MULTI-SYSTEM APPROACH TO SPEECH UNDERSTANDING

DR. RAJ REDDY

CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA

INTRODUCTION

In 1971, a group of scientists recommended the initiation of a five-year research program towards the demonstration of a large-vocabulary connected speech understanding system (Newell et al., 1971). Instead of setting vague objectives, the group proposed a set of specific performance goals (see Fig. 1.1 of Newell et al., 1971). The system was required to accept connected speech from many speakers based on a 1000 word vocabulary task-oriented grammar, within a constrained task. The system was expected to perform with less than 10% semantic errors, using about 300 million instructions per second of speech (MIPSS)* and to be operational within a five year period. The proposed research was a highly ambitious undertaking, given the almost total lack of experience with connected speech systems at that time.

The Harpy and Hearsay-II systems developed at Carnegie-Mellon University had the best overall performance at the end of the five year period. Figure 1 illustrates the performance of the Harpy system relative to the original specifications. It not only satisfies the original goals, but exceeds some of the stated objectives. It recognizes speech from male and female speakers using a 1011-word-vocabulary document retrieval task. Semantic error is 5% and response is an order of magnitude faster than expected. The Hearsay-II system achieves similar accuracy and runs about 2 to 20 times slower than Harpy.

Of the many factors that led to the final successful demonstration of these systems, perhaps the most important was the systems development methodology that evolved. Faced with prospects of developing systems with large number of unknowns, we opted to develop several intermediate "throw-away" systems rather than work towards a single carefully designed ultimate system. Many dimensions of these intermediate systems were deliberately finessed or ignored so as to gain deeper understanding of some aspect of the overall system. The purpose of this paper is to illustrate the incremental understanding of the solution space provided by the various intermediate systems developed at CMU.

*The actual specifications stated "a few times real-time" on a 100 MIPS (Million instructions per second) machine.

GOAL (Nov. 1971)	HARPY (Nov. 1976)
Accept connected speech	Yes
from many	5 (3 male, 2 female)
cooperative speakers	yes
in a quiet room	computer terminal room
using a good microphone	close-talking microphone
with slight tuning/speaker	20-30 sentences/talker
accepting 1000 words	1011 word vocabulary
using an artificial syntax	avg. branching factor = 33
in a constraining task	document retrieval
yielding 10% semantic error	5%
requiring approx. 300 MIPSS*	requiring 28 MIPSS
	using 256k of 36 bit words
	costing $5 per sentence processed

*The actual specifications stated "a few times real-time" on a 100 MIPS (Million instructions per second) machine.

Figure 1. Harpy Performance Compared to Desired Goals

Figure 2 illustrates the large number of design decisions which confront a speech understanding system designer*. For each of these 10 to 15 design decisions, we have 3 to 10 feasible alternative choices. Thus the solution space for speech systems seems to contain 10^6 to 10^8 possible system designs. Given the interactions between design choices, it is not possible to evaluate each design choice in isolation outside the framework of the total system.

SYSTEMS

Figure 3 shows the genealogy of the speech understanding systems developed at CMU. In this section we will briefly outline the interesting aspects of each of these systems and discuss their contributions towards the development of speech understanding systems technology. More complete descriptions of these systems can be found in the references listed at the end.

THE HEARSAY-I SYSTEM (Erman, Fennel, Lowerre, Neely, and Reddy)**

Hearsay-I (Reddy, Erman, and Neely 1973; Reddy, Erman, Fennel and Neely 1973), the first speech understanding system developed at Carnegie-Mellon University, was demonstrated in June of 1972. This system was one of the first connected speech understanding systems to use task dependent knowledge to achieve reduction of the search space. Recognition uses a best-first search strategy.

Model

Hearsay-I was the first system to utilize independent, cooperating knowledge sources and the concept of a global data base, or "blackboard", through which all knowledge sources communicate. Knowledge sources consist of the acoustic-phonetic, syntactic, and semantic modules. Each module operates in the "hypothesize-and-test" mode. Synchronous activation of the modules leads to a best-first search strategy. Several other systems have used this stratety (Forgie 1974). This system was one of the first to use syntactically derived word diagrams and trigrams, as anti-productions (Neely 1973), to predict forward and backward from "islands of reliability". Task dependent knowledge, such as a board position in the chess task, is used by the semantic module (Neely 1973), to reject meaningless partial parses early in the recognition process.

*Further discussion of many of these design choices can be found in Reddy (1976).

**The principle contributors towards the development of each of these systems are listed within parentheses.

Task characteristics

- speakers; number, male/female, dialect
- vocabulary and syntax
- response desired

Signal gathering environment

- room noise level
- transducer characteristics

Signal transformations

- digitization speed and accuracy
- special-purpose hardware required
- parametric representation

Signal-to-symbol transformation

- segmentation?
- level transformation occurs
- label selection technique
- amount of training required

Matching and searching

- relaxation: breadth-first
- blackboard: best-first, island driven
- productions: best-first
- Locus: beam search

Knowledge source representation

- networks
- procedures
- frames
- productions

System organization

- levels of representation
- signal processor/multi-processor

Figure 2. Design Choices for Speech Understanding Systems

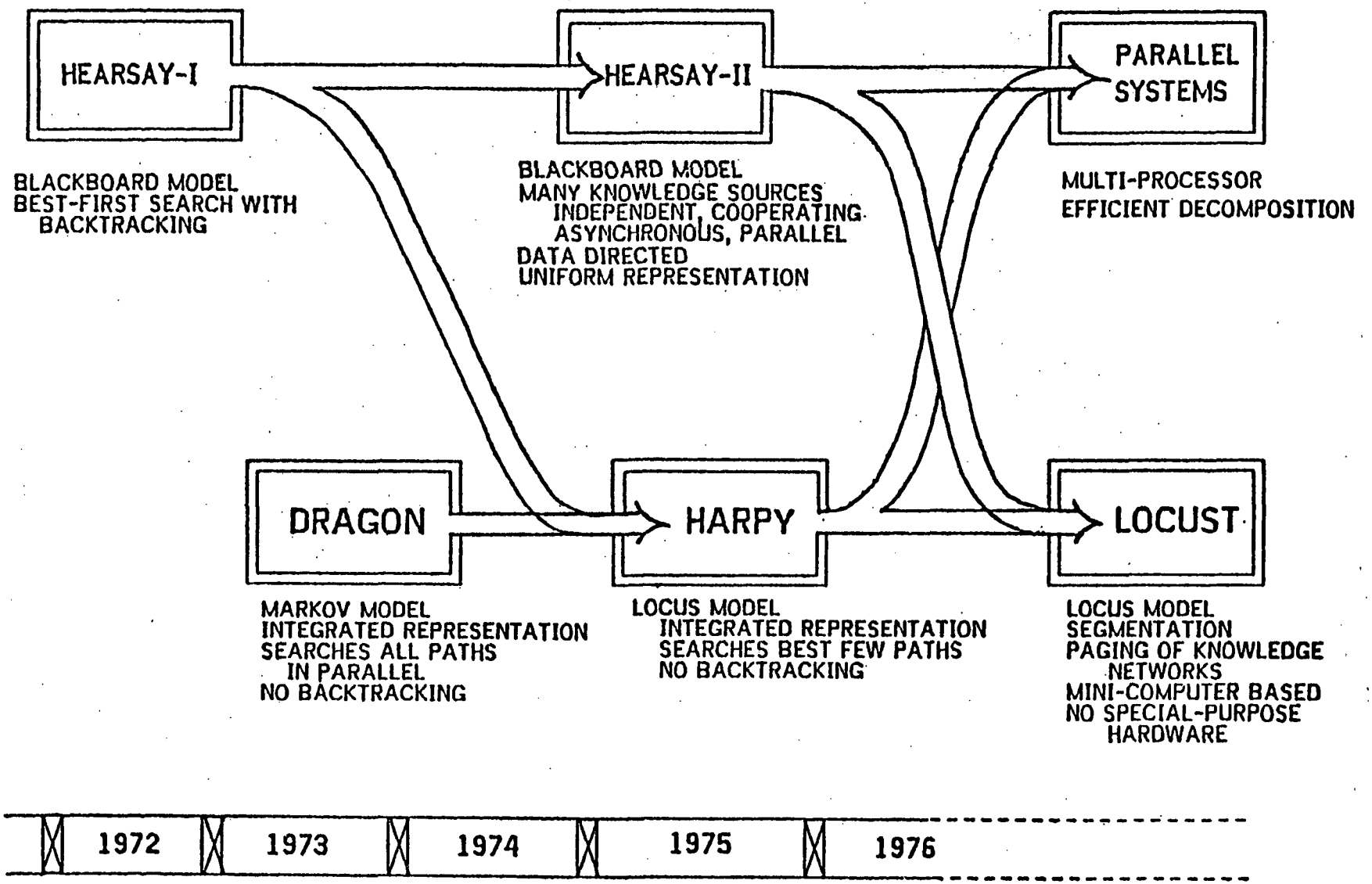

Figure 3. CMU Speech Understanding Systems Genealogy

The acoustic-phonetic module uses amplitude and zero-crossing parameters to obtain a multilevel segmentation into syllable-size and phoneme-size units (Erman, 1974).

Performance

Over a wide range of tasks, the average sentence error rate was 69% with a word error rate of 45%. Speed varied between 3 and 15 MIPSS over 162 utterances containing 578 words. Hearsay-I yields much higher accuracies on tasks with which it is carefully trained. For the chess task, for instance, average sentence and word error rates were 21 and 7 percent, respectively, with an average speed of 2 MIPSS.

Discussion

Hearsay-I, as a successful connected-speech understanding system, served to clarify the nature and necessary interaction of several sources of knowledge. Its flexibility provided a means for testing and evaluating competing theories, allowing the better theories to be chosen as a basis for later systems. In retrospect, we believe this system organization would have been adequate for the ARPA specifications given present acoustic-phonetic knowledge.

THE DRAGON SYSTEM (Baker)

Baker formulated the recognition process as a dynamic programming problem. The Dragon recognition system (Baker, 1975), based on this model was first demonstrated in April of 1974. The system was motivated by a desire to use a general abstract model to represent knowledge sources. The model, that of a probabilistic function of a Markov process, is flexible and leads to features which allow it to function despite high error rates. Recognition accuracy was greater with Dragon than with Hearsay-I, but the system ran significantly slower.

Model

Dragon was the first system to demonstrate the use of a Markov model and dynamic programming in a connected speech understanding system. It included several interesting features, such as delayed decisions and integrated representation, and is based on a general theoretical framework. The general framework allows acoustic-phonetic, syntactic, and semantic knowledge to be embodied in a finite-state network. Each path through this precomplied network represents an allowed pronunciation of a syntactically acceptable sentence. Recognition proceeds left-to-right through the network, searching all possible paths in parallel to determine the globally optimal path (i.e., the path which best matches the spoken utterance). Acoustic inputs are peak-to-peak amplitudes and zero-crossings from overlapping, one-third octave filters, sampled every centisecond.

Performance

Recognition accuracy was greater with Dragon than that obtained with Hearsay-I, but at a cost of speed, Dragon being approximately 5 to 10 times slower. Over a wide variety of tasks, the average sentence error rate was 51%. Speed ranged from 14 to 50 MIPSS. The computation is essentially linear with the number of states in the Markov network. Performance was later improved by Lowerre.

Discussion

Dragon, with more accurate performance than Hearsay-I, served to stimulate further research into factors that led to its improved performance. Many of the ideas motivating its design were important in the development of subsequent connected-speech understanding systems. Although later systems do not use the Markov Model and do not guarantee finding the globally optimal path, the concepts of integrated representation of knowledge sources and delayed decisions proved to be very valuable.

THE HARPY SYSTEM (Lowerre and Reddy)

The Harpy System (Lowerre 1976) was the first connected speech system to satisfy the original specifications given in the Newell report and was first demonstrated in September of 1976. System design was motivated by an investigation of the important design choices contributing to the success of the Dragon and Hearsay-I systems. The result was a combination of the "best" features of these two systems with additional heuristics to give high speed and accuracy.

Model

The Harpy system uses the locus model of search. The locus model of search, a very successful search technique in speech understanding research, is 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. This technique was instrumental in making Harpy the most successful connected speech understanding system to date. Harpy represents syntactic, lexical, and juncture knowledge in a unified network as in Dragon, but without the a-priori transition probabilities. Phonetic classification is accomplished by a set of speaker-dependent acoustic-phonetic templates based on LPC parameters which represent the acoustic realizations of the phones in the lexical portion of the network.

Performance

The system was tested on several different tasks with different vocabularies and branching factors. On the 1011-word task the system word error rate was 3% and the semantic error rate was 5% (see fig. 1). The system was also tested with connected digits recognition attaining a 2% word error rate. Using speaker-independent templates, error rate increases to 7% over 20 speakers including 10 new speakers. Using telephone input increases the error rate from 7% to 11% depending on the noise characteristics of the telephone system.

Discussion

Backtracking and redundant computation have always been problematic in AI systems. The Harpy system eliminates these in an elegant way, using the beam search technique. By compiling knowledge ahead of time, Harpy achieves a level of efficiency that is unattainable by systems that dynamically interpret their knowledge. This permits Harpy to consider many more alternatives and deal with error and uncertainty in a graceful manner.

THE HEARSAY-II SYSTEM (Erman, Hayes-Roth, Lesser and Reddy)

Hearsay-II has been the major research effort of the CMU speech group over the last three years. During this period, solutions were devised to many difficult conceptual problems that arose during the implementation of Hearsay-I and other earlier efforts. The result represents not only an interesting system design for speech understanding but also an experiment in the area of knowledge-based systems architecture. Attempts are being made by other AI groups to use this type of architecture in image processing and other knowledge-intensive systems.

Hearsay-II is similar to Hearsay-I in that it is based on the hypothesize-and-test-paradigm, using cooperating independent knowledge sources communicating through a global data structure (blackboard). It differs in the sense that many of the limitations and shortcomings of Hearsay-I are resolved in Hearsay-II.

Hearsay-II differs from the Harpy system in that it views knowledge sources as different and independent and thus cannot always be integrated into a single representation. Further, it has as a design goal the ability to recognize, understand, and respond even in situations where sentences cannot be guaranteed to agree with some predefined, restricted language model as is the case with the Harpy system.

Model

The main features of the Hearsay-II system structure are: 1) the representation of knowledge as self-activating, asynchronous, parallel processes, 2) the representation of the partial analysis in a generalized three-dimensional network; the dimensions being level of representation (e.g., parametric, segmental, syllabic, lexical, syntactic), time, and alternatives, with contextual and structural support connections explicitly specified, 3) a modular structure for incorporating new knowledge into the system at any level, and 4) a system structure suitable for execution on a parallel processing system.

Performance

The present system has been tested using about 100 utterances of the training data for the 1011-word vocabulary task. For a grammar with simple syntax (the same one used by Harpy), the sentence error rate is about 16% (semantic error 16%). For a grammar with more complex syntax the sentence error rate is about 42% (semantic error 26%). The system runs about 2 to 20 times slower than Harpy.

Discussion

Hearsay-II represents an important and continuing development in the pursuit of large-vocabulary speech understanding systems. The system is designed to respond in a semantically correct way even when the information is fuzzy and only partial recognition is achieved. Independent knowledge sources are easily written and added to Hearsay-II; knowledge sources may also be removed in order to test their effectiveness. The Hearsay-II system architecture offers great potential for exploiting parallelism to decrease recognition times and is capable of application to other knowledge-intensive AI problems dealing with errorful domains. Many more years of intensive research would be necessary in order to evaluate the full potential of this system.

THE LOCUST SYSTEM (Bisiani, Greer, Lowerre, and Reddy)

Present knowledge representation and search used in Harpy tend to require much memory and are not easily extendable to very large languages (vocabularies of over 10,000 words and more complex syntax). But we do not view this as an insurmountable limitation. Modified knowledge representation designed for use with secondary memories and specialized paging should overcome this difficulty. In addition, it appears larger-vocabulary speech understanding systems can be implemented on mini-computers without significant degradation in performance. Locust is designed to demonstrate the feasibility of these ideas.

Model

The model is essentially the same as the Harpy system except, given the limitations of storage capacity of main memory, the knowledge representation has to be reorganized significantly. The network is assumed to be larger than main memory, stored on secondary memory, and retrieved using a specialized paging mechanism. The choice of the file structure representation and clustering of the states into pages of uniform size are the main technical problems associated with the development of this system.

Discussion

A paging system for the 1011 word vocabulary is currently operational on a PDP-11/40E and has speed and accuracy performance comparable to Harpy on a PDP-10 (KA10). Simulation of various paging models is currently in progress. As memories with decreased access times become available, this class of systems is expected to perform as accurately and nearly as fast as systems requiring no secondary memory.

PARALLEL SYSTEMS (Feiler, Fennell, Lesser, McCracken, and Oleinick)

Response time for the present systems is usually greater than real-time, with indications that larger vocabularies and more complex syntax will require more time for search. One method of achieving greater speed is to use parallel processing. Several systems designed and developed at CMU exploit multi-processor hardware such as Cmmp and Cm*.

Models

Several systems are currently under development as part of multi-processor research projects which attempt to explore potential parallelism of Hearsay and Harpy-like systems. Fennell and Lesser (1977) studied the expected performance of parallel Hearsay systems and issues of algorithm decomposition. McCracken (1977) is studying a production system implementation of the Hearsay model. Oleinick (1977) and Feiler (1977) are studying parallel decompositions of the Harpy algorithm. Several of these studies are not yet complete, but preliminary performance results are very encouraging. Oleinick has demonstrated a version of Harpy that runs faster than real-time on Cmmp for several tasks.

Discussion

The main contribution of these system studies (when completed) will be to show the degree of parallelism which can reasonably be expected in complex speech understanding tasks. Attempts to produce reliable and cost-effective speech understanding systems would require extensive studies in this direction.

DISCUSSION

In the previous section we have briefly outlined the structure and contributions of various speech systems developed at CMU. In retrospect, it is clear that the slow rate of progress in this field is directly attributable to the large combinatorial space of design decisions involved. Thus, one might reasonably ask whether the human research strategy in solving this and other similar problems can benefit from search reduction heuristics that are commonly used in AI programs. Indeed, as we look around, it is not uncommon to find research paradigms analogous to depth-first exploration, breadth-first with shallow cut-off, backtracking, "jumping-to-conclusions", thrashing, and so on.

Our own research has been dominated by two such paradigms. First is a variant of best-first search: find the weakest link (and thus the potential for most improvement) in the system and attempt to improve it. Second is a variant of the beam search: when several alternative approaches look promising, we use limited parallel search with feed-foward. The systems shown in Figure 3 are examples of this type of system iteration and multi-systems approach.

Many system design decisions require an operational total systems framework to conduct experiments. However, it is not necessary to have a single system that permits all possible variations of system designs. Given enough working components, with well-designed interfaces, one can construct new system variants without excessive effort.

The success of the speech understanding research effort is all the more interesting because it is one of the few examples in AI research of a five year prediction that was in fact realized on time and within budget. It is also one of the few examples in AI where adding additional knowledge can be shown to lead to system speed-up as well as improved accuracy.

We note in conclusion that speech understanding research, in spite of the many superficial differences, raises many of the same issues that are central to other areas of AI. Faced with the problem of reasoning in the presence of error and uncertainty, we generate and search alternatives which have associated with them a likelihood value representing the degree of uncertainty. Faced with the problem of finding the most plausible symbolic description of the utterance in a large combinatorial space, we use techniques similar to those used in least-cost graph searching methods in problem solving. Given the problems of acquisition and representation of knowledge, and control of search, techniques used in speech are similar to most other knowledge intensive systems. The main difference is that given human performance the criteria for success, in terms of accuracy and response time, far exceed the performance requirements of other AI tasks except perhaps vision.

ACKNOWLEDGMENT

I would like to thank Gary Goodman and Lee Erman for their help and comments in the preparation of this paper.

REFERENCES

J.K. Baker (1975). "Stochastic Modeling as a Means of Automatic Speech Recognition", Ph.D. Dissertation, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

J.K. Baker (1975). "The Dragon System - An Overview", IEEE Trans. Acoustic., Speech, and Signal Processing, Vol ASSP-23, pp. 24-29, Feb. 1975.

J.K, Baker (1975). "Stochastic Modeling for Automatic Speech Understanding", in Speech Recognition, D.R. Reddy, (Ed.), Academic Press, New York, 1975.

Computer Science Speech Group (1976). "Working Papers in Speech Recognition IV - The Hearsay-II System", Tech. Report, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

L.D. Erman (1974). "An Environment and System for Machine Understanding of Connected Speech", Ph.D. Dissertation, Computer Science Dept., Stanford University, Technical Report, Computer Science Dept., Carnegie-Mellon University, Pittsburgh, PA.

R.D. Fennell and V.R. Lesser (1977). "Parallelism in AI Problem Solving: A Case Study of Hearsay-II", IEEE Trans. on Computers, C-26, pp. 98-111, Feb. 1977.

J.W. Forgie (1974). "An Overview of the Lincoln Laboratory Speech Recognition System", J. Acoust. Soc. Amer., Vol. 56, S27(A).

V.R. Lesser, R.D. Fennell, L.D. Erman, and D.R. Reddy (1975). "Organization of the Hearsay-II Speech Understanding System", IEEE Trans. ASSP-23, No. 1, pp. 11-23.

B.T. Lowerre (1976). "The HARPY Speech Recognition System", Ph.D. Dissertation, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

D. McCracken (1977). "A Parallel Production System for Speech Understanding", Ph.D. Thesis (in preparation), Comp. Sci. Dept., Carnegie-Mellon University, Pittsburgh, PA.

R.B. Neely (1973). "On the Use of Syntax and Semantics in a Speech Understanding System", Ph.D. Dissertation, Stanford University, Technical Report, Computer Science Dept., Carnegie-Mellon University, Pittsburgh, PA.

A. Newell, J. Barnett, J. Forgie, C. Green, D. Klatt, J.C.R. Licklider, J. Munson, R. Reddy, and W. Woods, Speech Understanding Systems: Final Report of a Study Group. North-Holland, 1973. Originally appeared in 1971.

D.R. Reddy, L.D. Erman and R.B. Neely (June 1973). "A Model and a System for Machine Recognition of Speech", IEEE Trans. Audio and Electroacoustics Vol. AU-21, (3), pp. 229-238.

D.R. Reddy, L.D. Erman, R.D. Fennell and R.B. Neely (1973). "The HEARSAY Speech Understanding System: An Example of the Recognition Process", Proc. 3rd Int. Joint Conf. on Artificial Intelligence, Stanford, CA., pp. 185-193.

D.R. Reddy (1976). "Speech Recognition by Machine: A Review", Proc. of the IEEE, Vol. 64, pp. 501-531, May 1976.