Shortfall and Density Scoring Strategies for Speech Understanding Control

W. A. woods
Bolt Beranek and Newman Inc.
Cambridge, Ma. 02138

Abstract

This note describes two methods of assigning priority scores to partially developed hypotheses about a speech utterance for determining which hypotheses to extend further. These methods guarantee the discovery of the best matching interpretation of the utterance, when used in an appropriate control framework. Although presented in the speech context, the algorithms are applicable to a general class of optimization and heuristic search problems. The density method is especially interesting since it is not an instance of the general A* algorithm of Hart, Nilsson, and Raphael, and appears to be superior to it in the domains in which it is applicable. Proofs of the guaranteed discovery of the best interpretation and some empir ical comparisons of the methods are given.

1. Introduction

This paper is concerned with control strategies governing the formation and refinement of partial hypotheses about the identity of an utterance in a continuous speech understanding system. We assume a system that contains the following components:

o) A Lexical ketrieva1 component that can find the k best matching words starting or ending at any given point in the utterance for any number k, and can be recalied to continue enumerating word matches in decreasing order of goodness at a given Position. We assume that this component is inter faced to appropriate signal processing, acoustic-phonetic and phonological analysis components as in (woods et al., 1976), and that it assiigns a "quality" score to each word match reflecting the goodness of the match.

b) A Linguistic component that, given any sequence of words, can determine whether that sequence can be parsed as a possible initial, final, or internal subsequence of a syntactically correct and semantically and pragmatically appropriate utterance, and can propose compatible classes of words at each end of such a sequence.

The HWIM speech understanding system developed at BBN [Woods et al., 1976; Wolf and Woods, 1977] has such capabilities. A control strategy for such a system must answer questions such as:

a) At which points in the utterance to call the Lexical Retrieval component, and when,

b) What number of words to ask for,

c) When to give subsequences of the results to the Linguistic component, and

d) When to recall the Lexical Retrieval component to continue enumerating words at a given point.

The goal of the control strategy is to d iscoyer__the__best scor inc;__sequence__of words that covers the entlre utterance and Ts acceptable_to "the LinguTstTc component. We will consider here a particular class of control strategies which we refer to as "island-dr iven".

2* Island-Driven Strateg ies

In an island-driven control strategy, partial hypotheses about the possible identity of the utterance are formed around initial "seed words somewhere in the utterance and are grown into larger and larger "island hypotheses by the addition of words to one or the other end of the island. Occasionally, two islands may "collide" by proposing and discovering the same word in the gap between them and may be combined into a single larger island. Each such island hypothesis is evaluated by the Lexical Retrieval component to determine its degree of match with the acoustic evidence and checked for syntactic, semantic, and pragmatic consistency by the Linguistic component. we will refer to a partial hypothesis that has been so evaluated and checked for consistency as a "theory". The strategies that we will consider operate by successively processing "events" on an event queue, where events correspond to suspended or dormant processes that may result in the creation of theories.

The general algorithm operates as follows:

(1) An initial scan of the utterance is performed by the Lexical Retrieval component to discover the n best matching words anywhere in the utterance according to some criterion of "best" and for some value n. An initial seed event is created for each such word and placed on the event queue. In addition, one or more continuation events, which can be

processed to continue the enumeration of successively lower scoring words, are created and placed on the queue. Each seed event is assigned a priority score (derived from the guality score" that the Lexical Retrieval component gave it in one of several ways to be described shortly), and each continuation event is assigned a priority score that can be guaranteed to bound the priority score of any word that can be generated by that event (e.g., derived from the score of the last word enumerated prior to the continuation). The events are ordered on the event queue by their priority scores and are processed in order of priority.

(2) The highest priority event is selected for processing, which consists of (i) creating the corresponding theory (a one-word theory in the case of a seed event), (ii) calling the Linguistic component to check the consistency of the theory and to make predictions for words and/or word classes that can occur adjacent to it, (iii) calling the Lexical Retrieval component to enumerate the k best matching words satisfying the predictions at each end of the theory, and (iv) generating a "word" event for each such word found. A word event is an event that will add one word to a theory to create a larger theory. Continuation events are also created that will continue the enumeration of successively lower scoring words adjacent to the theory. If island-collision is permitted as an operation, then each word event generated is checked against an island table to see if the same word (at the same position in the input) has been proposed and found in the other direction by some theory, and if so, an "island-collision" event is created that will combine the new word and the two theories on either side of it. Both word and island-collision events are assigned priority scores derived from the quality score of the new word and the scores of the theories to which it is being added and are inserted into the event queue according to their priorities.

(3) Continue selecting the top priority event from the event queue (step 2) until a theory is discovered that spans the entire utterance and is syntactically, semantically, and pragmatically acceptable as a complete sentence.

The main topic in this paper is the assignment of priority scores to the events in the above algorithm in order to guarantee that the first complete theory found will be the best scoring one that can be found. Using the quality scores assigned by the Lexical Retrieval component directly as priority scores does not ordinarily provide such a guarantee.

[3] • l£*£ §J2°_Lt££il £c°_li!21 Method

3.a) Assumptions

The shortfall method assumes that the quality scores assigned to word matches by the Lexical Matching component are additive, so that theories are appropriately assigned scores that are the sums of the scores of the word matches contained in them. It also assumes that word matches have associated beginning and ending positions that correspond to boundary positions in the input utterance. In the HWIM system, the quality scores are logarithms of estimates of the relative probabilities of the correctness of a theory given the acoustic evidence.

3.b) The Basic Shortfall ScojLiDJi. ProcecTure

Let t(i) be the time in milliseconds of the i-th boundary in the utterance; nsegs, the number of segments in the utterance; and seg(i) be the region of the input utterance from t(i-l) to t(i) for i from 1 to nsegs.

For a word match from position i to j with score q, we will allocate in some systematic way the total word score q to the segments seg(i-H) ... seg(j) covered by the word match. For this discussion, let us allocate it proportional to the durations of the segments.

For a given utterance, we will determine for each segment seg(i) the maximum score max(i) that can be allocated to that segment by any word match that covers the segment. The score for any word match from i to j will hence be bounded by the sum max(i+l)+ ... +max(j), and the maximum score for any complete theory will be bounded by T the sum from 1 to nsegs of max(i).

Every partial theory will consist of a contiguous sequence of word matches spanning a region from some boundary i to some boundary j. Each such theory will carry with it two scores m and q, where m is the sum of the max(i) for the segments covered by the sequence and q is the sum of the word scores of the theory. We will assign each theory a priority score p * T - m + q, which can be thought of as the maximum total score T for any theory minus the shortfal1 from this ideal to which one is committed by choosing this particular sequence of words (i.e., p * T-(m-q)). Alternatively, it can be thought of as the estimated best possible future score consisting of the score q which has already been achieved for the region covered plus the best potential score T-m for the region not yet covered

(i.e., p = qMT-m)). Because T-m is an upper bound on the possible score that can be achieved on the region not covered, the priority scores p have the characteristic that they are non-increasing as theories grow.

In the shortfall scoring strategy, the priority scores of the individual seed events are simply the shortfall scores of the words. A priority score for a continuation event that will be an upper bound on the priority score of any words that might result from the continuation can be computed as follows: Since the Lexical Retrieval component enumerates words in decreasing order of score, the quality score of any word that results from the continuation will be no greater than that of the last word enumerated so far. Moreover, we can derive from the lexicon 3 lower bound on the length of a word and from this we can deduce the shortest region of the utterance that such a word could cover, and hence the smallest possible m score that such a word could have. From these two numbers, we can bound the priority score (T-m+q) of any future word and use that as the priority score of the continuation event. (This bound is somewhat conservative, and in actual practice, it should be possible to derive a much tighter bound, but this argument is sufficient to guarantee that such a bound can be computed.)

As new theories arise from processing events that link an existing theory with a new word match, the m and q scores of an event and the new theory that it will create are simply the respective sums of the m and q scores of the old theory and the word being added to it. Thus, after assigning an m score to a word match by summing the max numbers for the segments that it covers, the m score of any new theory that includes it can be computed by a single addition.

3.d) Admissibi1ity of _the Method

Claim:

The first complete spanning theory found by the shortfall scoring method will be one of the best scoring complete theories (there could be more than one) that can be found by any strategy (i.e., the algorithm is "admissible" in the conventional terminology).

Proof;

At the time the first complete spanning theory has been processed, every other event on the event queue (including continuation events for finding lower scoring seeds or lower scoring words to add to the ends of islands) will already have fallen low enough in its partial score (q score) that no possible match sequence in the remaining region of the utterance can bring its total score above that of the spanning theory. Also, the presence of the continuation events in the queue makes the search process complete in the sense that any word in the vocabulary would be enumerated if the process were continued long enough. Thus there is no possible word sequence across the utterance that would not be considered by this search algorithm if it were run sufficiently far. Hence, any complete theory of the utterance will have a shortfall (m-q) at least as great as that of the first complete theory discovered. Since all spanning theories have the same maxscore m = T, it follows that the first spanning theory also has the maximum possible quality score (q) of any spanning theory.

3.e) Notes

Note that the process can be continued to obtain the second best complete theory, and so on. Note also that the admissibility holds for this method whether the process is left-to-right (i.e., seeds only at the left end of the utterance) or middle-out (seeds anywhere in the utterance), and that it does not require the island collision feature.

The shortfall method works with almost any type of grammar. it makes no assumptions that the grammar is finite-state, as do most Markovian strategies. In the middle-out modes, it does require the linguistic consultant to have a parser (such as the bidirectional ATN parser in the current HwIM system) that can take an arbitrary island fragment in the middle of an utterance and judge whether it is a possible subsequence of an acceptable sentence. In practice, it also helps immensely if the parser can also use the grammar to predict the acceptable words and classes adjacent to the island, and if the Lexical Retrieval component can use such predictions to constrain its search (as in HWIM), but this is not essential to the formal admissibility of the algorithm.

3•f) Avoiding Duplicate Theories

Note that in the middle-out, island-driven strategies there are many different ways of eventually arriving at the same theory. For example, if we have an island w with a possible word x on the left and a possible word y on the right,

then we can first form the theory (xw) and then (xwy) or we can form the theory (wy) and then derive (xwy) from that. Which of these two routes is taken will depend on the scores of the words, but it is quite possible (in fact, likely) that in the course of working toward a complete theory a strategy will arrive at the same subtheory several different times by alternate routes.

If we do not include checks for the duplication of theories, then we would often get two copies of the same theory. These would forever duplicate the same predictions and theory formations, giving rise to a rapid exponential explosion of the search process. If we include a test each time a theory is formed to determine whether that theory has been formed previously, tnen we can avoid this exponential process. In fact, if each time we are about to put an event on the event queue we check the event to see if the set of word matches that it uses is the same as that of some other event, then we can terminate this duplication before making the entry on the queue and consuming the queue space (and certainly before calling the Linguistic component to check it out and make further predictions) .

The check for duplication among all the events that have been created can amount-to a considerable amount of testing if done in a brute force exhaustive test, although it can be considerably reduced by indexing events by their beginning and end points or other tricks. However, if one can rely on the events being Generated in the order determined by the basic shortfall strategy, then the following simple check based only on the word matches at each end of an event can be used to determine whether an event is redundant (i.e., will produce the same theory as some event already generated):

> If the new word is at the left end and has the same or greater shortfall as the word at the right end, then this event is redundant.
>
> If the new word is at. the right end and has strictly greater shortfall than the word at the left end, then this event is redundant.

The argument for the validity of this test is as follows:

In the search space we are considering, it is possible, without a check for duplication, to derive a given theory with words $w^{\wedge},w_2\#...rWu$ in $2^{k\sim i}$ different ways one corresponding to each of the possible binary derivation trees starting with some one of the w- as a seed, and then successively adding words either to the right or the left end. (Proof either w-, or w^ was chosen last, hence there are two ways to derive a string of length k for every possible derivation of a string of length k-1. There is one possible way -- i.e., as a seed -- to derive a string of length 1.) Of all these derivation trees, the first one that will be found is the one that uses the w. with the smallest shortfall as a seed, and at subsequent steps adds the better (in terms of shortfall) of the two words at either end (assume for the moment that no two of the words have exactly the same score). Hence, any derivation that attempts to add a word to one end of an island when that word has a smaller shortfall than the word at the other end of the island will be duplicating a theory that has already been derived (or at least already has an event for it on the event queue). In the case of two competing seeds with the same shortfall or words at each end of an island that have the same shortfall, we have arbitrarily picked the leftmost as the preferred one, which we will permit the algorithm to follow fully, and we block the derivation of duplicates from the other one. Thus, if we have a word being added to the left end of a theory that has the same shortfall as the word at the right end, then this event is redundant, since the preferred order will generate an equivalent event that adds the left end word first.

Thus, a very simple check between the score of the word being added to a theory and the score of the word at the other end of the theory will suffice to eliminate the formation of redundant events.

3.g) Fuzzy Wor_d Matches

The above discussion does not explicitly mention the problem of finding the same word in essentially the same place but with slightly different end points and different scores. We have observed this kind of output from the Lexical Retrieval component of HWIM and indeed find it desirable to know the degree of variation possible in the end points of a word match and the appropriate degradation in score for each. However, it is wasteful to give several different events to the Linguistic component, all of which are adding word matches to a given theory that differ only in their endpoints and scores. For this reason, we have introduced a structure that groups together multiple equivalent word matches into a single entity called a .fuzzy word match (or "fuzzy" for short), which Ts given the score of its best member. A theory containing fuzzy word matches

actually represents a class of grammatically equivalent theories and carries the score of the best one.

When an event is created to add a word match to a theory containing a fuzzy word match at that end, the score of the event must be computed using a "rectified" score that takes into account the best member of the fuzzy that is compatible with the new word (i.e., has boundaries that hook up to the new word and satisfies appropriate phonological word boundary constraints). in general, when several fuzzies are adjacent, the best compatible sequence of members must be chosen, and when the new word match is itself a fuzzy, the best combination of one of its members with a corresponding rectified score for the theory must be taken. The event is thus given the score of the best of the grammatically equivalent, non-fuzzy events for which it stands.

It word matches returned by the Lexical Retrieval component are grouped into fuzzy matches whenever possible, and word events are given appropriately rectified scores, then the above admissibility result still holds (i.e., the first complete theory processed will be the best). The only difference (aside from the elimination of separate processing for grammatically equivalent theories) will be that certain word events (i.e., those using a less-than-best path through the existing theory) will be formed earlier than they otherwise would have. However, these events will still be placed on the queue with the correct score so that they will reach the top and be processed in exactly the same order as they would in the strategy without fuzzies.

3.h) Discussion

The shortfall scoring method is similar in some respects to the well-known branch and bound, technique, except for the characteristic in the middle-out version that the same partial interpretation may be reached by many different paths, and the fact that the space of possible solutions is determined by a grammar. It can also be modeled as an example of the A* algorithm of Hart, Nilsson, and Raphael [1968] for finding the shortest path through a graph, where, in this case, the nodes in the graph are partial interpretations of the utterance, and the connections in the graph correspond to the seed and word events. Consequently, it shares with that algorithm a certain kind of optimality that Hart, Nilsson, and Raphael prove. It is simpler than the general A* algorithm, however, in that we are looking for the best scoring node, and we are not interested in scores of paths leading to that node (in fact all such paths have the same score in our case). The simple argument given previously suffices to show the admissibility of the shortfall method, whereas the general A* algorithm is more complicated.

Measuring the shortfall from any profile that is a per word upper bound of quality score would be sufficient to assure the theoretical admissibility of the method. However, the tightness of the upper bound affects the number of events tried and partial theories created in the search for a successful interpretation (i.e., the "breadth" of the search). By assigning the upper bound as a segment-by-segment profile determined by allocated shares of actual word match scores, a fairly tight upper bound is achieved. A further effect of scoring the shortfall from such a maxscore profile is that the score differences in different parts of the utterance are effectively leveled out, so that events in a region of the utterance where there are not very good scoring words can hold their own against alternative interpretations in regions where there are high scoring words. This promotes the refocusing of attention from a region where there may happen to be high scoring accidental word matches to events whose word match quality may not be as great, but are the best matches in their regions. Thus, an apparently satisfactory and intuitively reasonable strategy for focusing attention emerges from the same strategy that guarantees to get the best scoring theory first.

When using the shortfa11 method for understanding an utteranee, the overwhelming tendency is for an event adding a new word to an island to pick up additional shortfall and fall some distance down in the queue, The result is that other events are processed before any additional work is done on that island. (Occasionally, the new word is the best word in its region and buys no additional shortfall, but this is a rarity.) The distance that this new event falls down the queue is determined by the amount of additional shortfall that it has just picked up and the shortfalls of the events that are competing with it on the queue. This distance directly affects the degree of "depth-first" vs. breadth-first" processing done by the algorithm. If the new word scores well, the event falls only slightly, and few, if any, alternate events are processed before it. In this case the algorithm is relatively depth first. If the new word scores badly, the event falls further down the queue, many more alternative events have priority over it and the algorithm is more breadth first.

The above characterization is only an intuitive approximation, since "the actual number of events processed before the new event is considered depends on the number of new events that will be generated by the intervening events that will still score higher than this one. In some cases, the number of such events can be extensive. The general effect, however, is that the shortfall scoring method provides a dynamically varying combination of depth-first and breadth-first search which is determined by the relative qualities of the events that are in competition.

4. Density Scoring with island Collisions

Density scoring consists of using the score of an event divided by the duration of the region covered by the event as the priority score. One way to view this strategy is to consider again the task of estimating the expected score to be achieved in the region not covered by an event and consider estimating this score as a direct extrapolation of the same score per millisecond that has already been achieved by the event -- i.e., add to the current score an estimated potential score consisting of the score density of the current event times the duration of the region not covered by the event. Since the resulting total estimated score is just the score density of the event times the total duration of the utterance, and the total duration of the utterance is a constant, we can compare only the score densities of the events themselves and achieve the same decisions.

When we think of the score density as an extrapolation of the score already achieved by an event into the region not covered by the event, we are clearly no longer obtaining an upper bound on the possible future score of an event, and the previous proof of admissibility used for the shortfall method no longer applies, in particular, whereas the shortfall is a monotonically increasing function as an island grows, the score densities can get smaller when a bad word is picked up and then get larger again as the theory grows and picks up better words, averaging the score of the bad word over a larger duration. Thus, it is not true that the score density of descendants of an event must be no greater than that of the event itself.

However, when used with the island collision feature that allows one to combine together in one step the word lists of two different events that are noticing the same word from opposite directions, the density method also guarantees that the first complete theory found is the best one. To prove this claim, we must use a different argument than for the basic shortfall strategy. The argument depends on the ability to derive the same theory in different ways from different seeds -- i.e., the middle-out control strategy is essential for the admissibility of the density scoring method.

Lemma:

Using the middle-out density scoring method and using island collision events, any theory covering any region of the utterance can be derived by a sequence of events ail of which have a score density no less than that of the theory itself.

Proof:

By induction on the number of words in the theory. (1) The hypothesis is trivially true for one-word theories by means of a seed event. (2) Suppose we have a theory of k + 1 words with density d. Let w be the word in the theory with the lowest score density (or one cvf them if there are several such), and let x and y be the word sequences on either side of w (one of which may be empty). Then the densities of both x and y must be at least as great as d and the density of w must be no greater than d. If either x or y is empty (i.e., w is at one end), then by the inductive hypothesis, the other has a derivation using events of density no less than its own and hence no less than d and the word event to add w will have density d. If w is not at the end, then by the inductive hypothesis, both x and y have derivations using events of density no less than their own, and hence no less than d. Therefore, before events of density less than d can reach the top of the stack, both of the theories x and y would have been processed, and both would have noticed the word w from opposite sides; hence, an island collision event would have been constructed for the combined theory and would have the combined density d.

Corollary:

When a spanning theory of some density has been found by the middle-out density scoring method with island collisions, any spanning theory of higher density could have been completely derived using events of higher density, and thus would have been found before the theory in question. Hence, the first complete spanning theory found by the density scoring strategy using island collisions will be one of the best possible interpretations.

5. Shortfall Density

The above proof of the admissibility of density scoring makes no assumptions about the scoring metric whose density is being taken other than that it be additive. Hence, the density method can be applied to either the original quality score assigned by the Lexical Retrieval component, or to the local shortfall described previously, giving rise to strategies which we refer to as quality density and shortfall density, respectively. Initial experimental comparison of the algorithms (see Sec. 7) suggests that the shortfall density method is superior to either the quality density or the shortfall method alone.

6 • Other heuristics

In addition to the basic choice of priority scoring metric used tor ranking the event queue, there are additional heuristics that can be used to improve the performance of the island-driven strategies without loss of admissibility guarantees. Two of these are the use of "ghost" words, and the selection of a preferred direction for events from a given theory.

6.a) Ghost Words

The ghost words option is a feature that can be added to any island-driven strategy without affecting the admissibility of the strategy to which it is added. Every time a theory is given to the linguistic consultant for evaluation, proposals are made on both sides of the resulting island (unless the island is already against one end of the utterance). Although events can only add one word at a time to the island, and this must be at one end or the other, eventually a word will have to be added to the other end, and that word cannot score better than the best word -that was found at that end the first time. The ghost words feature consists of remembering with each event the list of words found by the Lexical Retrieval component at the other end and scoring the event using the best of the ghost words as well as the words in the event proper. The result is that bad partial interpretations tend to get bad twice as fast, since they have essentially a one-word look-ahead at the other end that comes free from the linguistic consultant each time an event is processed. On the other hand, an event that has a good word match at the other end gets credit for it early, so that it gets processed sooner. The ghost words feature, thus, is an accelerator that causes extraneous events to fall faster down the event queue and allows the desired events to rise to the top faster. Experimental use of this feature has shown it to be very effective in reducing the number of events that must be processed to find the best spanning event.

6.b) Choo^LQSL a£ Preferred Direction

When a theory is evaluated by the linguistic consultant, predictions are made at both ends of the island. When one of the events resulting from these predictions is later processed, adding a new word to one end of the island, the predictions at the other end of the new island will be a subset of the predictions previously made at that end of the old island. In general, words found by this new island at that end will also have been found by the old island, and if the score of the new island is slightly worse than that of the old island (the normal situation), then the strategy will tend to revert to the old island to try events picking up a word at the other end. This leads to a rather frustrating derivation of a given theory by first enumerating a large number of different subsequences of its final word sequence.

Since any eventual spanning theory must eventually pick some word at each end of the island, one could arbitrarily pick either direction and decide to work only in that direction until the end of the utterance is encountered, and only then begin to consider events in the other direction. This would essentially eliminate the duplication described above, but could cause the algorithm to work into a region of the utterance where the correct word did not score very well without the benefit of additional syntactic support that could have been obtained by extending the island further in the other direction for a while.

Without sufficient syntactic constraint at the chosen end, there may be too many acceptable words that score fairly well for the correct poorly scoring word to occur within a reasonable distance from the top of the queue. By working on the other end, one may tighten that constraint and enable the desired word to appear (although this can never cause a better scoring word to appear than those that appeared for the shorter island).

A flag in the HWIM system causes the algorithm to pick a preferred or "chosen" direction for a given theory as the direction of the best scoring event that extends that theory, and to mark the events going in the other direction from that theory so that they can only be used for making tighter predictions for words at the chosen end. This is accomplished

by blocking any events for one of the ghost words at the inactive end of an event it that event is going counter to the chosen direction. This blocking, alone, eliminates a significant number of redundant generations of different ways to get to the same theory. An even greater improvement is obtained by rescoring the events that are going counter to the chosen direction by using the worst ghost at the other end rather than the best ghost. Since only word matches that score worse than any of the ghosts at that end arc being sought by these events, this is a much better estimate of the potential score of any spanning theories that might result from these events.

The effect of rescoring the events in the non-chosen direction using the worst ghost is that, in most cases, these events fall so low in the event queue as to be totally out of consideration. Only in those cases where there was little syntactic constraint in the chosen direction and the worst matching word at that point was still quite good, do these events stay in contention, and in those cases, the use of the worst ghost score provides the appropriate ranking of these events in the event queue.

7. Empirical Comparison of -t!2£ DT?f^er.^ent Strategies

In the HWIM Speech understanding system, approximations to the above algorithms have been implemented. The major approximation is that continuation events are not. implemented, but instead the initial values of n and k are chosen large enough that one believes that the correct interpretation of the utterance is found before any of the continuation events would have reached the top of the queue. If such is the case, then all of the decisions made by the approximation arc the same as those of the admissible theoretical algorithm, and hence the first complete theory found will still be guaranteed to be the best.

Details of the HwIM system and its general performance are found in [Woods et al., 1976]. Comparative performance results on a set of 10 utterances for the shortfall (S), shortfall density (SD), and quality density (QD) scoring strategies are shown in Table 1 below. The option of using the quality score (0) alone as a priority score is given for comparison.

These experiments were run using the ghosts, island-collision, and preferred direction heuristics with a resource limit of 150 theories to process before the system would give up with no response. The ten sentences used for the test were chosen at random from a test set of 124 recorded sentences.

Although a test set of only ten utterances is admittedly too small, 1 believe that the trends indicated in the-figure are generally correct. Specifically, while the quality density alone leads to a spanning interpretation in relatively few theories, it does so without any expectation of getting the best interpretation. In this case, only two out of three of its answers arc correct. All of the other methods spend additional effort in making sure that the best interpretation is found, and consequently found fewer spanning interpretations within the resource limitation. We did not try running the quality scoring strategy beyond the first interpretation to see if a better interpretation could be found since, among other things, it is not clear when to terminate such a process. Running in this mode, one could easily enumerate more theories than the other methods and still not have any guarantee that the best interpretation had been discovered.

None of the admissible algorithms found incorrect interpretations, so the reliability of their interpretations when they get them is 100% (providing the acoustic phonetic analysis of the . input utterance does not cause some incorrect interpretation to score higher than the correct one, a situation that occurs sometimes in the HVIM system, but is not a factor in this experiment). Unfortunately, the shortfall strategy alone is so conservative in doing this that it failed to find any interpretations within the resource limit. Both of the density methods are clearly superior to the straight shortfall method.

The shortfall density strategy ranked superior to the quality density strategy in terms of the number of events that needed to be processed to find the first spanning interpretation and consequently found more correct interpretations within the resource limitations.

The effects of the island collision (C), ghosts (G), and preferred direction (D) heuristics are shown in Table 2 (where SD+0 means shortfall density without collisions, ghosts, or chosen direction, SD+C means shortfall density with island collisions, etc.). The inclusion of a heuristic does not always guarantee that the system will understand an utterance in fewer theories, but the pooled results shown (note especially the series SD+0, SD+G, SD+GD, SD+GDC) suggest that the successively added heuristics produce improvements in both accuracy and number

ol theories required. (Note that our admissibility results hold only for the SD+C and SD+GDC cases above.)

8• Conclusion

he have presented two basic priority scoring methods, shortfall and density scoring, that provide admissible search strategies for finding the best matching interpretation of a continuous speech utterance. Moreover, the two methods can be used in conjunction, and the combined method appears to be more efficient than either of the methods by themselves. Although the methods are presented here in the context of speech understanding, analogous methods should be applicable to other perceptual tasks such as vision with appropriate generalizations of segment, word, and phrase. The density scoring strategy is especially interesting since it is not an instance of the A* algorithm and, at least for the speech understanding problem, appears to be superior to the corresponding A* algorithm (the shortfall method) in the number of hypotheses that need to be explored to obtain the optimum solution. It apparently gains this superiority by its ability to work on different parts of the solution independently and combine them by the mechanism of island collision. The density method is of course not applicable to as wide a class of problems as the general A* algorithm, but should be applicable to any search problem where scores are accumulated from partial hypotheses associated with some analog of a region. In particular it can be generalized to two-dimensional regions for such problems as vision [Woods, 1977].

Hart, P., N. Nilsson, and B. Raphael (1968)
"A Formal Basis for the Heuristic Determination of Minimum Cost Paths," IEBE Trans. Sys Sci Cybernetics, July, Vol. SSC-4, No. 2, pp. 100-107.

Wolf, J.J. and W.A. Woods (1977)
"The HWIM Speech Understanding System," Conference Record, IEEE International Conference on Acoustics S£eech and Signal pfocessing, Hartford, Conn., May.

Woods, W., M. Bates, G. Brown, B. Bruce, C. Cook, J. Klovstad, J. Makhoul, B. Nash-Webber, R. Schwartz, J. Wolf, V. Zue (1976)
"Speech Understanding Systems - Final Technical Progress Report," 30 October 1974 to 29 October 1976, BBN Report No. 3438 Vols. I-V, Bolt Bcranek and Newman Inc., Cambridge, Ma.

Woods, W.A. (1977)
"Theory Formation and Control in a Speech Understanding System with Extrapolations towards Vision," Proc. oj Workshop on Computer Vision Systems, University of Massachusetts, Amherst, June.

	Q	QD	S	SD
Correct first interpretation	4	3	0	5
Incorrect first interpretation	2	0	0	0
No response	4	7	10	5
Average number of theories processed	23	142	-	96

Table 1. Comparison of different priority scoring functions.

	SD+0	SD+C	SD+G	SD+GD	SD+GDC
Correct	3	3	3	4	5
Incorrect	0	0	0	0	0
No response	7	7	7	6	5
Average number of theories processed	162	121	145	130	96

Table 2. The effects of island collisions, ghosts, and direction preference.