STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

File Copy

THE SRI SPEECH UNDERSTANDING SYSTEM

by

Donald E. Walker
Artificial Intelligence Center

Technical Note 91

SRI Project 1526

Proceedings IEEE Speech Symposium, Carnegie-Mellon University, Pittsburgh, Pennsylvania, April 15-19, 1974.

The work reported herein was sponsored by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC04-72-C-0009 with the U. S. Army Research Office.

THE SRI SPEECH UNDERSTANDING SYSTEM*

Donald E. Walker
Stanford Research Institute
Menlo Park, California

Summary

This paper describes the structure of the SRI speech understanding system and presents the available data on its performance. The system is distinctive in the way that knowledge of various sources is coordinated by a "best-first" parser to predict the sequence of words in an utterance, and in the use of word functions--programs that represent the acoustic characteristics of a word--to test the predictions.

Introduction

SRI is participating with other ARPA/IPT contractors in a major program of research on the analysis of continuous speech by computer.[1] The goal is to develop, over a five-year period, a speech understanding system capable of engaging a human operator in a natural conversation about a specific task domain. During the first two years, the work of the SRI project was directed toward a simulation of the actions required to assemble and repair small devices such as faucets and pumps. The intent was to allow a person speaking to the system to direct its operations, to ask questions about available tools and parts or about other elements in the task domain, and to add information so that, for example, characteristics of a given device could be related to other devices with which the system has had experience in the past.

This paper is a progress report on the status of the system after two years; it emphasizes overall system design and performance, rather than the operation of particular components. Separate papers are presented in this symposium on the parser[2] and on our special procedures for phonetic and phonological analysis.[3] In addition, a paper on task-oriented dialogs[4] describes how protocol experiments can provide information to guide the development of models for a task domain.

System Structure

System Concept

The structure of the system is represented in Figure 1. Information from various sources of knowledge is coordinated by the parser to predict the sequence of words in an utterance spoken in the context of a particular task domain. On the basis of this information, a priority is assigned to each path branching from the current choice point in the grammar. In following a path, when a word is predicted for a particular place in the utterance, a word function is called. Each word function contains a representation of the acoustic features of that word based on its pronunciations in a variety of contexts. A test of the particular word function against acoustic data from the utterance returns a priority for assuming that the word is present. This value is part of the information used to decide which path to follow next. A complete analysis of an utterance produces a program that operates on a model of the world corresponding to the task domain. The execution of the program constitutes the "understanding" of the utterance; then, an appropriate response is made.

Task Domain

The task domain is a simulation of the actions required for assembly, test, and repair of small mechanical devices. The initial task is "repairing a leaky faucet." More information about it is provided in the section under World Model.

Basic System Components

Acoustic Preprocessing. An utterance is recorded in a double-walled IAC booth, using a 1-inch condenser microphone and a studio-quality tape recorder. The signal is bandpass-filtered at 80-8000 Hz by an active 70-dB/octave filter; it is then digitized at 20,000 sps with an 11-bit quantization, providing a signal-to-noise ratio of about 40 dB. The digitized wave form is monitored on a CRT for excessive peak clipping or underusage of dynamic range. The digitization is performed on a PDP-11, and the results are transferred by DEC-tape to the PDP-10 for the next steps (an interim expedient).

Classification--The raw signal plus the outputs from four digital filters--80-200 Hz, 300-1000 Hz, 500-2800 Hz, and 3200-6800 Hz--are used to classify each 10-ms interval as vowel-like, voiced stop, voiced turbulence, unvoiced turbulence, silence, or transient/unknown. Three other digital filters--1500-8000 Hz, 1500-3000 Hz, and 4000-8000 Hz--are used to classify turbulent intervals as s, sh, f-th, z-zh, or v-dh.

Spectral Analysis--An LPC analysis of the voiced intervals provides frequency and half-bandwidth values for the first five formants. The LPC is performed over a 15-ms interval, using 3 dB/octave preemphasis, 28 coefficients, a Hamming window, and 128 estimation points in the frequency continuum.

Word Functions. A word function is prepared after a detailed examination of acoustic data from that word in typical contexts from a variety of utterances. Judgments are made about the acoustic parameters that are most relevant and the variations in the parameters that are most reasonable. Each word function consists of a series of Fortran subroutines that use data from a variety of sources: the acoustic preprocessing of the utterance; algorithms for level (volume) detection, formant smoothing, detecting formant discontinuities,

*The work reported herein was sponsored by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC04-72-C-0009 with the U.S. Army Research Office.

fitting formant trajectories, and identifying formant bandwidths; and, specially designed digital filters or LPC analyses.

Spectrograms and the results of acoustic preprocessing are available for almost 300 utterances. When the interactive speech analysis system is used, a graphic display of these results ean be seen for each utterance; specifically, the classification of 10-ms intervals and the positions of each of the first three formants. A photograph of the display developed on transparent acetate and overlaid on a spectrogram of the utterance provides a permanent form that can be photocopied for ease in handling. A more detailed representation of the acoustic information is available in printouts containing all of the data determined for each 10-ms interval during the acoustic preprocessing.

The data base also can be processed by an interactive program exerciser that can produce the values of any specified parameters for portions of those utterances in the data base that contain occurrences of a particular word. In addition, it is possible to test out the various algorithms, digital filters, or special LPC analyses over these same intervals.

On the bases of these sources of information, decisions are made about the analysis techniques to use in a word function, their order of application, and the appropriate parameters. The exerciser then is used to test the word function against utterances in the data base to determine where it succeeds and fails. Threshold changes can be made on-line, interactively in the exerciser; or other analysis techniques can be selected and the exerciser run again. Forty-two word functions are currently available.

A more detailed description and evaluation of phonetic and phonological analysis through word functions is presented in Ref. 3.

Parser. The parser performs a dual role. In addition to handling the usual parsing functions, it calls on the other components and coordinates information from them.

The parser executes a top-down, "best-first" strategy. Each new path resulting from a choice point is assigned a priority according to its estimated likelihood of arriving at a correct parse. These paths are added to the set of all possible paths created but not yet extended during the parse. The system follows the highest priority path until its priority drops or a choice point is reached. At this time, the cycle repeats. When the highest priority path requires testing for the presence of a particular word, the appropriate word function is called. A complete analysis of an utterance produces a program that references procedures and data in the world model.

Several sources of knowledge currently affect the value of a priority: the grammar (parameters are assigned to alternative branehcs at a choice point, reflecting our judgment about their relative likelihood), the world model, and the results of the word function test.

The basic programs for the parser are written in INTERLISP, but an interpreter has been added to handle multiprocessing control structures and to faeilitate sharing information among the competing processes looking for a parse. A single family of proeesses acts as the sole producer of certain constructions (currently just simple noun phrases); processes needing such constituents provide the contexts for establishing priorities within the family and aet as consumers of structures produced by it.

A debugging facility for the multiprocessing control structures has been developed and integrated into the standard LISP debugging package. It allows us to trace the overall parse, to check the history of various processes, and to establish break points in the grammar.

The operation of the parser is described in more detail in two other papers.[3,5]

World Model. The world model contains data reflecting the current configuration of objects and procedures for operating on that configuration. The program resulting from a completed parse will contain references to specific objects corresponding to definite descriptions (definite noun phrases and proper nouns), procedures for finding objects satisfying indefinite descriptions (other noun phrases), and procedures for testing or establishing relations (given by verbs and prepositions). Executing the program for an imperative produces a description of actions performed in carrying out that command. Executing the program for an interrogative produces a set of objects satisfying the queried relation.

The faucet world contains a faucet, screws, washers, tools, and other objects relating to plumbing. The objects can be of different sizes and colors; their location, type, and condition can be specified. Class membership, superset relations, and various kinds of connections among the objects are possible. A variety of actions can be performed that entail moving, picking up, screwing, unscrewing, and the like.

The current implementation in QLISP[6] provides capabilities for updatable model manipulation, including associative storage and retrieval of procedures and data.

Response. The response of the system depends on the form of the utterance, the state of the world, and the result of executing the program. For imperatives, the system simply lists the actions performed. For interrogatives, the type of the question determines the general form of the response. For example, a "how many" question is answered by giving the size of the answer set returned by the program for the utterance and then by describing the members of that set. In forming a description of the objects, the world model is used to find identifying properties (such as size, color, and type), and the utterance phrase corresponding to the object (e.g., 'What bolt?') is used to form a natural abbreviation (e.g., 'The small one.'). Currently, responses are typed out on the terminal.

Sources of Knowledge

Grammar and Semantics. The grammar contains clause, noun-group, and verb-group subgrammars, and a case component. The clauses are: major--declarative, imperative, and interrogative; adjunct; sentence complement; and noun-qualifying. The verb-group subgrammar allows: present and past tense, active and passive voice, and auxiliary preceding or separate from the main verb. The noun-group subgrammar allows: determiner, quantifier, adjective, pronoun, noun, qualifying phrase, and qualifying clause. In the case component, verb functions establish for each verb sense its obligatory and optional cases--causal actant, instrument, theme, locus, source, goal, loc, and time. Each verb function calls a paradigm that specifies the allowable clause positions for the different case arguments and maps these arguments into a goal state to be achieved in the world model.

The current vocabulary can handle 54 words, including 11 plural and five past tense forms. (This number is determined by the number of word functions written; there are syntactic and semantic specifications for almost 300 words.) Each entry can include word type, syntactic and semantic features, a call to the paradigm (for verbs), a function for resolving anaphoric reference, information specifically related to the world model, a function for evaluating its priority, and a call to its acoustic word function. Currently, we can handle anaphora for pronouns (e.g., 'it' and 'one'), and definite noun phrases that are in case arguments in the main clause of the sentence.

World Model Information. In addition to its role as a basic system component, the world model functions as a source of knowledge to affect the value of priorities in the parsing. At present, its use in this role is to determine whether definite noun phrases or prepositional phrases correspond to some portion of the model.

The presence of a model of the task would be a valuable addition. Therefore, toward this goal we have specified the sequences of steps involved in repairing a faucet that has a worn washer, but this information has not been incorporated into the world model.

User and Discourse Models. A user of the system is presumed to be task-oriented, in a quiet room, and speaking in a clear, "normal" manner without repetitions or noisy hesitations. Currently, threshold parameters in the word functions are adjusted for the voices of two different speakers; in only a few cases has it been necessary to provide different values for each. To extend the system to larger numbers of users, speaker-dependent information would have to be incorporated for adequate discrimination.

The person using the system is assumed more likely to use certain constructions. The initial parameters for the priorities of different rules in the grammar are set according to our intuitions about these likelihoods. However, we recognize the need to model the discourse more accurately; for example, the parameters should vary depending on progress in the performance of the task. Consequently, we have been conducting protocol experiments to gather data about the form and sequence of the utterances likely to occur in dialogs with the system.[4] Eventually, it may be possible to incorporate speaker-dependent information about the relative frequencies of use of different constructions. Robinson[7] provides an extended discussion of linguistic performance that is clearly relevant to modeling differences among speakers.

Acoustic-Phonetic Data. Prosodic information can be critically important in a system of this design. Accordingly, we have been using data from the protocol experiments to try to correlate intonation cues with sentence type, to relate stress cues to the distinction between old and new content elements in the dialog, and to identify phrase boundaries within an utterance. Our studies of utterances parsed by the system indicate clearly that this information could be used to increase the efficiency of the analysis.

System Performance

General

The system has been running for only a few months. During this time, we have made many modifications to ensure that the various acoustic, syntactic, semantic, and pragmatic processing steps are working satisfactorily. Thus, although we have now processed 71 utterances, not all of them constitute "tests" of the system. That is, in developing the algorithms for a word function, the acoustic data for an utterance may have been used in establishing threshold values. Our experience with word functions is still sufficiently limited so that modifications do not necessarily generalize to accommodate the occurrence of a given word in a different context. In contrast, changes in the grammar and the semantics, made so that a particular utterance can be understood, contribute to the overall improvement of the system in a cumulative fashion. Before discussing the results for the total set of utterances, it is appropriate to consider the results for one block of ten that were processed as a test set.

Analysis of a Test Set of Utterances

Four sentences were recorded:

- Z0 Put one washer in the faucet.
- Z1 Grasp the crescent wrench.
- Z2 Is it in it?
- Z3 What little brass parts are in the box?

Three speakers were used: the two for whom thresholds in the word functions had been written (B and P), and one whose voice had not been recorded for the system before (K). Each speaker recorded the first three sentences; the fourth was recorded just by the third speaker, since the other two had recorded it previously.

The results for the test set were examined in a two step procedure. First, the word functions were checked against the acoustic data for the utterances; these results are presented in Table 1. For three utterances--coincidentally, one for each speaker--all of the word functions worked correctly. For four

Table 1

ACOUSTIC RESULTS FOR THE TEST SENTENCES*

	Speaker B	Speaker P	Speaker K
ZØ	put	put, one, faucet	washer, in
Z1	wrench	OK	crescent, wrench
Z2	OK	in	OK
Z3			little, part
Subtotals:			
	12/14 = 86%	10/14 = 71%	17/23 = 74%
Total:		39/51 = 77%	

*Threshold changes were made for the words not underlined; algorithm changes would be required for the underlined words.

others, adjustments in thresholds were made. For the last three, a change in one of the algorithms would have been required.

Some examples will be given of the kinds of threshold adjustments made. The valuea of the upper limit for the first formant for several of the vowels had to be changed. For 'put,' the value had been set at 500, but in utterance ZØ for Speaker B, it was actually 518. Corresponding pairs of values for 'in' were 500 and 553, and for 'crescent' 600 and 602. For 'faucet,' the upper limit on the variance about the line fitted to the second formant had been set at 5; in utterance ZØ for Speaker K it was actually 5.9. Because of the relatively little experience we have had with setting thresholds for word functions, these adjustments are not considered to be significant errors in the system. Actually, a change in the form of the threshold cut-off function from absolute to graded probably would have accommodated most of the differences.

Changing an algorithm in a word function is more significant; it indicates a failure to account for a certain kind of acoustic event. For example, in utterance Z3 for Speaker K, the second liquid in 'little' had such a reduced flap-D that it could not be detected by our current algorithms. More adequate detectors would provide the discrimination needed. In compiling statistics for the acoustic section, these requirements for algorithmic changes are considered failures.

Given these interpretations, the acoustic results may be summarized as indicated at the bottom of Table 1. For the three speakers, the word functions were correct for 12/14 = 86%, 10/14 = 71%, and 17/23 = 74%, respectively; therefore, the aggregate value is 39/51 = 77%.

The seven utterances for which the word functions were considered to work satisfactorily were parsed. First, however, two minor changes were made. It was necessary to raise the priority for a construction with a null determiner (in ZØ), because, through an oversight in setting the initial parameters, it had an unrealistically low value. The word 'is' was included among verbs that could take two anaphoric references; this modification had not been made when that addition to our anaphoric routines had been introduced for the other verbs. With these changes, five of the utterances were analyzed eorrectly, as shown in Table 2. For one, the system recognized 'a' rather than 'the'; the response from the system would be the same in either case. In the last utterance, the correct analysis was found, but the presence of vocal fry at the end of the utterance resulted in a lowering of the priority of that analysis; so, it continued to look for a longer word, and subsequently, it accepted an interpretation for which the values of the word functions actually were lower. To summarize, the system parsed six of the scven utterances for a rating of 86%. If we include the three utterances that failed on acoustic grounds, the system analyzed six out of ten, or 60%.

Table 2

PARSING RESULTS FOR THE TEST SENTENCES

	Speaker B	Speaker P	Speaker K
ZØ	OK	'a'/'the'	OK
Z1	--	OK	OK
Z2	OK	--	failed
Z3			--
Subtotals:			
	2/2 = 100%	2/2 = 100%	2/3 = 67%
Total:		6/7 = 86%	

The results from this set of test utterances represent our first attempt at an assessment of the system. They are more valuable to us as guides to further work, than as benchmarks of achievement. From this perspective, the implications of their analysis will be considered in the next section together with the results from all of the utterances that have been processed by the system.

Analysis of All Processed Utterances

Of the almost 300 utterances we have recorded relating specifically to the faucet world task domain (there also were a large number done for the "blocks world"), 71 have been run through the system. For most of the remainder, there are not enough word functions written to handle the vocabulary. About three dozen for which there are word functions have not been run, either because they contain constructions that we have not yet included in the grammar, because our semantics cannot yet process them, or because they do not make sense. Utterances in this last category were recorded primarily to provide alternate acoustic contexts that could be used in preparing word functions without concern for their syntactic or semantic adequacy. (The larger set of utterances for which we do not have word

functions also contains instances of each of these categories.) Eventually, we would want the system to respond to every input in some constructive manner, but we believe that efforts in that direction are best deferred until we do better on the ones the system should be able to process.

Of the 71 utterances run, the system returned a complete parse with a response for 51. Of these, 44 were understood correctly. This number includes three instances in which 'a' was recognized instead of 'the'; one with 'the' for 'a'; one with 'in' for 'on'; and one with 'one' for 'the.' In all six cases, the response of the system was appropriate for the utterance as recorded.

The other seven utterances were processed incorrectly. For three, a path for the correct parse was also present, but audible sounds on the recording after the end of the last word (e.g., vocal fry) caused the system to look for an alternative analysis that would be longer. The results were: 'Pick up a big wrench' instead of 'Pick up a big one'; 'How many big brass parts are the handle?' instead of 'How many big brass parts are there?' (clearly, the grammar should have precluded that result); 'Is there a wrench?' instead of 'Is it in it?' (the priorities for the correct path were higher, except for the penalty caused by the faulty identification of the end of the utterance). In another, the path for the correct parse was present and the acoustic score for the correct word was actually higher; however, the combination with the plural form was slightly lower: 'How many wrenches are in the box?' instead of 'How many washers are in the box?' Two others accepted incorrect words without ever considering the correct ones: 'How many big tools are on it?' instead of 'How many big tools are there?'; 'Is there a little one in the faucet?' instead of 'Is there a little one in the box?' (the path leading to 'faucet' was so high that it compensated for a relatively low rating for the acoustic match). In the last utterance, a word function failed, and a word with a relatively low acoustic score was accepted: 'Is there a little handle in a box?' instead of 'Is there a little one in the box?'

Twenty utterances never parsed. For 12 of these, the word functions failed to identify words that were present; with two exceptions ('big' and 'one'), these were function words ('a,' 'the,' 'in,' 'there') and words with liquids (three instances of 'tool' and two of 'little'). A more refined acoustic analysis will help with liquids, but function words have long been recognized as problems for speech recognition and speech understanding, and they will continue to be. We believe that it may be possible to develop syntactic strategies that will compensate for acoustic failures to some extent.

Another six utterances terminated by exceeding the limit on the number of processes that could be created (set for 500); however, each contained a path representing a correct analysis of the utterance up to the point of termination. Most of these cases resulted from failures by the word functions to reject words that were not present--with a consequent proliferation of paths in the analysis. Although better word functions would also help here, other sources of knowledge could be used advantageously to lower the priority functions for paths that were inappropriate for semantic or pragmatic reasons.

One utterance terminated by exceeding the process limit because a large gap between successive words reduced the priority of the path; procedures for handling interword coarticulation--required in any case--should help. The last utterance failed because the initial priority for 'one' as a noun group was set too low; however, the word itself was accepted at the appropriate position as an ordinal.

Conclusions

This paper has described the structure of the SRI speech understanding system and presented the available data on its performance. The system is distinctive in the way that knowledge of various sources is coordinated by the parser to predict the sequence of words in an utterance, and in the use of word functions--programs that represent the acoustic characteristics of a word--to test the predictions.

In processing 71 utterances, the system responded as follows: 44 (62%) correctly understood, 7 (10%) incorrectly understood, and 20 (28%) not understood. As indicated above, it is difficult to know how to interpret these results, since not all of them constitute valid tests. The system was designed to make use of many sources of knowledge in the analysis of an utterance. However, the current performance reflects the use of only primitive capabilities; consequently, these results could be interpreted as a lower bound on the power of the system.

More important for further system development, the analysis of each utterance provides guidance for modifications. In addition, we know how to refine and augment each of the system components to handle inadequacies we already recognize. Therefore, we believe that our experiences with the SRI speech understanding system will prove valuable in our further efforts toward satisfying the specifications for system performance described in the ARPA Study Group Report.[1]

Acknowledgments

A large number of people at SRI have contributed to this project over the past two years. The most recent work has been sustained primarily by Sharon Baranofsky, Dick Becker, Barbara Deutsch, Grant Hoyt, Bill Paxton, Tito Poza, Ann Robinson, and Jane Robinson. We also are indebted in various ways to other participants in the ARPA Program on Speech Understanding Research.

References

1. Newell, Allen et al., Speech Understanding Systems, North-Holland Publishing Co., Amsterdam (1973).

2. Paxton, William H., A Best-First Parser, presented at the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, 15-19 April 1974.

3. Becker, Richard W., and Poza, Fausto, Acoustic Processing in the SRI Speech Understanding System, presented at the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, 15-19 April 1974.

4. Deutsch, Barbara, The Structure of Task-Oriented Dialogs, presented at the IEEE Symposium on Spcech Recognition, Carnegie-Mellon University, 15-19 April 1974.

5. Paxton, William H., and Robinson, Ann E., A Parser for a Speech Understanding System, International Joint Conference on Artificinl Intelligence, Stanford, California, 20-23 August 1973, Advance Papers of the Conference, Stanford Research Institute, Menlo Park, California, 1973, pp. 216-222.

6. Reboh, Rene, and Sacerdoti, Earl, A Preliminary QLISP Manual, Technical Note 81, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California, August 1973.

7. Robinson, Jane J., Performance Grammars, presented at the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, 15-19 April 1974.

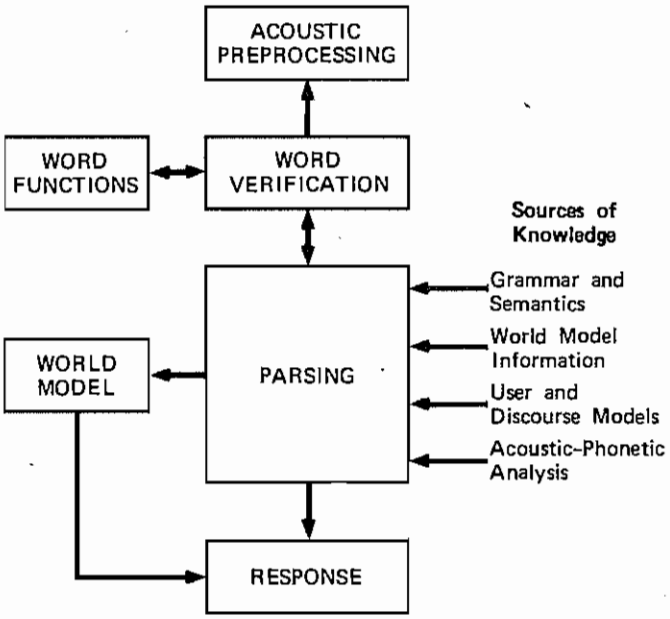

Figure 1. Structure of the SRI Speech Understanding System