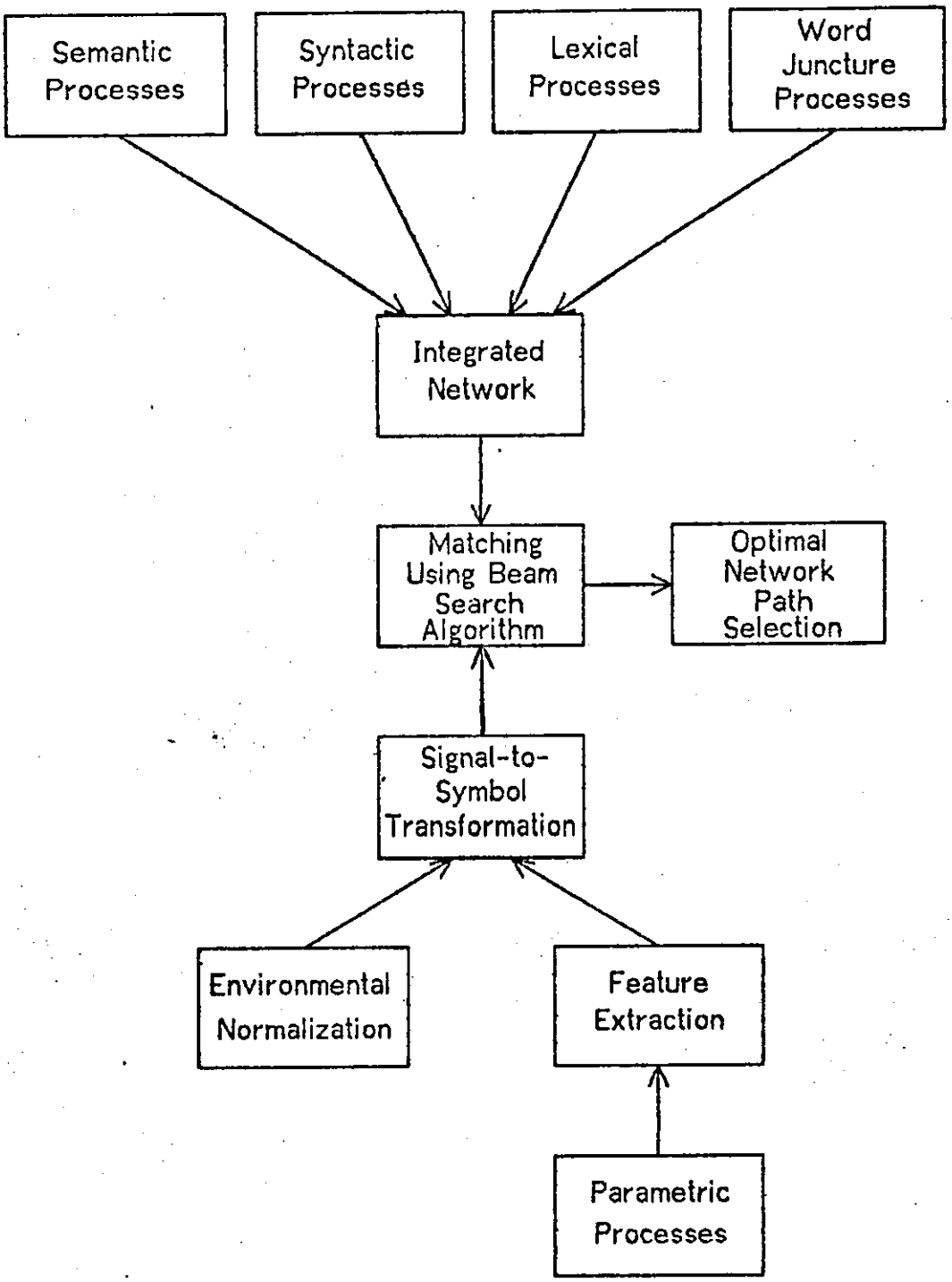

Figure 10-6 The Locus Model

and was first demonstrated in September of 1976. A complete description and discussion of this system may be found in Lowerre & Reddy (this volume).

10-4. KNOWLEDGE SOURCE ACTIVATION

The goal of a speech understanding system is to find the most plausible interpretation consistent with knowledge at every level, including the input utterance (signal level). The role of acoustic-phonetic knowledge is to propose plausible interpretations of the signal using pattern classification techniques (Part II, this volume; Itakura, 1975; Makhoul, 1975; and Schafer & Rabiner, 1975). Because these techniques are not capable of making unique choices with perfect accuracy, several interpretations must be considered plausible. Figure 10-7 shows a speech waveform with multiple interpretations at the phone level. If acoustic-phonetic knowledge were perfect, only one alternative would exist for each segment, there would be only one interpretation of the utterance at the phone level, and recognition could proceed in a straightforward manner.

Pattern matching techniques utilize distance functions to decide on the plausibility of a particular choice. The result is a value which represents relative plausibility and can be used to order the alternatives. In the absence of other knowledge the most plausible interpretation would be formed by selecting the best of each set of alternatives. Since the techniques are less than perfect this generally leads to nonsense such as /ah r m aw t/. Higher level knowledge must be utilized to resolve the ambiguity. The goal is to find the most plausible sequence of candidates which is also consistent with the higher level knowledge. It would be possible to evaluate the path likelihood of every possible sequence, as in the Dragon recognition system (Baker, 1975). Since this can be very time consuming, some form of pruning strategy is usually desirable. The knowledge applicable at each level serves to constrain the search by considering only those sequences which are consistent with the knowledge. In the example of Fig. 10-7, lexical knowledge has constrained the last word to be "but", "out", or "about". The word "mutt" does not appear because it is not in the language and therefore inconsistent with lexical knowledge.

Here, we will briefly describe search mechanisms for activating knowledge. For clarification and more detail see Nilsson (1971) and Winston (1977).

10-4.1 Basic Search Mechanisms

During a recognition, partial interpretations are built in an attempt to find the most plausible one. The ways in which knowledge sources are activated are governed by the search strategies employed. As alternative interpretations are generated, the choice of the next alternative to attend to must be made. In a depth-first search, the most recent alternative is the one chosen. In a speech understanding system this means examining the implications of some hypothesis at every level before attending to other hypotheses. In the example of Fig. 10-7, suppose segmental alternatives were ordered (/o/, /ah/, /aw/). Segment /aw/ would be chosen to be explored next. This would lead to the generation of /aw/ at the phone level. Then lexical knowledge would be activated. If, in this instance, there are no words starting with /aw/ which also start a sentence of the language, the lexical search would generate no new

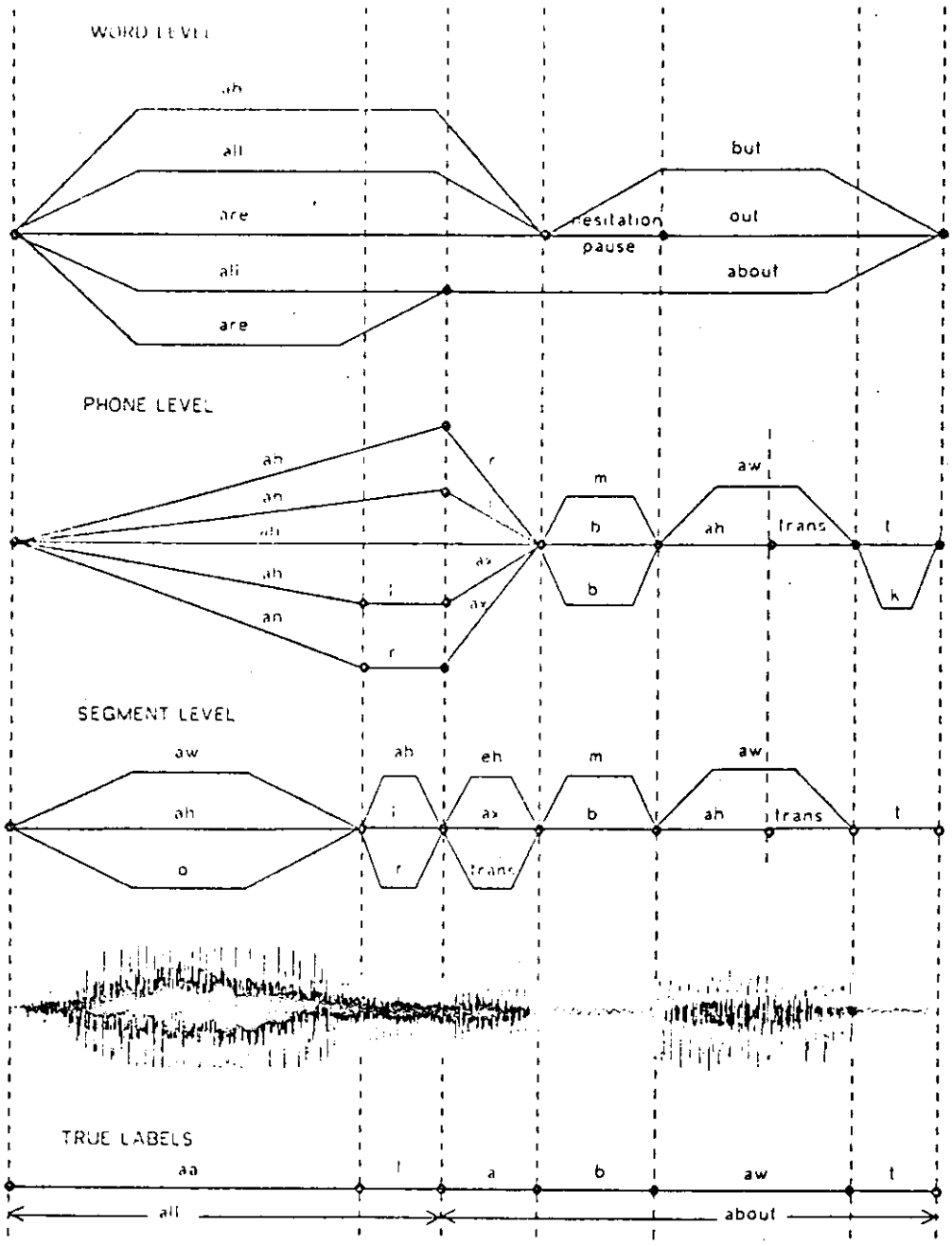

Figure 10-7

Example of Alternative Interpretations at Various Levels (From Reddy and Erman, 1975)

hypotheses. The depth-first premise would then select /ah/ at the segment level as the most recent, unattended alternative. This hypothesis would generate /ah/ at the phone level and "all" and "are" at the word level. Depth-first search is risky in the sense that the search may waste much time following a partial interpretation which is eventually found to be inconsistent with some knowledge.

Breadth-first is another form of basic search. This method of search examines all options at any level before proceeding to another level. Alternatives are explored uniformly so that the true path receives as much attention as any other path. In order not to overlook the correct path, many alternative labelings must be kept at every level, leading to combinatorial explosion.

10-4.2 Probabilistic Search

Search efficiency can be greatly improved if pattern matching knowledge is used to order the alternatives so that the most promising are explored first. If the selected alternatives are pursued in a depth-first manner, this is called best-first search. Suppose, in the example of Fig. 10-7, that the scores resulting from a signal-to-symbol transformation for (/aw/, /ah/, and /o/) are (.7, .9, .05). Implications of /ah/ at the word level would be explored before those for /aw/ or /o/ were considered. The words generated would be likewise ordered so that the most promising would be pursued. The best-few method expands breadth-first, evaluates the new hypotheses, and continues to explore only the most promising ones in a breadth-first fashion. In the example above /o/ would be discarded as unlikely with /aw/ and /ah/ being explored further.

The effectiveness of these methods depends highly on the accuracy with which the matching procedure orders the alternatives. Current signal-to-symbol matching yields 50-70% first choice accuracy (Klatt, 1977). Given this, the method works well only when moderate lexical and syntactic constraint is available.

10-5. KNOWLEDGE SOURCE FOCUSING

The search methods discussed in the previous section were concerned with examining alternatives at various levels. However, the search space may be regarded as having 3 dimensions: level, alternative, and time. In a global sense, each partial interpretation is a tree structure made up of alternatives at various levels and covering some portion of the utterance. Focus of attention mechanisms are control structures for deciding which of the competitive partial interpretations should be extended. A simple focusing strategy, used by Harpy, is to perform a best-few search moving left-to-right in time. Of the many ways in which search may proceed, there are two popular viewpoints: left-to-right and island driving (Erman & Lesser, this volume; Woods, 1977). This latter form of search begins by establishing anchor points which are portions of the utterance where the credibility of a word or sequence of segments is very high. These small partial interpretations are then extended in a best-first manner. The idea is to build interpretations covering larger time periods until an interpretation is found which covers the entire utterance.

Proponents of island driving argue that extending the globally best interpretation is more efficient since it approaches the recognition goal in the obvious, direct manner. Further, accuracy is better because the method does not consider portions of the utterance with low credibility until they become possible extensions of the current best interpretation, whereas a left-to-right strategy is forced to deal with unpromising portions as they occur in the utterance. If this happens at the beginning of an utterance, a left-to-right strategy may consume a great amount of time examining interpretations which look good initially, but cannot be completed.

Proponents of left-to-right strategy argue that it is much simplier, requiring far less bookkeeping, and thus leads to greater efficiency. Also, this method can achieve the same accuracy by using a best-few search which explores more alternatives during portions of the utterances where credibility is low.

Speech systems developed during the ARPA project used some form of probabilistically guided search. Hearsay-I (Reddy, Erman, & Neely, 1973), and the speech systems developed at Lincoln Labs (Forgie, et al., 1974) and SDC (Ritea, 1975) used a best-first left-to-right search with backtracking. Hearsay-II and HWIM (Woods, et al., 1976) also employed best-first search but with an island driving strategy. Harpy (Lowerre, this volume) moves left to right with a best-few search, called "beam search", which requires no backtracking.

Control strategies for speech systems which use island driving may be classified as explicit or distributed. Explicit strategies, after deciding which interpretation to extend, call a predefined sequence of knowledge sources to extend and rate the new interpretation. Distributed strategies are necessary when knowledge sources are independently activated.

An example of an explicit focusing strategy is the Shortfall Scoring Method used in the HWIM system (Woods, 1977). This method assigns priorities to partial interpretations, called islands, by comparing the actual score obtained for an island with the maximum attainable score for the time period covered by the island. An island is a sequence of words which is part of a legal sentence. The maximum attainable score is computed by summing the best scores for all the sub-units (words) in the associated time period. Each island is assigned a priority equal to the actual score for the island plus the maximum attainable score(s) for the region(s) not covered by the island. Refering to the diagram of Fig. 10-8, Priority(Island) = Actual Score(Island) + Max-attainable(Region1) + Max-attainable(Region3). This represents an optimistic estimate of an island's final score. The algorithm extends the island having the highest priority. This priority scheme is interesting because it guarantees "admissibility"; i.e., it guarantees the discovery of the best matching interpretation of the utterance. Note also, that the method will work not only for island-driven strategies, but for left-to-right and right-to-left strategies as well.

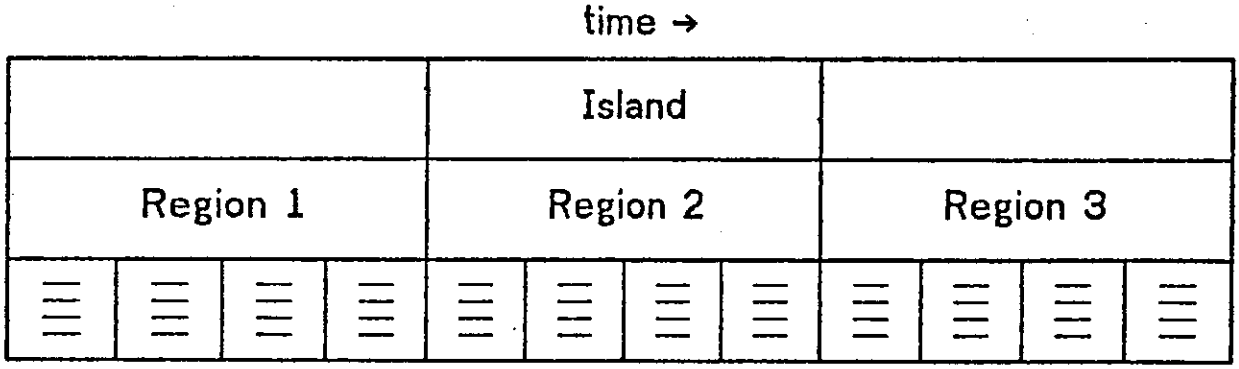

Figure 10-8

Example of Regions Used in Shortfall Scoring

A distributed focus mechanism was used in the Hearsay-II system (Hayes-Roth & Lesser, 1976). In this system, knowledge sources are data-directed and invoked whenever certain preconditions are observed in the data base. The control problem is to execute first those knowledge sources which are more likely to lead to a successful recognition. The approach was to lay down several principles for control: (1) The Competition Principle: the best of several alternatives should be explored first; (2) The Validity Principle: more processing should be given to knowledge sources operating on more valid data; (3) The Significance Principle: more processing should be given to knowledge sources whose expected results are more significant; (4) The Efficiency Principle: more processing should be given to knowledge sources that perform more reliably and inexpensively; (5) The Goal Satisfaction Principle: more processing should be given to knowledge sources whose responses are more likely to satisfy processing goals. Knowledge source priorities were based upon the degree to which they satisfied these principles. By tuning various weighting factors, a desirable balance between breadth- and depth-first search was achieved.

10-6. DISCUSSION

The complexity of speech understanding requires the use of many diverse sources of knowledge cooperating to achieve a solution. Ambiguity inherent in the speech signal necessitates search and the computational complexity of the search demands that it be carefully controlled. The designer of a speech understanding system has to make many decisions which affect the nature of control. He would like to know which kinds of knowledge should be used, how knowledge sources should interact, and how they should be activated and focused. In this chapter we have discussed different solutions to these problems. Obviously, we have not specified the best combination of design choices. In fact, there may be no single best set of choices. Each task probably requires a different combination of control decisions. I.e., if the task is simple, such as the recognition of digit sequences, a left-to-right best-first strategy may be adequate. If it is as complex as the "unrestricted English" task, some form of blackboard model with island driving and best-few strategy might be necessary.

One might suggest that the appropriate action is to build the best system possible using the most advanced forms of knowledge application. This is a worthy goal, but a flexible general system capable of handling unrestricted English will, in general, require too much space, too much time, and is not likely to be cost effective for simple tasks. Thus, the choice of optimal control strategies is affected by various aspects of the task; such as, connected speech versus isolated words or phrases, the degree of semantic and syntactic constraint, vocabulary size, and the degree of phonetic similarity of lexical items.

10-7. ACKNOWLEDGEMENTS

We would like to thank Jack Mostow and B. Yegnanarayana for their helpful comments on this manuscript and John Zsarnay for his help in generating the illustrations.